FROM TOKENS TO POLICY: *Causal and Interpretable Heterogeneous Treatment Effects Identification*

Anonymous authors
Paper under double-blind review

ABSTRACT

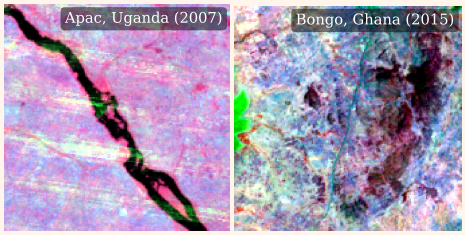
Heterogeneous Treatment Effect (HTE) identification is essential to explain the impact of an intervention and optimize our policies accordingly. Existing approaches trade expressivity for interpretability, but, if some active heterogeneity drivers are unmeasured, methods at both ends of this spectrum 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., multi-modal 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 a provable selection guarantee. We deploy NEXIS on two anti-poverty programs in Africa, augmenting each with satellite imagery capturing previously unmeasured environmental effect modifiers, leading to new *candidate* prescriptive guidelines to optimize the programs' next iterations.

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 1). Standalone demographic covariates are insufficient (Burke et al., 2021; Athey & Imbens, 2017).



(a) Perennial river (b) Ephemeral waterway

Figure 1: New candidate environmental interactors in anti-poverty programs.

Such causal characterization has been historically out of 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

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

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., 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., 2003b; Aliferis et al., 2010) for the effect heterogeneity over the learned dictionary, introducing *Neural EXposure Interaction Search* (NEXIS), a forward-backward Markov-blanket 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 Heterogeneous Treatment Effect

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 Heterogeneous effect characterization via Direct Effect Modifiers

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 no interference (SUTVA), define $\tau := Y(1) - Y(0)$. The treatment effect can vary across units due to the treatment interaction with the complete 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:

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

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

i. It represents the *minimal* and *sufficient* summary of the effect heterogeneity: by definition $\boldsymbol{W}^{\mathrm{dir}}$ d-separates every other 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:

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

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

But Equation 1 is not the unique effect heterogeneity characterization, and $\boldsymbol{W}^{\mathrm{dir}}$ is generally 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 $\boldsymbol{W}^{\mathrm{dir}}$, from their ancestors *indirect* modifiers $\boldsymbol{W}^{\mathrm{ind}}$, the *proxies* $\boldsymbol{W}^{\mathrm{prx}}$ descending from direct modifiers, and modifiers *by common cause* $\boldsymbol{W}^{\mathrm{cc}}$ sharing a common ancestor with a direct modifier, as illustrated in the causal model in Figure 2. Among these, direct modifiers provide the minimal, direct causal characterization targeted here: indirect modifiers can affect τ only through mediation by a direct modifier, 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.

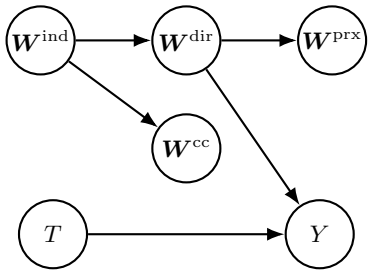


Figure 2: Effect modification taxonomy in a causal model (VanderWeele & Robins, 2007). Note: The illustrated causal relationships are *only* necessary conditions for the described taxonomy. A causal model without structural equations is not sufficient to characterize effect modifications and interactions, e.g., distinguishing between direct effect modifiers from prognostic variables.

Furthermore $\boldsymbol{W}^{\mathrm{dir}}$ 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 1 (Measurement Sufficiency)**.** *The measured pre-treatment variables $\boldsymbol{X}$ entangle all the heterogeneous treatment effect information, i.e.,*

$$\mathbb{E}[\tau \mid \boldsymbol{X}] = \mathbb{E}[\tau \mid \boldsymbol{W}^{\mathrm{dir}}] \qquad \textit{a.s.} \tag{3}$$

Assumption 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 for completeness: if violated, some heterogeneity is unmeasured, so no method can recover its full causal characterization or extrapolate it reliably.

Existing methods can estimate causal CATEs from complex measurements, but generally do not distinguish direct modifiers from correlated proxies (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

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

targets for the first time a causal HTE characterization from complex observations, i.e., identifying the direct modifiers $\boldsymbol{W}^{\mathrm{dir}}$, which is also interpretable by construction due to causal mechanisms sparsity (Schölkopf et al., 2021) and representation minimality.

2.2 Valid representation for Direct Effect Modifiers identification

Pretrained foundation models offer versatile representations of complex pretreatment measurements and often linearly encode semantic attributes (Chen et al., 2020; Dosovitskiy et al., 2020; Radford et al., 2021; Caron et al., 2021); whether they preserve a study's direct effect modifiers is task-dependent. Furthermore, learning end-to-end a meaningful representation on a target experiment is generally statistically infeasible due to the limited data. We then propose to represent the pre-treatment measurements constructing an encoder $\psi : \mathcal{X} \to \mathbb{R}^m$ (with $m \ll |\mathcal{X}|$), e.g., 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[2]. The resulting dictionary coordinates

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

are sparse candidates for post-hoc interpretation (Oquab et al., 2023; Zhai et al., 2023; Park et al., 2023), and each can be labeled, e.g., by inspecting its top-activating inputs via a vision-language model. Two distinct desiderata on the learned representation arise naturally:

Assumption 2 (Representation Sufficiency)**.** *All the heterogeneous-treatment-effect information in the pre-treatment measurements $\boldsymbol{X}$ is preserved in the representation $\boldsymbol{Z}$, i.e.,*

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

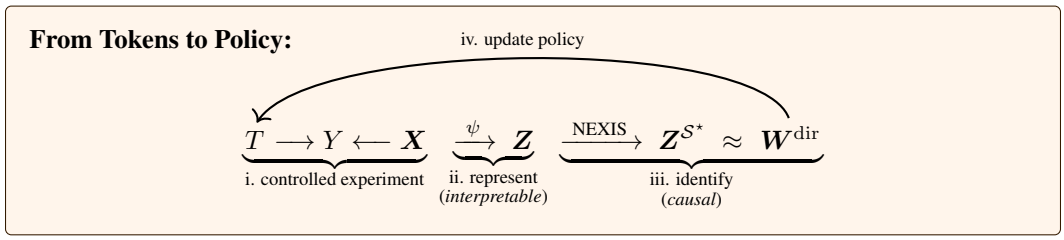
In autoencoders, reconstruction loss encourages information preservation but does not certify Representation Sufficiency, which is necessary for full HTE functional identification.

Assumption 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 $\mathcal{S}^\star := \mathrm{Im}(\pi) = \{j_1, \ldots, j_r\} \subseteq [m]$ the set of principal proxy coordinates *for the (latent) direct effect modifiers, and $\boldsymbol{Z}^{\mathcal{S}^\star} = \{Z^{j_1}, ..., Z^{j_r}\}$ the corresponding* principal proxies*.*

Principal Alignment is a sufficiency, not purity, condition: a distinct principal coordinate screens each direct modifier from the rest, while the coordinate may remain polysemantic and the modifier correlated with other coordinates. 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) (Appendix C.3). Principal Alignment maps the latent characterization of Equation 2 to the observed-space target $\mathcal{S}^\star$. Under faithfulness, Theorem 4.1 makes it the unique mean Markov boundary of the CATE. Faithfulness makes every principal proxy marginally relevant, but not the converse: effect modification is broader than direct interaction, and entanglement transmits its signal to non-principal coordinates. Section 3 shows why greater experimental power widens this gap; Section 4 introduces NEXIS with a selection guarantee, completing the policy-design loop below:



[2]Different hypotheses generation strategies can be considered provided satisfying the method requirements.

3 On the distinction between Effect Modification and Interaction

Marginal effect-modification testing is the natural first attempt to recover $\mathcal{S}^\star$. Faithfulness ensures population recall: every principal proxy is marginally relevant. But the converse fails. Effect modification need not reflect direct interaction, and learned representations entangle modifier signal across coordinates, so non-principal coordinates can be genuinely relevant. The target is the minimal sufficient representation of the CATE, not all marginal modifiers.

Effect modification entanglement. In learned representations, a coordinate can principally proxy a non-direct modifier and non-principally proxy many others through entanglement (Chanin et al., 2024), regardless of Principal Alignment. Let

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

the marginal effect-modifier coordinates within the learned representation. Every $j_k \in \mathcal{S}^\star$ belongs to $\mathcal{M}^\star$[3], but so does any entangled coordinate carrying real modifier signal. In a reconstruction-rich dictionary with $m \gg r$, non-principal proxies vastly outnumber the principal ones, making $\mathcal{M}^\star$ a redundant, non-causal superset of the target (VanderWeele, 2009). Marginal activation magnitude cannot separate them: a non-principal proxy of a strong direct modifier can outrank the principal proxy of a weaker one, and the cutoff is unknown.

Experimental Power Paradox. Marginal testing fails structurally, not from low power. As power grows through larger n, stronger signals, or more sensitive tests, it converges to $\mathcal{M}^\star$ and *accumulates* non-direct modifiers backed by genuine evidence. It is consistent for the wrong target: power widens the visible gap from $\mathcal{S}^\star$. Recovering $\mathcal{S}^\star$ instead requires conditional tests that rule out redundant candidates, the design principle behind NEXIS.

Empirical evidence. Similarly to Mencattini et al. (2025), we design a semi-synthetic benchmark simulating different latent data generating processes pairing them with real-world face images from CelebA (Liu et al., 2018) as observed pre-treatment measurements. A TopK SAE (m=9,216 codes) trained on SigLIP 2 patch representations (Tschannen et al., 2025) provides the candidate dictionary; Appendix C.1 details its construction and the supervised principal-coordinate labeling used to assess alignment. Figure 3 reports Precision and Recall in recovering the two principal coordinates for `wearing a hat` and `wearing eyeglasses` with 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 recovers $\mathcal{S}^\star$ increasingly well as power grows. Experiment details and extended ablations are reported in Appendix C.

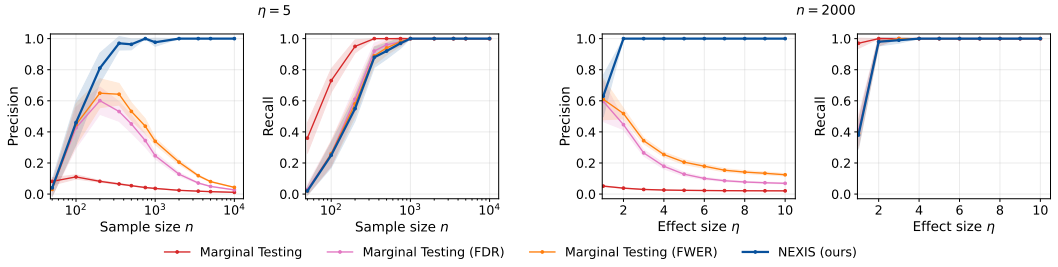


Figure 3: *Experimental power paradox* on CelebA: Increasing the sample size (n) or the effect magnitude (η) of the experiment, marginal screening accumulates non-direct modifiers (false discoveries) and diverges from $\mathcal{S}^\star$ recovery, i.e., precision collapse. NEXIS robustly mitigates such behavior by iterative conditioning.

4 Neural EXposure Interaction Search

Given a randomized experiment and pre-treatment representation $\boldsymbol{Z}$, the task is to recover $\mathcal{S}^\star$ without observing $\boldsymbol{W}^{\text{dir}}$. *Neural EXposure Interaction Search* (NEXIS) adds the coordinate carrying the strongest residual heterogeneity signal, conditional on the current selection, then removes any coordinate redundant given a subset of the others. It uses *CATE-equivalence tests*: for candidate $j \in [m]$ and selection $S \subseteq [m] \setminus \{j\}$, $p_j(S)$ is a valid p-value for

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

Instantiation details and suggested defaults are presented in Appendix B.

[3]Assuming local mean-adjacency faithfulness clause (see Assumption 4 with $A = \emptyset$).

Algorithm 1 Neural EXposure Interaction Search (NEXIS)

Input: randomized experiment with pre-treatment representations; significance $\alpha \in (0, 1)$
$S \leftarrow \emptyset,\ S_{\text{prev}} \leftarrow \{-1\}$
while $S \neq S_{\text{prev}}$ and $|S| < m$ **do** ▷ forward search
  $S_{\text{prev}} \leftarrow S, \quad \bar{S} \leftarrow [m] \setminus S$
  $j^\star \leftarrow \arg\min_{j \in \bar{S}} p_j(S)$
  if $p_{j^\star}(S) \leq \alpha/|\bar{S}|$ **then** $S \leftarrow S \cup \{j^\star\}$
  end if
end while
$S \leftarrow \{j \in S : \max_{A \subseteq S \setminus \{j\}} p_j(A) \leq \alpha/m\}$ ▷ terminal backward certification
return S

NEXIS adapts classic forward–backward Markov-blanket discovery (Tsamardinos et al., 2003b; Aliferis et al., 2010) to the mean Markov boundary of the unobserved CATE over a learned representation. Its forward step is IAMB's grow phase (Tsamardinos et al., 2003b); after it stops at $\widetilde{S}$, the terminal backward step applies the MMPC/HITON-PC subset rule (Tsamardinos et al., 2003a; Aliferis et al., 2010) at level α/m. On the recall event, a false coordinate survives only by rejecting the fixed true null $H_0(j \mid \mathcal{S}^\star)$, so precision needs validity only there. The terminal step costs at most $|\widetilde{S}|\, 2^{|\widetilde{S}|-1}$ tests, practically feasible by the causal mechanism sparsity principle. The Bonferroni forward gate (Bonferroni, 1936) can be relaxed, e.g., to FDR (Benjamin et al., 1995); Appendices B.3 and B.4 discuss the optional interleaved backward step and practical spectral-gap default ρ.

Theorem 4.1 (Causal Identification)**.** *In a randomized experiment with Principal Aligned pre-treatment measures with faithfulness and test validity/consistency, NEXIS' outcome $\widehat{\mathcal{S}}_n$ satisfies:*

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

and, additionally assuming Measurement and Representation Sufficiency:

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

Proof sketch. Principal Alignment and composition faithfulness guarantee $\mathcal{S}^\star$ sufficiency, while local mean-adjacency faithfulness guarantees its necessity. Test consistency guarantees asymptotic recall, while the terminal step and test validity bound asymptotic false retention by α, testing only at the fixed set $\mathcal{S}^\star$ rather than at data-dependent ones (post-selection inference). Measurement and Representation Sufficiency supply the causal interpretation. Full proof in Appendix A. □

Principal Alignment[4], faithfulness, and test validity and consistency ensure recovery of the principal proxies; Measurement and Representation Sufficiency further guarantee that such proxies capture all the heterogeneity information, thereby identifying the joint (latent direct modifiers) interventional effect. As such, Theorem 4.1 supports policy hypotheses and counterfactual comparisons, e.g., across bundled program configurations. See Appendix B for implementation details and practical 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 require 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

[4]Without Principal Alignment, the one-coordinate-per-modifier target need not be unique.

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: Heterogeneous Treatment Effect estimation properties by family.

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

Empiricist: data-driven feature extraction from raw pre-treatment observations.
Interpretable: human-readable HTE characterization.
Prescriptive: targets treatment-interactive factors.
✓ = fully, ✗ = not fully.

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 apply NEXIS to two cash-transfer evaluations augmented with satellite imagery. In each case, it surfaces interpretable candidate effect modifiers, including landscape features absent from the original surveys. With no ground truth, these results are hypotheses for field validation, not established causal modifiers or policy prescriptions. For quantitative validation on semi-synthetic experiments, see Appendix C.

Results: ***From demographic to environmental anti-poverty programs impact heterogeneity.*** In both programs, the selected modifiers are mainly environmental factors, coarsely proxied or absent in survey data, while no individual-level demographic is selected.

6.1 Case study 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 535 groups (439 in our analytic sample). 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[5] 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) (Bai et al., 2025).

(a) Latent direct modifiers of *skilled employment*. (b) Latent direct modifiers of *business assets*.

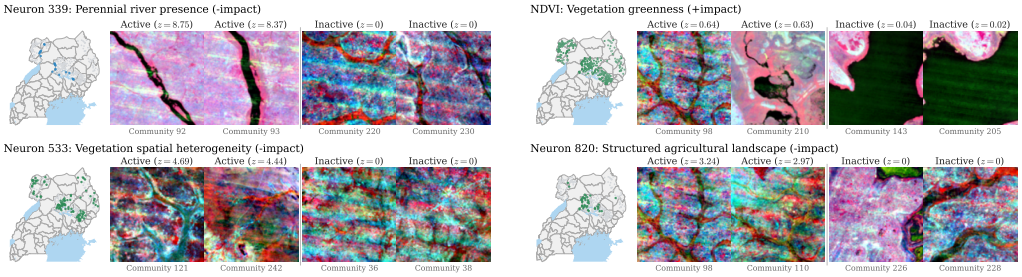


Figure 4: Satellite-derived discoveries on YOP program impact heterogeneity by NEXIS, with VLM interpretations obtained by contrasting top- and bottom-activating tiles; neurons 339 and 820 are candidates. Each row pairs the Uganda activation map of a feature with its two top- and bottom-activating Landsat tiles in false-color composite (NIR, Green, SWIR).

Results. NEXIS retains two ethnolinguistic groups and one environmental modifier for the first target and NDVI for the second; three additional *candidates* fail the backward step (Figure 4, Appendix D.4). No individual-level demographic (e.g., age, sex, parental education), all very homogeneous by experimental design, is selected.

(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 and the candidate 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 (a candidate) 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 agricultural landscape, a candidate, instead dampens returns, plausibly reflecting

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

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. By contrast, 71 and 45 of the 170 coordinates pass an uncorrected marginal test, mostly correlated proxies of these few direct modifiers: a real-world instance of the experimental power paradox discussed in Section 3.

6.2 Case study 2: Livelihood Empowerment Against Poverty 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 Northern 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 an 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).

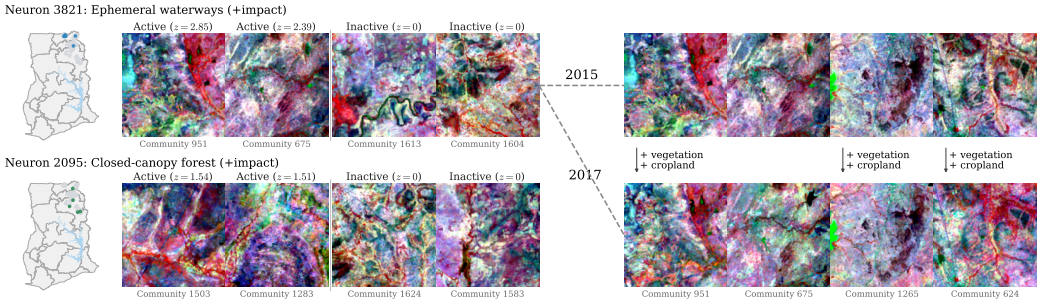


Figure 5: *Left*: Satellite-derived discoveries on LEAP1000 programme impact heterogeneity by NEXIS, with VLM interpretations by contrasting top- and bottom-activating Landsat tiles in false-color composite. *Right*: Temporal evolution of land use in ephemeral-waterway-active communities, showing cropland extension.

Results. NEXIS retains two satellite coordinates labelled ephemeral waterways and closed-canopy forest (Figure 5), and no demographics; 18 of 167 coordinates pass an uncorrected marginal test. In communities with seasonal water corridors, estimated impact is several times the average; a similarly amplified pattern appears in rare forest-patch communities. We hypothesize *complementarity* (De Mel et al., 2008; Prifti et al., 2020): water-adjacent smallholders can invest transfers in irrigation-season agriculture, while forest-edge households can combine cash with non-timber livelihoods. We further support the former hypothesis by tracking cropland expansion and denser riparian vegetation in VLM comparisons of 2015 and 2017 imagery. No demographic modifier is detected at this threshold; this does not establish invariance.

7 Conclusion

Identifying causal and interpretable HTE characterizations is essential for trustworthy policy design. We show that this target is within reach by combining extensive measurements with learned representations, and introduce the first algorithm for the task. Applications to two anti-poverty programs uncover previously overlooked environmental heterogeneity and motivate future adaptations. NEXIS recovers only modifiers that the measurements capture, so future evaluations should invest in richer, multi-source pre-treatment measurement. Structural assumptions determine the target and causal interpretation; valid and consistent tests determine the selection guarantee. Violations weaken either conclusion, so discoveries require cautious interpretation and experimental validation.

AI USE STATEMENT

In this work, we used generative AI tools as coding assistants to support the implementation and modification of the codebase and for language editing. These tools were not used to autonomously generate research results or replace our scientific judgment. We determined the research methodology and experimental design, reviewed and validated all AI-assisted implementations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We collected no new data and did not interact with participants. The two applications reuse de-identified microdata from completed evaluations run under their own institutional approvals: the Youth Opportunities Program data come from the public replication release of Jerzak et al. (2023), and the LEAP 1000 data are proprietary and were used under agreement with the data owners. Satellite imagery enters only at the level of community tiles, and no household record or precise location is released. NEXIS discoveries are candidate effect modifiers, certified only under assumptions that cannot be fully verified in practice; they should inform the design of future programs and be validated experimentally, not used to exclude communities from assistance.

REPRODUCIBILITY STATEMENT

The assumptions are stated in the main text, with their full formal statements and all proofs in Appendix A; Appendix B details every component of NEXIS with a minimal implementation. We release code that reproduces every semi-synthetic experiment of Appendix C end to end, one command per block, with seeds and compute reported. The applications rely on proprietary survey data and on satellite imagery that cannot be redistributed, so we cannot release their full analyses, but every step (data, representation, NEXIS configuration and interpretation) is documented in Appendices D and E.

REFERENCES

Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. 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.

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

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

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

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

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

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

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. OpenAI tech report, 2023.

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

Philippe Boileau, Ning Leng, Nima S Hejazi, Mark van der Laan, and Sandrine Dudoit. 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.

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

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

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

Riccardo Cadei, Ilker Demirel, Piersilvio De Bartolomeis, Lukas Lindorfer, Sylvia Cremer, Cordelia Schmid, and Francesco Locatello. Prediction-powered causal inferences. *arXiv preprint arXiv:2502.06343*, 2025.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

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

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Bruno Crépon, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. Do labor market policies have displacement effects? evidence from a clustered randomized experiment. *The quarterly journal of economics*, 128(2):531–580, 2013.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

Dennis Egger, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael Walker. General equilibrium effects of cash transfers: experimental evidence from kenya. *Econometrica*, 90(6): 2603–2643, 2022.

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

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

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 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`.

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

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *nature*, 585(7825):357–362, 2020.

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

Oliver Hines, Karla Diaz-Ordaz, and Stijn Vansteelandt. Variable importance measures for heterogeneous causal effects. *arXiv preprint arXiv:2204.06030*, 2022.

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

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

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

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

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

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

Junning Li and Z Jane Wang. Controlling the false discovery rate of the association/causality structure learned with the pc algorithm. *Journal of Machine Learning Research*, 10(2), 2009.

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

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

Anton Rask Lundborg, Ilmun Kim, Rajen D Shah, and Richard J Samworth. The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52(6):2851–2878, 2024.

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

Aleksandar Makelov, George Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv preprint arXiv:2405.08366*, 2024.

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

Tommaso Mencattini, Riccardo Cadei, and Francesco Locatello. Exploratory causal inference in saence. *arXiv preprint arXiv:2510.14073*, 2025.

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.

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

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

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*, 2025.

Joseph Paillard, Angel Reyero Lobo, Vitaliy Kolodyazhniy, Bertrand Thirion, and Denis A Engemann. Measuring variable importance in heterogeneous treatment effects with confidence. *arXiv preprint arXiv:2408.13002*, 2024.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

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

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

Ervin Prifti, Silvio Daidone, Noemi Pace, and Benjamin Davis. 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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Thomas S Richardson and James M Robins. 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.

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

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

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

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

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

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

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

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

Dario Simionato and Fabio Vandin. Bounding the family-wise error rate in local causal discovery using rademacher averages. *Data Mining and Knowledge Discovery*, 38(6):4157–4183, 2024.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. 2000.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Orsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique De Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

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

Ioannis Tsamardinos and Laura E Brown. Bounding the false discovery rate in local bayesian network learning. In *AAAI*, pp. 1100–1105, 2008.

Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 673–678, 2003a.

Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. Algorithms for large scale markov blanket discovery. In *FLAIRS*, volume 2, pp. 376–81, 2003b.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

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

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

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

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

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

Dingling Yao, Shimeng Huang, Riccardo Cadei, Kun Zhang, and Francesco Locatello. The third pillar of causal analysis? a measurement perspective on causal representations. *arXiv preprint arXiv:2505.17708*, 2025a.

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

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

Appendix

TABLE OF CONTENTS

A Proof of Theorem 4.1 **17**
A.1 Markov-blanket formulation . . . 18
A.2 Selection guarantee . . . 19
A.3 Causal Heterogeneous Treatment Effect identification . . . 20

B Method Details **22**
B.1 CATE-equivalence test . . . 22
B.2 Error control . . . 24
B.3 Spectral-gap gate . . . 24
B.4 Backward steps . . . 25
B.5 Minimal Python Implementation Snippet . . . 26

C Semi-synthetic experiments (CelebA) **28**
C.1 Experimental setup . . . 28
C.2 Main results . . . 30
C.3 Testing Principal Alignment with supervision . . . 33
C.4 Model ablations: SAE design choices . . . 35
C.5 Method ablations: NEXIS design choices . . . 37
C.6 Additional data generating process ablations . . . 42
C.7 Practical guidance and summary . . . 45
C.8 Computational budget . . . 46

D Application 1: Youth Opportunities Program (Uganda) **47**
D.1 Trial design and outcomes . . . 47
D.2 Satellite imagery pipeline . . . 48
D.3 NEXIS instantiation . . . 48
D.4 Discovered modifiers . . . 49
D.5 Marginal screening vs. NEXIS . . . 50
D.6 VLM interpretation protocol . . . 51
D.7 Limitations . . . 51
D.8 Computational budget . . . 52

E Application 2: LEAP 1000 Programme (Ghana) **53**
E.1 Study design and identification strategy . . . 53
E.2 NEXIS under quasi-experimental identification . . . 54
E.3 Satellite data and SAE pipeline . . . 55
E.4 NEXIS configuration . . . 55
E.5 Discovered effect modifiers . . . 56
E.6 VLM interpretation protocol . . . 57
E.7 Exploratory analysis . . . 57
E.8 Limitations . . . 58
E.9 Computational budget . . . 58

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 (Faithfulness)**.** *The following two conditions hold. First, the component-wise conditional independences between the direct modifiers and the non-principal coordinates compose jointly:*

$$W^{\mathrm{dir},k} \perp\!\!\!\perp \boldsymbol{Z}^{[m]\setminus\mathcal{S}^\star} \mid \boldsymbol{Z}^{\mathcal{S}^\star} \quad \textit{for every } k \in [r] \quad \Longrightarrow \quad \boldsymbol{W}^{\mathrm{dir}} \perp\!\!\!\perp \boldsymbol{Z}^{[m]\setminus\mathcal{S}^\star} \mid \boldsymbol{Z}^{\mathcal{S}^\star}. \tag{11}$$

Second, the CATE is locally mean-adjacency faithful to the principal proxies: for every $j_k \in \mathcal{S}^\star$ and every $A \subseteq [m] \setminus \{j_k\}$, the empty set included,

$$\mathbb{P}\Big(\mathbb{E}[\tau \mid \boldsymbol{Z}^{A\cup\{j_k\}}] \neq \mathbb{E}[\tau \mid \boldsymbol{Z}^{A}]\Big) > 0. \tag{12}$$

Principal Alignment is a screening condition, not a relevance condition: a principal proxy may be polysemantic and its modifier correlated with other coordinates, provided that the principal proxy screens the modifier from the rest of the dictionary. The first faithfulness clause is a restricted composition property, implied by standard DAG faithfulness, that combines these coordinate-wise screening statements (Spirtes et al., 2000). The second is the local mean-level analogue of adjacency faithfulness used in local causal discovery (Aliferis et al., 2010; Simionato & Vandin, 2024): no subset of the other dictionary coordinates can render a principal proxy conditionally mean-null, including through exact substitution or cancellation. This uniform form is stronger than marginal relevance and is required because the terminal backward step tests every subset. It is specific to the conditional-mean functional targeted by NEXIS; targeting a different heterogeneity functional would require the corresponding local non-redundancy condition.

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 valid, i.e., $\limsup_n \mathbb{P}(p_j(S) \le t) \le t$ for every $t \in (0,1)$;

(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 proof invokes clause (a) only at the fixed set $S = \mathcal{S}^\star$, i.e., for the $m - |\mathcal{S}^\star|$ true nulls $H_0(j \mid \mathcal{S}^\star)$, and clause (b) only for the principal proxies, against $H_0(j_k \mid A)$ for every $A \subseteq [m] \setminus \{j_k\}$; the assumption can be weakened to these instances. We state it for all fixed S because $\mathcal{S}^\star$ is unknown, so in practice a test is justified uniformly; for the linear interaction test, the weaker form needs correct specification only at $\mathcal{S}^\star$. Clause (a) supplies precision, while clause (b) supplies forward recall and retention of the principal proxies by the terminal backward step; neither argument requires independence among the p-values. The assumption is otherwise agnostic to the test instantiation. By default we adopt a linear treatment-interaction test, which satisfies it under a correctly specified conditional linear model and standard moment, nonsingularity, and variance-estimation conditions, and is consistent against alternatives with a nonzero conditional linear interaction. For nonlinear conditional CATEs, the projected covariance measure can target a broader alternative class when its projection learner is sufficiently rich (Lundborg et al., 2024). Appendix B.1 states the validity and power conditions of each construction, and Appendix C.5 reports their ablations.

Theorem (Causal Identification)**.** *Given a randomized experiment $\{(\boldsymbol{z}_i, t_i, y_i)\}_{i=1}^n$ with Principally Aligned pre-treatment measurements $\boldsymbol{z}_i \in \mathbb{R}^m$ (Assumption 3), under faithfulness (Assumption 4) and tests validity and consistency (Assumption 5), NEXIS' outcome $\widehat{\mathcal{S}}_n$ satisfies:*

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

and, additionally assuming Measurement Sufficiency (Assumptions 1) and Representation Sufficiency (Assumptions 2):

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

The proof proceeds in three steps:

i. **Markov-blanket formulation:** Under Principal Alignment and faithfulness, $\mathcal{S}^\star$ is the mean Markov boundary of the CATE on the learned dictionary $[m]$: Lemma A.1 gives sufficiency using the composition clause, while the local mean-adjacency clause gives non-redundancy given any subset of the other coordinates. *Measurement and Representation Sufficiency are not invoked here:* these are conditions on the dictionary-level target alone.

ii. **Selection guarantee:** Combined with test validity at fixed conditioning sets, these facts give asymptotic recall of the forward step and retention of $\mathcal{S}^\star$ by the terminal backward step, while every false coordinate the terminal backward step keeps must reject the fixed true null $H_0(j \mid \mathcal{S}^\star)$. This bounds the asymptotic probability of any false discovery by α, giving the exact-recovery guarantee.

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 the composition clause of Assumption 4, the principal proxies are sufficient for the CATE over the dictionary: for every $S \supseteq \mathcal{S}^\star$ (with $S \subseteq [m]$),*

$$\mathbb{E}[\tau \mid \boldsymbol{Z}^S] = \mathbb{E}[\tau \mid \boldsymbol{Z}^{\mathcal{S}^\star}] = \mathbb{E}[\tau \mid \boldsymbol{Z}] \quad \textit{a.s.} \tag{13}$$

Proof. It is enough to establish the following statement for $R = [m] \setminus \mathcal{S}^\star$, since decomposition then gives it for every $R \subseteq [m] \setminus \mathcal{S}^\star$:

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

which directly implies $\mathbb{E}[\tau \mid \boldsymbol{Z}^{\mathcal{S}^\star \cup R}] = \mathbb{E}[\tau \mid \boldsymbol{Z}^{\mathcal{S}^\star}]$ a.s. Equation 13 follows 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 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. Repeated application of Equation 11 combines these per-k statements 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 a Markov blanket of the treatment-effect heterogeneity (Section 2.1), $\tau \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{W}^{\mathrm{dir}}$. Weak union gives $\tau \perp\!\!\!\perp \boldsymbol{Z}^{[m]\setminus\mathcal{S}^\star} \mid (\boldsymbol{W}^{\mathrm{dir}}, \boldsymbol{Z}^{\mathcal{S}^\star})$. Combining this with Equation 17 by symmetry, contraction, and decomposition yields Equation 14 for $R = [m] \setminus \mathcal{S}^\star$, and therefore for every subset R. □

Reading Lemma A.1. The lemma says $\mathcal{S}^\star$ captures all the heterogeneity-relevant information $\boldsymbol{Z}$ carries about τ, and adding non-principal coordinates does not destroy sufficiency. Together with the local mean-adjacency clause of Assumption 4, this makes $\mathcal{S}^\star$ the unique mean Markov boundary of the CATE over the dictionary. Indeed, if a sufficient A omitted j_k, then conditioning $\mathbb{E}[\tau \mid \boldsymbol{Z}^A] = \mathbb{E}[\tau \mid \boldsymbol{Z}]$ on $\boldsymbol{Z}^{A\cup\{j_k\}}$ would give $\mathbb{E}[\tau \mid \boldsymbol{Z}^{A\cup\{j_k\}}] = \mathbb{E}[\tau \mid \boldsymbol{Z}^A]$, making $H_0(j_k \mid A)$ true and contradicting Equation 12. These facts do not require Measurement or Representation Sufficiency, which determine how much of the true heterogeneity the dictionary captures rather than its internal structure around $\mathcal{S}^\star$.

A.2 Selection guarantee

Proof of Equation 9. Let $\widetilde{S}_n$ be the selection at which the forward step of Algorithm 1 stops ($\widetilde{S}$ in Section 4 and Appendix B.4), so that the terminal backward step returns

$$\widehat{\mathcal{S}}_n = \big\{j \in \widetilde{S}_n : p_j(A) \le \alpha/m \text{ for every } A \subseteq \widetilde{S}_n \setminus \{j\}\big\}. \tag{18}$$

We prove (i) recall of the forward step, $\mathbb{P}(\mathcal{S}^\star \subseteq \widetilde{S}_n) \to 1$, then (ii) that the terminal backward step keeps $\mathcal{S}^\star$, then (iii) precision, $\limsup_n \mathbb{P}(\mathcal{S}^\star \subseteq \widetilde{S}_n,\, \widehat{\mathcal{S}}_n \not\subseteq \mathcal{S}^\star) \le \alpha$, and finally combine.

(i) Recall of the forward step. Fix any $S \subseteq [m]$ with $\mathcal{S}^\star \not\subseteq S$, let $\bar{S} = [m] \setminus S$, and pick any $j_k \in \mathcal{S}^\star \setminus S$. The local mean-adjacency clause of Assumption 4, applied with $A = S$, gives that $H_0(j_k \mid S)$ 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 at selection S 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 $p_{j_k}(S)$. Define the favorable event $\mathcal{E}_n$ as the intersection of the events $\{\min_{j\in\bar{S}} p_j(S) \le \alpha/|\bar{S}|\}$ over all such S. This is a finite intersection, so $\mathbb{P}(\mathcal{E}_n) \to 1$. The forward step adds one coordinate per round and never removes one, so it stops after at most m rounds, and on $\mathcal{E}_n$ it cannot stop at a selection that misses a coordinate of $\mathcal{S}^\star$. Hence $\mathbb{P}(\mathcal{S}^\star \subseteq \widetilde{S}_n) \to 1$.

(ii) The terminal backward step keeps $\mathcal{S}^\star$. By the local mean-adjacency clause of Assumption 4, $H_0(j_k \mid A)$ is false for every $j_k \in \mathcal{S}^\star$ and every $A \subseteq [m] \setminus \{j_k\}$. Since m is fixed, this is a finite family of fixed alternatives, and Assumption 5(b) with a union bound yields

$$\mathbb{P}\big(\exists\, j_k \in \mathcal{S}^\star,\ \exists\, A \subseteq [m] \setminus \{j_k\} : p_{j_k}(A) > \alpha/m\big) \to 0. \tag{20}$$

The terminal backward step tests each $j_k \in \widetilde{S}_n$ only at subsets $A \subseteq \widetilde{S}_n \setminus \{j_k\} \subseteq [m] \setminus \{j_k\}$, and makes all removals simultaneously, so these are the only tests that decide on j_k. Hence, with probability $\to 1$, every $j_k \in \mathcal{S}^\star \cap \widetilde{S}_n$ is kept, and by (i), $\mathbb{P}(\mathcal{S}^\star \subseteq \widehat{\mathcal{S}}_n) \to 1$.

(iii) Precision. Let $\mathcal{R}_n := \{\mathcal{S}^\star \subseteq \widetilde{S}_n\}$ and fix $j \in [m] \setminus \mathcal{S}^\star$. On $\mathcal{R}_n$, if $j \in \widehat{\mathcal{S}}_n$, then $j \in \widetilde{S}_n$ and $\mathcal{S}^\star \subseteq \widetilde{S}_n \setminus \{j\}$, so $A = \mathcal{S}^\star$ is one of the subsets at which j is tested, and $p_j(\mathcal{S}^\star) \le \alpha/m$. By Lemma A.1,

$$\mathbb{E}[\tau \mid \boldsymbol{Z}^{\mathcal{S}^\star\cup\{j\}}] = \mathbb{E}[\tau \mid \boldsymbol{Z}^{\mathcal{S}^\star}] \quad \text{a.s.}, \tag{21}$$

so $H_0(j \mid \mathcal{S}^\star)$ holds at the population level. Its conditioning set $\mathcal{S}^\star$ is fixed, not selected from the data. A union bound and Assumption 5(a) over the $m - |\mathcal{S}^\star|$ coordinates outside $\mathcal{S}^\star$ yield

$$\limsup_n \mathbb{P}\big(\mathcal{R}_n,\, \widehat{\mathcal{S}}_n \not\subseteq \mathcal{S}^\star\big) \;\le\; \sum_{j\notin\mathcal{S}^\star} \limsup_n \mathbb{P}\big(p_j(\mathcal{S}^\star) \le \alpha/m\big) \;\le\; \frac{m - |\mathcal{S}^\star|}{m}\,\alpha \;\le\; \alpha. \tag{22}$$

(iv) Combination. Since $\widehat{\mathcal{S}}_n \subseteq \widetilde{S}_n$, the event $\{\mathcal{S}^\star \subseteq \widehat{\mathcal{S}}_n\}$ is contained in $\mathcal{R}_n$, so

$$\mathbb{P}\big(\widehat{\mathcal{S}}_n \neq \mathcal{S}^\star\big) \;\le\; \mathbb{P}\big(\mathcal{S}^\star \not\subseteq \widehat{\mathcal{S}}_n\big) + \mathbb{P}\big(\mathcal{R}_n,\, \widehat{\mathcal{S}}_n \not\subseteq \mathcal{S}^\star\big). \tag{23}$$

By (ii) the first term vanishes, and by (iii) the second has $\limsup$ at most α, so $\liminf_n \mathbb{P}(\widehat{\mathcal{S}}_n = \mathcal{S}^\star) \ge 1 - \alpha$. No step uses test validity at a data-dependent conditioning set: (i) and (ii) apply Assumption 5(b) to finitely many fixed pairs (j, A), and (iii) applies Assumption 5(a) at the fixed set $\mathcal{S}^\star$ only. □

Role of the corrections. The Bonferroni gate in the forward step is not needed for the stated error bound: with fixed m, test consistency and any fixed positive admission threshold suffice for recall. The correction that supplies precision is the terminal level α/m, used in step (iii) to control the $m - |\mathcal{S}^\star|$ fixed nulls $H_0(j \mid \mathcal{S}^\star)$. No additional correction over the exponentially many subset tests is needed for this argument, because a false survivor is bounded using only its test at $\mathcal{S}^\star$. Another terminal rule could replace α/m if it provided the same family-wise bound over these fixed nulls. The fixed-m asymptotic treats the learned dictionary as fixed before the experiment; allowing m to grow with n would require uniform rates for the tests used in the finite-intersection and union-bound arguments above.

Role of the terminal backward step. The terminal backward step is what supplies precision. Without it, NEXIS would still satisfy recall by (i), 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. The terminal backward step re-tests it at every subset of $\widetilde{S}_n \setminus \{j'\}$, which on $\mathcal{R}_n$ includes the fixed set $\mathcal{S}^\star$, where Assumption 5(a) applies. Bounding a false coordinate by its p-value at a fixed true conditioning set is the device behind existing error-control bounds for local causal discovery (Tsamardinos & Brown, 2008; Li & Wang, 2009). The interleaved backward step of Appendix B.4, which re-tests each j at $S \setminus \{j\}$ after every forward addition, is not required for the guarantee, and it is not sufficient alone: the current selection S is chosen by the data, so $H_0(j \mid S \setminus \{j\})$ is a data-dependent null, outside the scope of Assumption 5, which is the post-selection issue. It remains a finite-sample variant for settings with many or entangled modifiers.

Role of the forward step. By the local mean-adjacency clause of Assumption 4, the terminal backward step alone, applied to the whole dictionary $[m]$, would already satisfy steps (ii)–(iv): every $j_k \in \mathcal{S}^\star$ is non-null given any subset, and a false j must reject the fixed true null $H_0(j \mid \mathcal{S}^\star)$. Exhaustive subset testing of this kind underlies family-wise error control in local causal discovery (Simionato & Vandin, 2024), but here it would take $m\,2^{m-1}$ tests, conditioning on up to $m-1$ coordinates with n observations. The forward step restricts the terminal step to at most $|\widetilde{S}_n|\,2^{|\widetilde{S}_n|-1}$ tests on smaller conditioning sets. In the intended sparse-mechanism regime (Schölkopf et al., 2021), few direct modifiers imply $|\mathcal{S}^\star| \ll m$, and the forward step is expected to leave $|\widetilde{S}_n| \ll m$ in practice. This is a computational and finite-sample motivation, not an additional condition for the guarantee. Appendix C.3 tests a controlled violation of Principal Alignment, in which one principal coordinate is split in two and the Markov boundary grows to three coordinates; NEXIS still recovers it, with less power at small n.

A.3 Causal Heterogeneous Treatment Effect identification

Proof of Equation 10. Chaining Lemma A.1 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, $\boldsymbol{W}^{\mathrm{dir}}$ contains every pre-treatment variable with a direct causal effect on τ (Section 2.1); every other pre-treatment variable affects τ, if at all, only through $\boldsymbol{W}^{\mathrm{dir}}$. Hence the joint intervention on $\boldsymbol{W}^{\mathrm{dir}}$ leaves no open back-door path to τ (Pearl, 2009; Richardson & Robins, 2013), 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 interventional is its characterization by conditioning on a non-direct modifier: indirect modifiers can affect τ only through $\boldsymbol{W}^{\mathrm{dir}}$, while proxies and common-cause modifiers have no causal pathway to τ. Hence conditioning on a generic non-direct modifier does not in general identify intervening on it, i.e., $\mathbb{E}[\tau \mid do(\cdot)] \neq \mathbb{E}[\tau \mid \cdot]$. 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 step followed by a terminal backward step, built around a generic CATE equivalence test, with a Bonferroni-style multiple-testing correction in the forward step. 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.5, and to provide a minimal Python implementation. The four axes we cover are (i) the CATE equivalence test (Appendix B.1), (ii) the multiple-testing correction in the forward step (Appendix B.2), (iii) the spectral gap ρ, which is not part of Algorithm 1 but is our practical default against the residual entanglement of real dictionaries (Appendix B.3), and (iv) the backward steps, i.e., the terminal backward step and the optional interleaved backward step (Appendix B.4). The implementation snippet of Appendix B.5 uses the simplest combination, the linear test with Bonferroni correction in the forward step followed by the terminal backward step, which is also the default we recommend for routine use based on the controlled experiments, together with $\rho = 0.5$.

B.1 CATE-EQUIVALENCE TEST

Each NEXIS search step is built around a CATE equivalence test: 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 valid under the null and consistent against fixed alternatives (Assumption 5); any such test plugs in. We consider three test families, spanning the parametric–nonparametric spectrum: the linear test, and the doubly-robust GCM and PCM, each of the latter two with a quadratic and a LightGBM variant.

(i) 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 t-test on the interaction coefficient 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. What matters is that the conditional CATE $\mathbb{E}[\tau \mid \boldsymbol{Z}^{S\cup\{j\}}]$ be close to linear in the coordinates. This is a property of the data-generating process that a dictionary can expose but not create: when τ is linear in the direct modifiers and the dictionary aligns them to near-indicator coordinates, the linear test is well specified. The controlled CelebA experiments are in this regime: the linear test reaches recall 0.95 at $n = 750$ when $\eta = 5$ and at $\eta = 2$ when $n = 2000$, against $n = 2000$ and $\eta = 4$ for both GCM variants below, and $n = 3500$ for both PCM variants, which never reach it at $n = 2000$ (see Figure 13 and the discussion in Appendix C.5). We adopt it as the default throughout.

(ii) Doubly-robust GCM. When linearity is unwarranted, we replace Equation 27 with a doubly-robust pseudo-outcome plus a residual-product test in the spirit of Shah & Peters (2020). The R-learner residualization (Robinson, 1988; Nie & Wager, 2021)

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

with $e := \mathbb{P}(T = 1)$ known by design and $\widehat{m}$ a cross-fitted estimator of $\mathbb{E}[Y \mid \boldsymbol{Z}^S]$, 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 two variants differ only in the outcome-side nuisances $\widehat{m} = \mathbb{E}[Y \mid \boldsymbol{Z}^S]$ and $\widehat{\mathbb{E}}[\widehat{\varphi} \mid \boldsymbol{Z}^S]$, both cross-fitted by K-fold splitting to keep the nuisance estimate independent of the residual at each unit: degree-2 polynomial (Ridge) regressions in the quadratic variant, gradient-boosted trees (Ke et al., 2017) in the LightGBM variant. In both, the Z^j-on-$\boldsymbol{Z}^S$ residualization is kept linear and computed in a single vectorized pass over all candidates. Provided $\mathbb{E}[Z^j \mid \boldsymbol{Z}^S]$ is well-approximated by its linear projection, $T_n(j \mid S)$ is asymptotically standard normal under the null and consistent against alternatives whose conditional-mean heterogeneity covaries linearly with Z^j given $\boldsymbol{Z}^S$. The quadratic variant adds curvature while keeping the model auditable; the LightGBM variant is the most flexible in the nuisances and the most expensive, since each test fits two boosting models. Neither variant removes the blind spot of the linear test: the GCM targets the expected conditional covariance of $\widehat{\varphi}$ and Z^j given $\boldsymbol{Z}^S$, which vanishes under $H_0(j \mid S)$ but also under any alternative whose conditional-mean contrast is uncorrelated with Z^j, e.g., a U-shape, and such alternatives are missed at any sample size.

(iii) Projected covariance measure (PCM). We use a variant of the PCM of Lundborg et al. (2024), which tests the conditional mean independence of an observed response from a covariate given a conditioning set. Our variant replaces the unobserved τ by the pseudo-outcome: since $\mathbb{E}[\widehat{\varphi} \mid \boldsymbol{Z}] = \mathbb{E}[\tau \mid \boldsymbol{Z}]$, $H_0(j \mid S)$ is equivalent to $\mathbb{E}[\widehat{\varphi} \mid \boldsymbol{Z}^{S\cup\{j\}}] = \mathbb{E}[\widehat{\varphi} \mid \boldsymbol{Z}^S]$, the PCM null with response $\widehat{\varphi}$, covariate Z^j and conditioning set $\boldsymbol{Z}^S$. The sample is split into two halves, with $\widehat{m}$ cross-fitted within each half. On the first half, a projection $\widehat{f}_j$ of the conditional-mean contrast $\mathbb{E}[\widehat{\varphi} \mid \boldsymbol{Z}^{S\cup\{j\}}] - \mathbb{E}[\widehat{\varphi} \mid \boldsymbol{Z}^S]$ is learned, without the conditional-variance weighting of the original, and rescaled to unit variance. On the second half, the residual product

$$R^{\mathrm{PCM}}(j \mid S) \;:=\; \big(\widehat{\varphi} - \widehat{\mathbb{E}}[\widehat{\varphi} \mid \boldsymbol{Z}^S]\big) \cdot \big(\widehat{f}_j - \widehat{\mathbb{E}}[\widehat{f}_j \mid \boldsymbol{Z}^S]\big) \tag{31}$$

is formed and studentized as in Equation 30. Because $\widehat{f}_j$ is learned on independent data, the statistic is asymptotically standard normal under $H_0(j \mid S)$ however poor the projection, provided the regressions on $\boldsymbol{Z}^S$ are accurate enough (Lundborg et al., 2024, Theorem 4). Under an alternative, R^{PCM} has positive mean whenever the residualized projection is positively correlated with the contrast, so the test is one-sided, with power tending to one against the alternatives where this correlation holds with high probability (Lundborg et al., 2024, Theorem 5), including the U-shapes invisible to the GCM. The price is sample splitting: the projection only sees half of the data, and the statistic is evaluated on the other half. The two halves then swap roles, and the two one-sided p-values are combined as $\min\{1, 2\min(p_1, p_2)\}$, a Bonferroni combination that is valid whatever their dependence; Lundborg et al. (2024) instead average the statistic over repeated splits. The quadratic variant fits $\widehat{f}_j$ by regressing $\widehat{\varphi}$ on $(Z^j, (Z^j)^2)$ after partialling out $[1, \boldsymbol{Z}^S]$, vectorized over all candidates, and residualizes it on $\boldsymbol{Z}^S$ linearly. The LightGBM variant additionally refits $\mathbb{E}[\widehat{\varphi} \mid Z^j, \boldsymbol{Z}^S]$ with gradient-boosted trees for the 32 candidates with the largest explained sum of squares on the first half, and residualizes these fits on $\boldsymbol{Z}^S$ with the same learner. Both variants keep quadratic nuisances for $\widehat{m}$ and $\widehat{\mathbb{E}}[\widehat{\varphi} \mid \boldsymbol{Z}^S]$.

Trade-offs and default. The three families trade power for the class of alternatives they see. Linear is fast, easy to implement, and the most efficient when the conditional CATE is approximately linear in $\boldsymbol{Z}$, as in our controlled benchmark; we adopt it as the default. The GCM is a compromise: more expressive than the linear test, since it is robust to misspecified nuisances, and more efficient than the PCM when the heterogeneity is monotone. However, it estimates $\mathbb{E}[\mathrm{Cov}(\widehat{\varphi}, Z^j \mid \boldsymbol{Z}^S)]$, which vanishes under $H_0(j \mid S)$ but also under any alternative where $\mathbb{E}[\tau \mid \boldsymbol{Z}^{S\cup\{j\}}]$ varies with Z^j while being uncorrelated with Z^j given $\boldsymbol{Z}^S$, e.g., a U-shape: against such alternatives it has no power at any sample size (Lundborg et al., 2024). The PCM targets conditional mean independence itself and is consistent against any alternative its projection can learn; the price is the sample split. On the CelebA benchmark, both assumption-lean families pay a measurable cost in power when the CATE is linear, and only the PCM recovers the modifiers of a U-shaped CATE (Figure 13, Table 3). We therefore use the linear test by default, the GCM when the heterogeneity may be nonlinear but

monotone, and the PCM when non-monotone heterogeneity cannot be ruled out and enough data are available. Switching test is a one-line change in the implementation of Appendix B.5; all other components of NEXIS are unchanged.

B.2 Error control

The forward step accepts the most significant candidate, provided its p-value passes a threshold, and that threshold decides how many candidates enter $\widetilde{S}$. The terminal backward step keeps its own level α/m whatever the forward correction (Appendix B.4); the interleaved backward step, when used, applies the forward correction against $|S|$. We consider three options.

(i) 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 and enter $\widetilde{S}$, leaving the terminal backward step to remove them. The semi-synthetic experiments (Figure 14) show that it does: precision reaches 0.95 almost as early as with the corrected variants, but recall comes later ($n = 1000$ instead of $n = 750$ at $\eta = 5$, and $\eta = 8$ instead of $\eta = 7$ at $n = 500$). The unadjusted gate is therefore not a recall-friendly choice in disguise: it delays recall.

(ii) FDR via Benjamini–Hochberg. Under its usual independence or positive-dependence conditions, the Benjamini–Hochberg procedure (Benjamin et al., 1995) controls the expected proportion of false discoveries among rejections at level α. At each forward step, candidate p-values are ordered $p_{(1)} \le \cdots \le p_{(|\bar{S}|)}$ and the largest k such that $p_{(k)} \le k\alpha/|\bar{S}|$ is identified; the candidates with p-value at most $p_{(k)}$ are eligible to enter S. If this rejection set is nonempty, NEXIS admits its most significant candidate. Although only one coordinate enters, this is not generally equivalent to the Bonferroni rule: several jointly small p-values can make $k > 1$ and admit a minimum above $\alpha/|\bar{S}|$. The interleaved backward step, when used, treats its batch of retained coordinates symmetrically. 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.

(iii) FWER via Bonferroni. Bonferroni (Bonferroni, 1936) controls the family-wise error rate by comparing each p-value against the conservative threshold $\alpha/|\bar{S}|$ (or $\alpha/|S|$ in the interleaved backward step). It is the most stringent of the three. The selection guarantee does not require a Bonferroni correction in the forward step: with fixed m, all three gates above admit every remaining principal proxy with probability tending to one by test consistency. Precision instead comes from the terminal level α/m, which corrects over the dictionary coordinates; no correction over all tested subsets is needed because the proof bounds each false survivor only at the fixed set $\mathcal{S}^\star$ (Appendix A.2).

Trade-offs and default. Empirically (Figure 14) FDR and FWER are essentially indistinguishable in our setting because the Benjamini–Hochberg rejection set is effectively a singleton at the leading discoveries. We therefore adopt FWER as the more conservative default. In larger truth-set regimes FDR may admit several jointly significant candidates and be the better operating point; 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 principal coordinate that leaves no residual modifier information in the others after conditioning on it. This is a population-level idealisation. Real learned dictionaries exhibit *partial* entanglement (Chanin et al., 2024): 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 terminal backward step always catching it, because experimental power amplifies finite-resolution misalignment in exactly the same way it amplifies genuine signals.

To guard against this residual signal, we add a relative-magnitude check to the existing p-value gate. For $S \neq \emptyset$, let

$$|T_n(j^\star \mid S)| \;\ge\; \rho \cdot \min_{j \in S} |T_n(j \mid S \setminus \{j\})|, \qquad \rho \in [0, 1], \tag{32}$$

where T_n is the statistic of the chosen CATE-equivalence test; the first candidate is governed only by the p-value gate. For $\rho > 0$, a later candidate $j^\star$ is retained only if its conditional test statistic is within a factor ρ^{-1} of the *weakest* coordinate already in S, while $\rho = 0$ disables the check. The intuition is residual-by-construction: principal modifiers with comparable detectability produce comparable standardized signals; a residual proxy of an already-selected principal should produce a weaker one, with the gap increasing as alignment improves. In this comparison the common $\sqrt{n}$ factor cancels, but the coordinate-specific standard-error estimates generally do not. Thus ρ compares standardized conditional signal strengths, which combine CATE contrast, coordinate scale, and residual variability; it is not in general a ratio of raw CATE contrasts. When the standard errors are comparable, the ratio approximately tracks relative contrast magnitudes. For example, $\rho = 0.5$ requires a new candidate to carry at least half the standardized evidence of the weakest selected coordinate.

Trade-offs and default. 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 $\rho = 0$ disables the gate, 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 steps

NEXIS uses two kinds of backward step. The *terminal backward step*, the last step of Algorithm 1, runs once, after the forward step has stopped at a selection $\widetilde{S}$: it keeps $j \in \widetilde{S}$ only if $p_j(A) \le \alpha/m$ for every $A \subseteq \widetilde{S} \setminus \{j\}$, the empty set included, and makes all removals simultaneously. The *interleaved backward step* is an optional variant: at every round, after the forward step, it re-tests each retained coordinate j against the rest of the current selection S, and removes it if $p_j(S \setminus \{j\}) > \alpha/|S|$.

(i) Terminal backward step. This step delivers the precision part of Theorem 4.1, without requiring test validity at data-dependent conditioning sets. On the recall event $\mathcal{S}^\star \subseteq \widetilde{S}$, the subsets tested for a coordinate $j' \in \widetilde{S} \setminus \mathcal{S}^\star$ include $A = \mathcal{S}^\star$. The null $H_0(j' \mid \mathcal{S}^\star)$ holds by Lemma A.1 and does not depend on the data, so Assumption 5(a) applies to it directly, and j' survives only if $p_{j'}(\mathcal{S}^\star) \le \alpha/m$. A union bound over the at most m coordinates outside $\mathcal{S}^\star$ bounds the probability that any of them survives by $\alpha + o(1)$. The price is the number of tests: at most $|\widetilde{S}| \cdot 2^{|\widetilde{S}|-1}$, exponential in $|\widetilde{S}|$ and determined by the size reached by the forward step rather than directly by m. This is feasible when the treatment effect has few direct modifiers, so that the forward step stops at a small $\widetilde{S}$. We motivate this practical regime by analogy with the sparse mechanism shift principle of causal representation learning (Schölkopf et al., 2021), which posits that small distribution changes tend to affect only a few factors of the causal factorization, not all of them simultaneously: we likewise expect the heterogeneity of a treatment effect to run through few direct modifiers. The measured sizes agree: over the 6,300 CelebA runs of the three dictionaries at $\rho = 0.5$, $|\widetilde{S}|$ has median 2 and maximum 7, and the forward step alone gives the same median and maximum on the main setting. Since a coordinate is dropped at its first failing subset, trying the largest subsets first, the same runs needed a median of 4 and a maximum of 192 tests, against the bound $7 \cdot 2^6 = 448$. Since the terminal backward step runs at level α/m, discarding the coordinates rarely active in the sample before running NEXIS, as in Appendices D and E, raises this level, and the guarantee applies to the retained coordinates, with m their number, because the screen uses only $\boldsymbol{Z}$, not the treatment or the outcome.

(ii) Interleaved backward step. The forward step followed by the terminal backward step mirrors the grow and shrink phases of IAMB (Tsamardinos et al., 2003b); interleaving a backward step after each forward addition mirrors Inter-IAMB (Tsamardinos et al., 2003b), and HITON interleaves in the same way (Aliferis et al., 2010). The interleaved step is not required asymptotically: recall follows from the forward step and precision from the terminal backward step. It can help in finite samples, when a coordinate that mixes several modifiers enters before their principal coordinates and becomes redundant once they are in. Removing it keeps $\widetilde{S}$ small, which cuts the cost of the terminal step and spares the principal coordinates the extra subset tests that include it. The interleaved step

alone, however, is not theoretically sufficient. It tests j at $S \setminus \{j\}$, where the current selection S is itself chosen by the data, so $H_0(j \mid S \setminus \{j\})$ is a data-dependent null: Assumption 5 covers fixed conditioning sets only, and the union bound behind the precision guarantee does not apply. This post-selection issue is what the terminal backward step resolves.

Trade-offs and default. In the backward-step ablation of Appendix C.5 (Figure 16), the terminal backward step raises precision for a small delay in recall, while the interleaved backward step changes no run: with two well-aligned modifiers on sparse SAE codes, no mixing coordinate enters ahead of the principals. The terminal step is therefore the default; we keep the interleaved step as an option for settings with more, or more entangled, modifiers.

B.5 Minimal Python Implementation Snippet

We provide a self-contained Python implementation of NEXIS as in Algorithm 1: the linear treatment-interaction test of Equation 27, the forward step with Bonferroni multiple-testing correction, and the terminal backward step, with no spectral-gap gate ($\rho = 0$) and no interleaved backward step. The implementation relies on `pandas` (McKinney et al., 2011) for tabular operations, `NumPy` (Harris et al., 2020) and `SciPy` (Virtanen et al., 2020) for numerical routines, and `statsmodels` (Seabold et al., 2010) for the OLS interaction t-test. Switching to a GCM or PCM variant of Appendix B.1, swapping Bonferroni for Benjamini–Hochberg, enabling the spectral-gap gate of Appendix B.3, or adding the interleaved backward step of Appendix B.4 is a localised modification of the routines below.

Neural Exposure Interaction Search. The forward step admits the most significant candidate until none passes the Bonferroni gate, and collects the selected coordinates in `S_candidate` (the $\widetilde{S}$ of Appendix B.4). The terminal backward step then filters `S_candidate` into the output `S`: it keeps each candidate only if it passes the gate α/m given every subset of the other candidates, largest subsets first, with all removals made simultaneously. The subset generator is lazy and `all` stops at the first failing subset, while `S_candidate` remains unchanged until every candidate is assessed; thus the implementation matches Algorithm 1 and uses at most $|\widetilde{S}|2^{|\widetilde{S}|-1}$ terminal tests. The procedure is parameterized by a generic CATE equivalence test `cci_test`, which returns a p-value per candidate.

```
from itertools import combinations

def NEXIS(T, Y, Z, alpha=0.05, cci_test=cci_test_linear):
    m, S_candidate = Z.shape[1], []
    def p(A, j):  # p-value of H0(j | A)
        return cci_test(T, Y, Z, A, [j]).loc[j, "p_value"]
    # forward step with Bonferroni gate
    while len(S_candidate) < m:
        bar_S = [c for c in Z.columns if c not in S_candidate]
        tests = cci_test(T, Y, Z, S_candidate, candidates=bar_S)
        j_star = tests["p_value"].idxmin()
        if tests.loc[j_star, "p_value"] > alpha / len(bar_S):
            break
        S_candidate.append(j_star)
    # terminal backward step over every subset A of S_candidate\{j}
    S = []
    for j in S_candidate:
        rest = [c for c in S_candidate if c != j]
        subsets = (list(A) for r in range(len(rest), -1, -1)
                   for A in combinations(rest, r))
        if all(p(A, j) <= alpha / m for A in subsets):
            S.append(j)
    return S
```

Linear treatment-interaction test (default). Fits Equation 27 via OLS and tests the candidate's treatment interaction δ_j via a t-test.

```python
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()
        t_test = model.t_test(f"_T_x_{j} = 0")
        out.append((j, float(t_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 (Liu et al., 2018) 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 the 19,867 images of the official CelebA validation split (aligned and cropped) with **SigLIP 2** (Tschannen et al., 2025), a contrastively pretrained ViT-B/16 (timm `vit_base_patch16_siglip_224`, weights `v2_webli`), at 224 px, taking the 196 final-layer patch tokens (768-dimensional) and not using the attention-pooling head. These images are used for SAE training, for the ground-truth labels below, and as the pool from which the benchmark samples its units. 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** (Gao et al., 2024) of hidden width $m = 9{,}216$ ($= 12 \times 768$) on the individual SigLIP 2 patch tokens of the same images ($19{,}867 \times 196 \approx 3.89 \times 10^6$ training vectors per epoch), following the protocol of Mencattini et al. (2025): a linear encoder followed by ReLU and TopK, unit-norm decoder rows, mean-squared reconstruction loss, Adam with learning rate 5×10^{-4} and gradient-norm clipping at 1, 20 epochs, batches of 20 images, initialization seed 0. The SAE is trained without labels, treatments or outcomes, but on the same images the benchmark samples from; we check reproducibility with the $k = 20$ SAE retrained on an independent sample of 19,867 training-split images that shares no identities with the validation split (Appendix C.2). Two sparsity levels are evaluated, $k = 5$ and $k = 20$ active features. Each image is encoded by applying the trained encoder to the mean $\bar{\boldsymbol{h}}$ of its 196 patch tokens, and for each level we expose two views of the representation:

- $\boldsymbol{Z} = \mathrm{TopK}(\mathrm{ReLU}(\boldsymbol{W}_{\mathrm{enc}}\bar{\boldsymbol{h}} + \boldsymbol{b}_{\mathrm{enc}}))$, the sparse *post*-TopK codes (near-orthogonal, $L_0 \le k$);
- $\boldsymbol{Z}_{\mathrm{pre}} = \boldsymbol{W}_{\mathrm{enc}}\bar{\boldsymbol{h}} + \boldsymbol{b}_{\mathrm{enc}}$, the dense, continuous linear *pre*-activations before ReLU and TopK (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.4.

Data-generating process. We fix two CelebA attributes, $W_1 =$WEARING_HAT ($\sim 4.7\%$ prevalence) and $W_2 =$EYEGLASSES ($\sim 7.0\%$ 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:

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

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 19,867 images of the pool. 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 19,867

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 $\mathcal{S}^\star$.

Baselines. Classic CATE methods, e.g., forests (Wager & Athey, 2018), meta-learners (Künzel et al., 2019), causal trees (Athey & Imbens, 2016) and rule ensembles (Bargagli-Stoffi et al., 2020), do not target the characterization sought here: they return pointwise effects, partitions or subgroups, not a set of coordinates. The task-aligned baseline is therefore marginal interaction screening, the standard multiple-testing pipeline for this question. 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)* (Benjamin et al., 1995), 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 in the forward step, spectral gap $\rho = 0.5$, and the forward step followed by the terminal backward step. These are the defaults recommended in Appendix B. The GCM and PCM nuisances are cross-fitted with 3 folds, and their LightGBM variants use 50 trees of depth at most 4 with at most 15 leaves.

What we test. Appendix C.2 reports the main DGP (Figure 6) and its reproduction with an independently trained SAE (Figure 7). Appendix C.3 tests Principal Alignment with supervision, and NEXIS under a controlled violation of it (Figure 10). Appendices C.4 and C.5 ablate the dictionary (sparsity k, $\boldsymbol{Z}$ vs. $\boldsymbol{Z}_{\mathrm{pre}}$) and NEXIS (test, correction, ρ, backward steps); the test ablation adds a non-monotone U-shape DGP (Table 3). Appendix C.6 varies the number of direct modifiers, $r \in \{0, 1, 3\}$. Every DGP includes prognostic main effects of the attributes, $\beta_{W_1} = 0.3$ and $\beta_{W_2} = -0.2$ in the main setting. The CATE-equivalence test targets the treatment interaction, so its null holds for a coordinate that carries only prognostic signal: at $r = 1$, EYEGLASSES keeps only its prognostic effect and NEXIS never selects its coordinate, and at $r = 0$ the share of runs with any false discovery is at most 0.015 at every n.

C.2 MAIN RESULTS

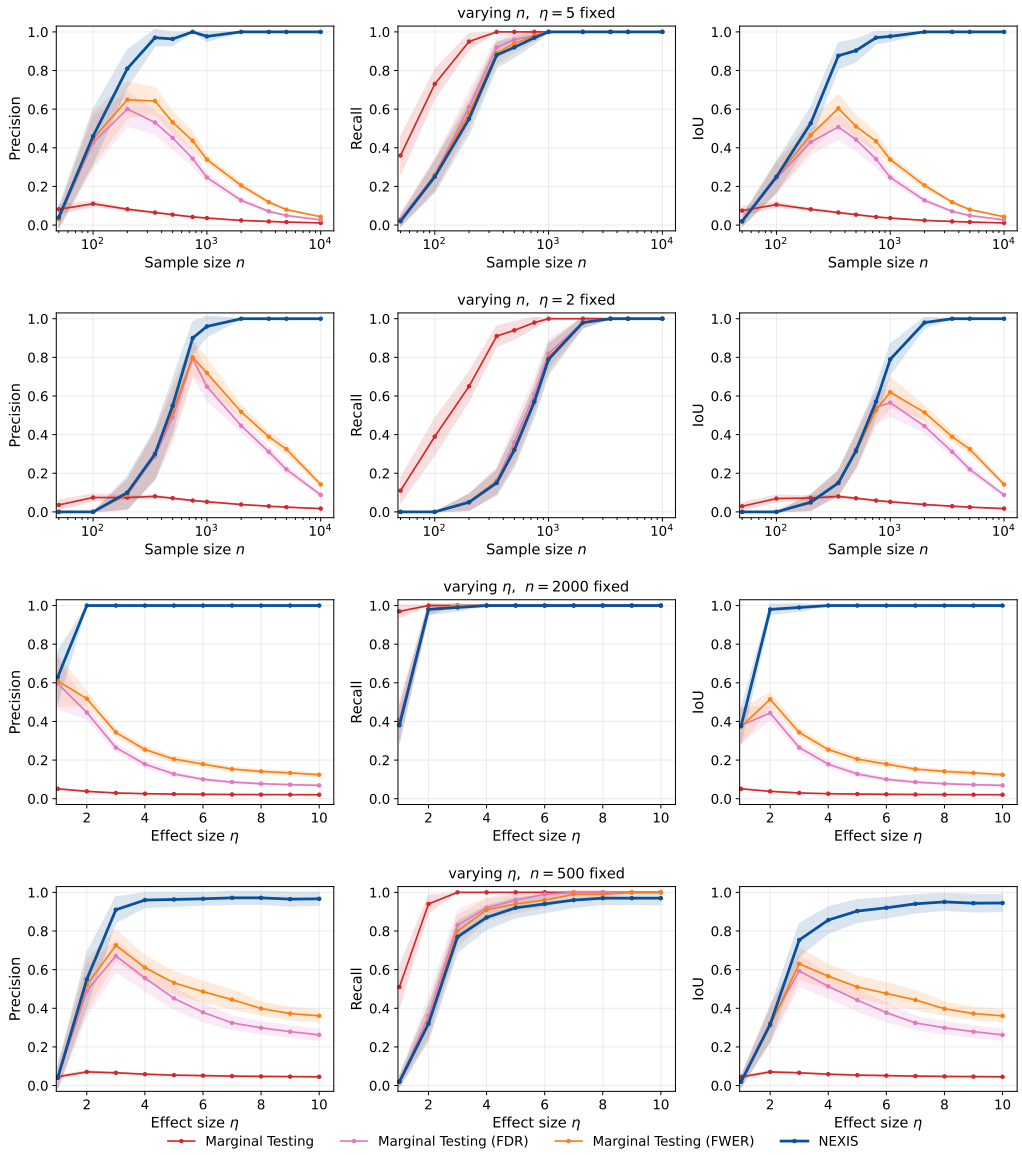


Figure 6: **Reference figure.** NEXIS vs. marginal baselines on the main setting ($k = 20$, $\boldsymbol{Z}$, FWER, $\rho = 0.5$, linear test, terminal backward step). Rows traverse the four DGP conditions; columns show precision, recall, IoU. NEXIS achieves both high recall and high precision in the high-power regime ($\eta \geq 2$ at $n = 2000$, or $n \geq 750$ 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.

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 $\eta = 7$ when $n = 500$. Precision reaches 0.95 before recall ($n = 350$ at $\eta = 5$, $\eta = 4$ at $n = 500$), then stays above it. 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 recall onset from $n = 750$ to $n = 2000$, with no qualitative change in shape. Comparing the $n = 2000$ and $n = 500$ rows, NEXIS shifts its recall onset from $\eta = 2$ to $\eta = 7$ and saturates at a slightly lower precision plateau (~ 0.97 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.

Reproducibility. The SAE of the main setting is trained on the images the benchmark samples from. To check that the results do not depend on this, we retrain the $k = 20$ TopK SAE with the same recipe on an independent sample of $19{,}867$ images of the CelebA training split, which shares no identity with the validation split, and rerun the full main grid unchanged, with the same DGP, seeds and baselines; $\mathcal{S}^\star$ is again the best-F_1 coordinate of each attribute in the new dictionary. Figure 7 shows that, even when the concepts are not perfectly disentangled, NEXIS still identifies them. Recall is essentially unchanged (macro recall 0.729 on the replica against 0.750 on the main dictionary, averaged over the 42 cells of the figure), while macro precision and IoU drop from 0.808 and 0.745 to 0.745 and 0.680, and the IoU at $n = 2000$, $\eta = 5$ from 1.000 to 0.920. NEXIS keeps dominating the marginal baselines, whose best macro precision on the replica is 0.312. The remaining precision gap at large n comes from coordinates of the replica dictionary that carry residual signal of the same concepts and count as false discoveries against the coordinate-level $\mathcal{S}^\star$.

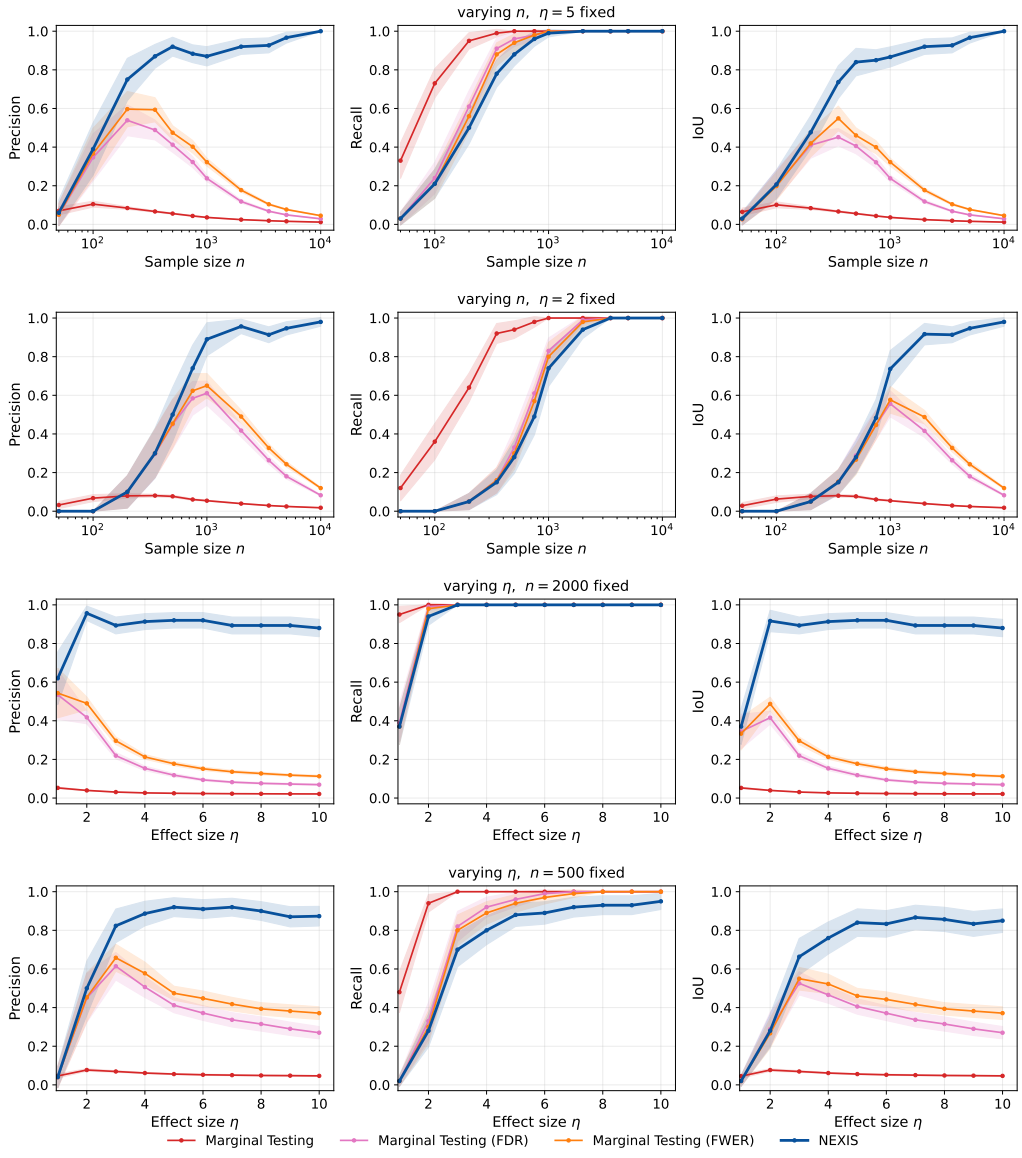


Figure 7: **Reproducibility: independently trained SAE.** Same layout as Figure 6, with the $k = 20$ TopK SAE retrained with the same recipe on 19,867 CelebA training-split images that share no identity with the validation split; same DGP, seeds and baselines. Recall matches the main dictionary and NEXIS dominates the marginal baselines; precision saturates later because a few coordinates retain residual signal of the same concepts. Reference: Figure 6.

C.3 Testing Principal Alignment with supervision

Principal Alignment (Assumption 3) cannot be checked without supervision, but in the CelebA benchmark the direct modifiers are observed attributes, so we test it on the main dictionary (TopK SAE, $k = 20$, $m = 9{,}216$, codes $\boldsymbol{Z}$ of the $19{,}867$ pool images).

Marginal alignment. Each principal coordinate is the most aligned coordinate by a wide margin: its activation separates WEARING_HAT with AUC 0.969 and EYEGLASSES with AUC 0.956, against 0.755 and 0.714 for the runner-up among the other $9{,}215$ coordinates (Figure 8), and 90% and 95% of its 100 most activating images carry the attribute, against prevalences of 4.7% and 7.0% (Figure 9). The alignment is not monosemantic: 9 and 11 other coordinates still reach AUC 0.6, and some narrower coordinates reach a higher top-100 purity while covering only a small part of the attribute.

Screening-off. To test the screening-off of Equation 6 itself, we compare held-out logistic probes of W^k on Z^{j_k} alone (a linear term plus indicators at the deciles of its positive values, to allow a nonlinear link) with probes that also include the $K \in \{10, 50, 200\}$ other coordinates most associated with W^k given Z^{j_k}, selected on a disjoint half of the images (Table 2, 20 random splits). The probes are ℓ_2-penalized ($C = 1$) and fitted on one half; AUCs are computed on the other half, on which the likelihood-ratio test compares unpenalized fits with and without the K extra coordinates. The AUC rises only from 0.970 to at most 0.988 for WEARING_HAT and from 0.956 to at most 0.989 for EYEGLASSES, and the principal coordinate alone carries 94–96% of the held-out information the augmented probe extracts. With about $10{,}000$ held-out images, a likelihood-ratio test of the K extra coordinates still rejects exact conditional independence ($p < 10^{-14}$ in every split), so Principal Alignment holds only approximately, as a population-level idealization. This residual leakage does not affect recovery: in the same setting NEXIS returns exactly $\mathcal{S}^\star = \{5348, 5537\}$ in all 50 runs at $n = 2000$ and $\eta = 5$, with mean IoU 0.999 over $\eta \in \{3, \dots, 10\}$ at $n = 2000$.

Table 2: **Screening-off test of Principal Alignment.** Held-out AUC of logistic probes of W^k on the principal coordinate Z^{j_k} alone and on Z^{j_k} plus the K other coordinates most associated with W^k given Z^{j_k}, selected on the training half; mean over 20 random 50/50 splits, with the 2.5%–97.5% range over splits for ΔAUC. Info. share: share of the held-out log-loss reduction of the augmented probe that the principal coordinate alone achieves.

Modifier	K	AUC Z^{j_k}	AUC Z^{j_k} + K others	ΔAUC	info. share
WEARING_HAT	10	0.970	0.981	+0.011 [+0.007, +0.014]	0.96
	50		0.985	+0.015 [+0.010, +0.019]	0.95
	200		0.988	+0.018 [+0.009, +0.024]	0.95
EYEGLASSES	10	0.956	0.983	+0.026 [+0.020, +0.030]	0.95
	50		0.989	+0.033 [+0.028, +0.036]	0.94
	200		0.989	+0.033 [+0.029, +0.037]	0.94

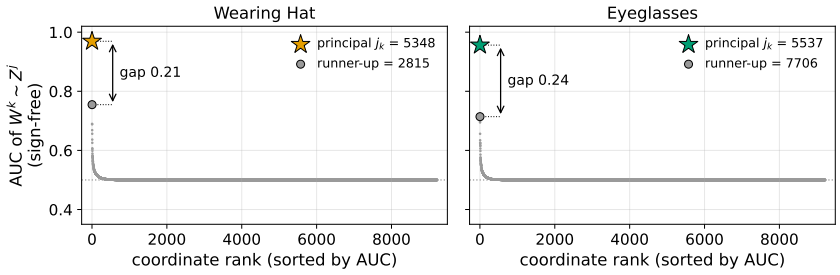


Figure 8: **Alignment spectrum.** Each point is one coordinate Z^j of $\boldsymbol{Z}$, with the sign-free AUC, $\max(\mathrm{AUC}, 1-\mathrm{AUC})$, of Z^j as a score for W^k, WEARING_HAT (left) and EYEGLASSES (right), over the $19{,}867$ pool images; coordinates are sorted by AUC. The principal coordinate (star) leads the runner-up (dot) by 0.21 and 0.24.

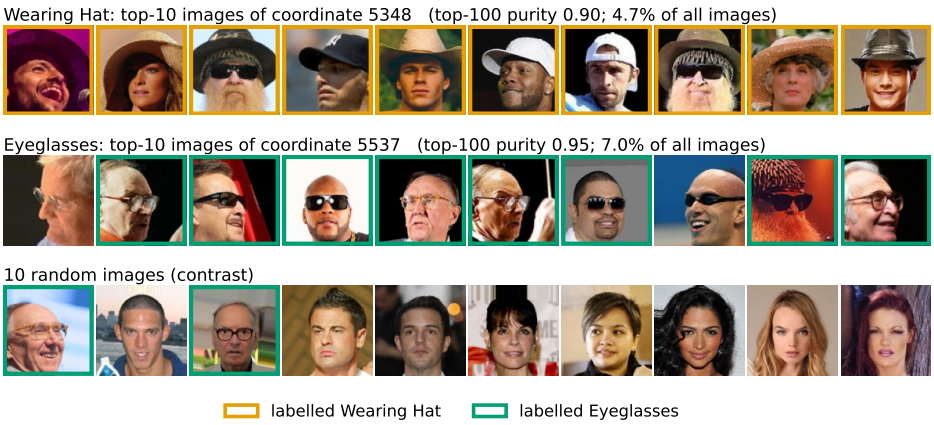


Figure 9: **Top-activating images of the principal coordinates.** The 10 most activating images of coordinate 5348 (WEARING_HAT, top row) and 5537 (EYEGLASSES, middle row), and 10 random images for contrast (bottom row). Framed images carry the attribute.

Controlled violation. To see what happens when Principal Alignment fails, we split the WEARING_HAT principal coordinate j_1 in two: for each pool image we draw $U \sim \text{Bernoulli}(0.5)$ once, set $Z^{j_1} \leftarrow UZ^{j_1}$, and append $Z^{j_{\text{new}}} = (1-U)Z^{j_1}$. The two halves sum to the original coordinate, but neither screens off the modifier alone, so no map as in Assumption 3 exists and the mean Markov boundary of the CATE grows from $\{j_1, j_2\} = \{5348, 5537\}$ to $\{j_1, j_{\text{new}}, j_2\}$. On the draws of the main setting at $\eta = 5$ (Figure 10), NEXIS selects both halves in 48 of 50 runs at $n = 2000$ and returns exactly $\{j_1, j_{\text{new}}, j_2\}$ in all 50 runs at $n = 10^4$, whereas Marginal Testing (FWER) keeps the three target coordinates but selects 47.9 coordinates on average (precision 0.06). The violation costs power, since each half is active on only about half of the images that activate the original coordinate: NEXIS needs $n = 3500$ instead of $n = 750$ to reach a mean IoU of 0.95, and in the other 2 runs at $n = 2000$ the spectral-gap gate stops the forward step before either half enters. Theorem 4.1 assumes Principal Alignment; in this experiment, where it fails, NEXIS still recovers the larger Markov boundary once n is large enough.

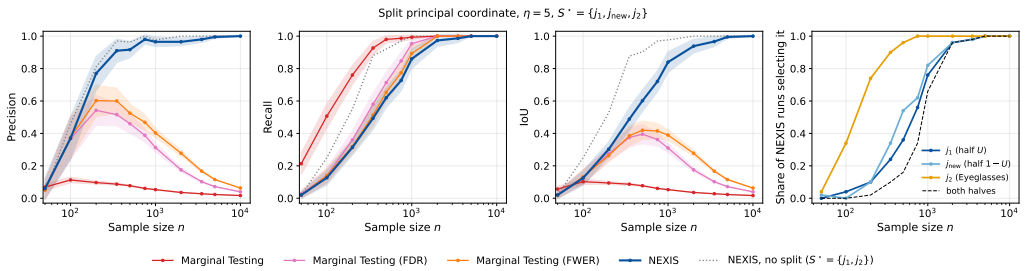


Figure 10: **Controlled violation of Principal Alignment.** The WEARING_HAT principal coordinate is split into two complementary halves, so $\mathcal{S}^\star = \{j_1, j_{\text{new}}, j_2\}$; $\eta = 5$. Left: precision, recall and IoU against n; dotted: NEXIS on the unsplit dictionary. Right: share of NEXIS runs selecting each target coordinate.

C.4 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 $\boldsymbol{Z}_{\text{pre}}$. Both deviate from the main setting along a single axis and are read against Figure 6.

Sparsity level: $k = 5$ vs. $k = 20$. Figure 11 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 $n = 350$ 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.

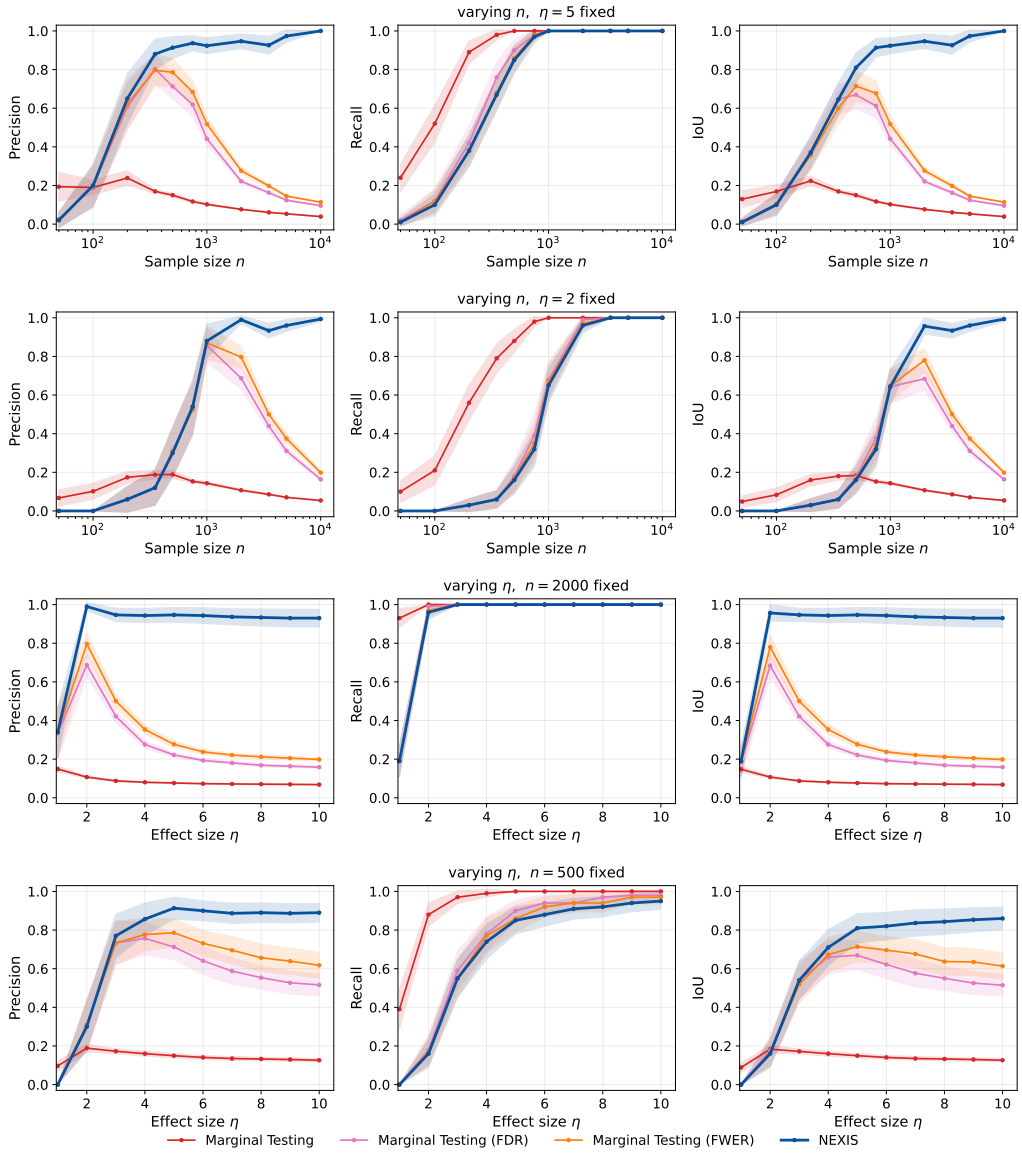


Figure 11: **Ablation: SAE sparsity level $k = 5$.** Same layout as Figure 6. Recall thresholds match the main setting; precision saturates more slowly because the lower-k dictionary spreads attribute information across more coordinates. Reference: Figure 6.

Feature view: Z vs. Z_{pre}. Figure 12 repeats the reference comparison using the dense pre-activations Z_{pre} at $k = 20$. Recall crosses 0.95 at the same point ($n = 750$ at $\eta = 5$); precision saturates at the same plateau but requires modestly more data ($n = 500$ at $\eta = 5$, vs. $n = 350$ on Z). This is consistent with the geometry: the post-TopK codes Z are near-orthogonal, so conditioning on the already-selected set introduces little noise and the Bonferroni gate remains tight; Z_{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 Z are the preferable input.

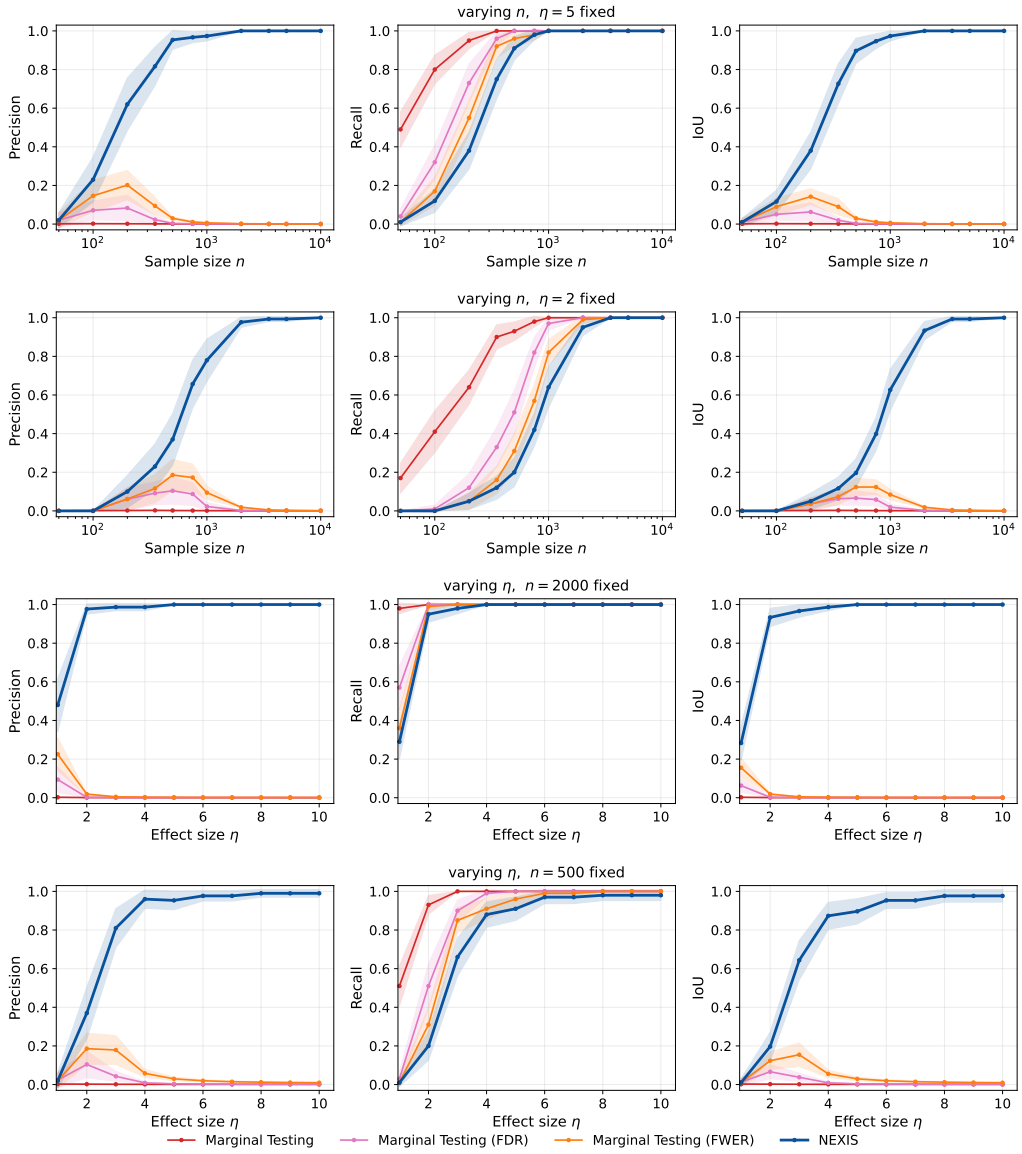


Figure 12: **Ablation: pre-TopK activations Z_{pre}.** Same layout as Figure 6, with the SAE feature view replaced by the dense pre-activations Z_{pre} at $k = 20$. On this dense view every coordinate is a linear projection of the same embedding and inherits marginal heterogeneity, so the FWER baseline selects about 600 coordinates on average and its precision collapses; the precision and IoU of NEXIS shift slightly to the right relative to the sparse-code reference, and the qualitative ordering is unchanged. Reference: Figure 6.

C.5 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 13 compares the linear treatment-interaction test (default, Equation 27) to the doubly-robust GCM and to the projected covariance measure (PCM) of Lundborg et al. (2024) (Appendix B.1), each with a quadratic and a LightGBM variant (of the nuisance regressions for the GCM, of the projection for the PCM). As expected, the linear test is the most sample-efficient: it crosses recall 0.95 at $n = 750$ when $\eta = 5$ and at $\eta = 2$ when $n = 2000$. Both GCM variants need $n = 2000$ at $\eta = 5$ and $\eta = 4$ at $n = 2000$. Both PCM variants need $n = 3500$ at $\eta = 5$ and never reach recall 0.95 at $n = 2000$ on the grid $\eta \leq 10$ (recall about 0.6 at $n = 2000$, $\eta = 5$): the PCM pays for the sample split, since in each direction the statistic is evaluated on $n/2$ observations, which scales it by about $1/\sqrt{2}$. At $n = 500$, the recall of all four stays at or below 0.04 up to $\eta = 10$. The linearity the test needs is that of the conditional CATE $\mathbb{E}[\tau \mid \boldsymbol{Z}^{S\cup\{j\}}]$ in the coordinates, and here it holds approximately: τ in Equation 33 is by construction *linear* in the binary attributes W_1, W_2, and the principal coordinates Z^{j_k} act close to indicators of these attributes. Linearity is a property of the data-generating process; a dictionary can at most expose it, when its aligned coordinates are near-indicators of the modifiers, but cannot create it. When τ is linear in the modifiers and the dictionary aligns them in this way, the linear test is well specified. The GCM and PCM variants pay a power penalty for a nonparametric flexibility that this DGP does not reward, and the nuisance (or projection) model barely matters: the quadratic and LightGBM curves of each test nearly coincide. The two GCM variants even select the same sets in every cell: the outcome-side nuisances enter only once $S \neq \emptyset$, where $\boldsymbol{Z}^S$ is one or two sparse coordinates that are zero on about 90% of the samples, so both fits give nearly identical statistics that never fall on opposite sides of the gate.

The split buys power against non-monotone heterogeneity. Table 3 adds a U-shape DGP (panel B) that keeps the prognostic effects of Equation 33 but sets $\tau = \tau_0 + \eta\left(g_1(Z^{5348}_{\mathrm{pre}}) - g_2(Z^{5537}_{\mathrm{pre}})\right)$, where $g_k(z)$ is the square of the standardized pre-activation, residualized on $(1, z)$ over the pool and scaled to unit variance, so that τ varies with both principal coordinates while having zero covariance with each. The linear test and both GCM variants are blind to it: their recall never exceeds 0.10 on this grid and is zero from $n = 2000$ on. The PCM recovers both modifiers: its IoU reaches 0.95 at $n = 1000$ with the LightGBM projection (1.00 at $n = 2000$), and at $n = 3500$ with the quadratic one (0.77 at $n = 2000$). The choice of test is thus a trade-off between power and the class of alternatives it sees. The linear test is the most efficient when the conditional CATE is linear in the coordinates. The GCM is a compromise: more expressive than the linear test, since it is robust to misspecified nuisances, and more efficient than the PCM when the heterogeneity is monotone, but blind to non-monotone heterogeneity at every n. The PCM sees non-monotone heterogeneity at the cost of the split. We therefore use the linear test by default, the GCM when the heterogeneity may be nonlinear but monotone, and the PCM when non-monotone heterogeneity cannot be ruled out and enough data are available.

Table 3: **CATE equivalence test under a linear and a non-monotone CATE.** $n^\star$: smallest n on the grid with mean IoU at least 0.95 at $\eta = 5$ (–: never); IoU: mean IoU at $n = 2000$, $\eta = 5$. A: main setting, τ linear in WEARING_HAT and EYEGLASSES, codes $\boldsymbol{Z}$. B: U-shape DGP, pre-activations $\boldsymbol{Z}_{\mathrm{pre}}$. 50 seeds per cell, default NEXIS otherwise.

	A. linear CATE		B. U-shape	
Test	$n^\star$	IoU	$n^\star$	IoU
linear (default)	750	1.00	–	0.00
GCM: quadratic	2000	0.98	–	0.00
GCM: LightGBM	2000	0.98	–	0.00
PCM: quadratic	3500	0.60	3500	0.77
PCM: LightGBM	3500	0.59	1000	1.00

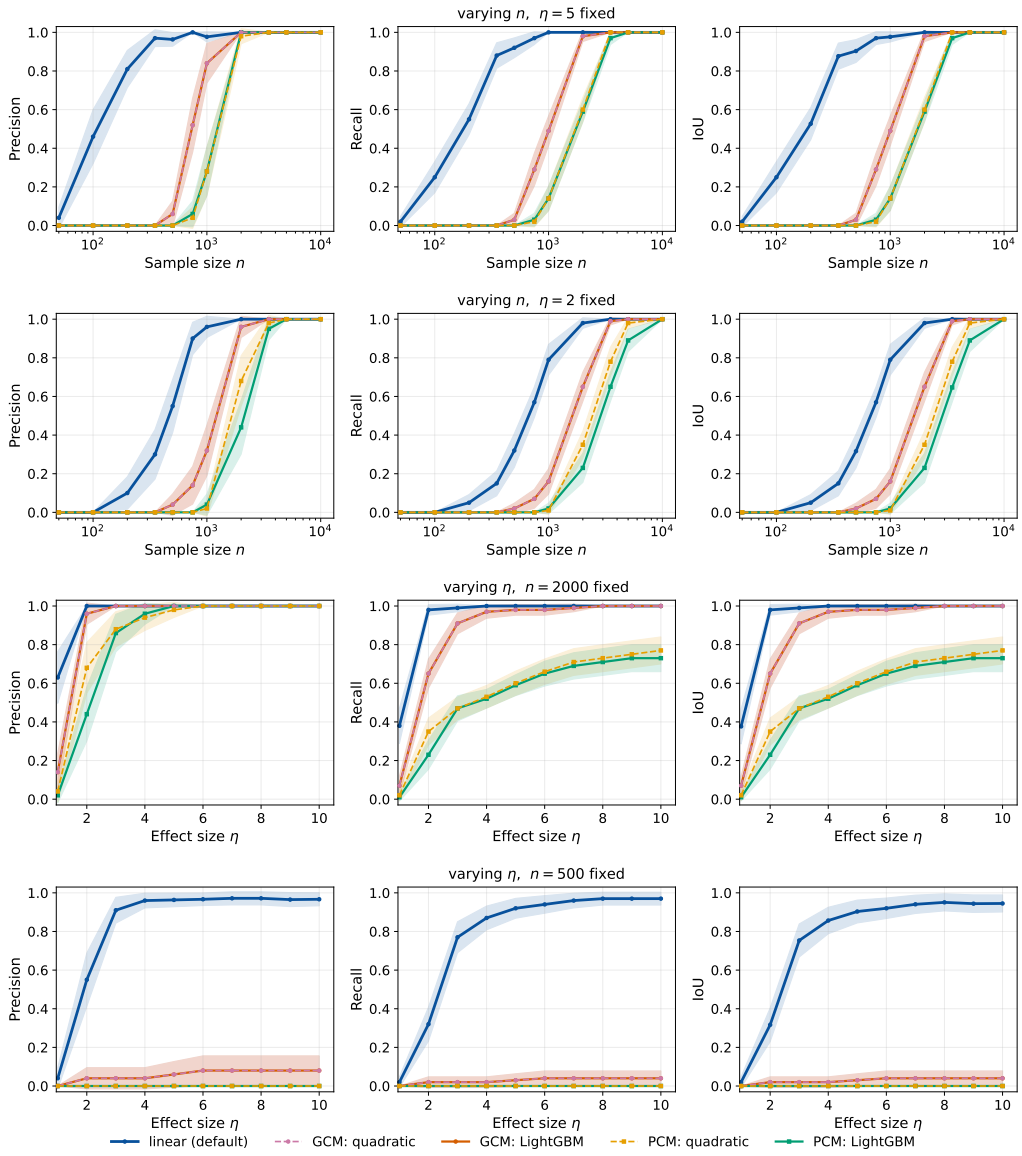


Figure 13: **Ablation: CATE equivalence test.** Linear (default), GCM and PCM, each with a quadratic or a LightGBM variant (of the nuisance regressions for the GCM, of the projection for the PCM), on the main setting. The linear test is close to well specified, since τ in Equation 33 is linear in the attributes and the principal coordinates act as near-indicators of them, and dominates in this regime. Reference: Figure 6.

Multiple-testing correction. Figure 14 compares the FWER (Bonferroni, default) gate of the forward step to the FDR (Benjamini–Hochberg) gate and to the unadjusted run; the terminal backward step keeps its level α/m throughout. With no correction, the terminal backward step removes the extra coordinates admitted by the forward step, so precision reaches 0.95 almost as early as with FWER ($\eta = 2$ at $n = 2000$; $n = 500$ instead of $n = 350$ at $\eta = 5$), but recall comes later: $n = 1000$ instead of $n = 750$ at $\eta = 5$, and $\eta = 8$ instead of $\eta = 7$ at $n = 500$. FDR and FWER are essentially indistinguishable here because the Benjamini–Hochberg rejection sets at the leading discoveries are effectively singletons, so the two gates coincide. We keep FWER as the more conservative default; in larger truth-set regimes FDR may be preferable.

Spectral gap ρ. Figure 15 sweeps $\rho \in \{0, 0.2, 0.5, 0.8\}$, where $\rho = 0$ disables the gap entirely (Equation 32 of Appendix B.3). The lower-ρ runs (0 and 0.2) admit correlated companions of the

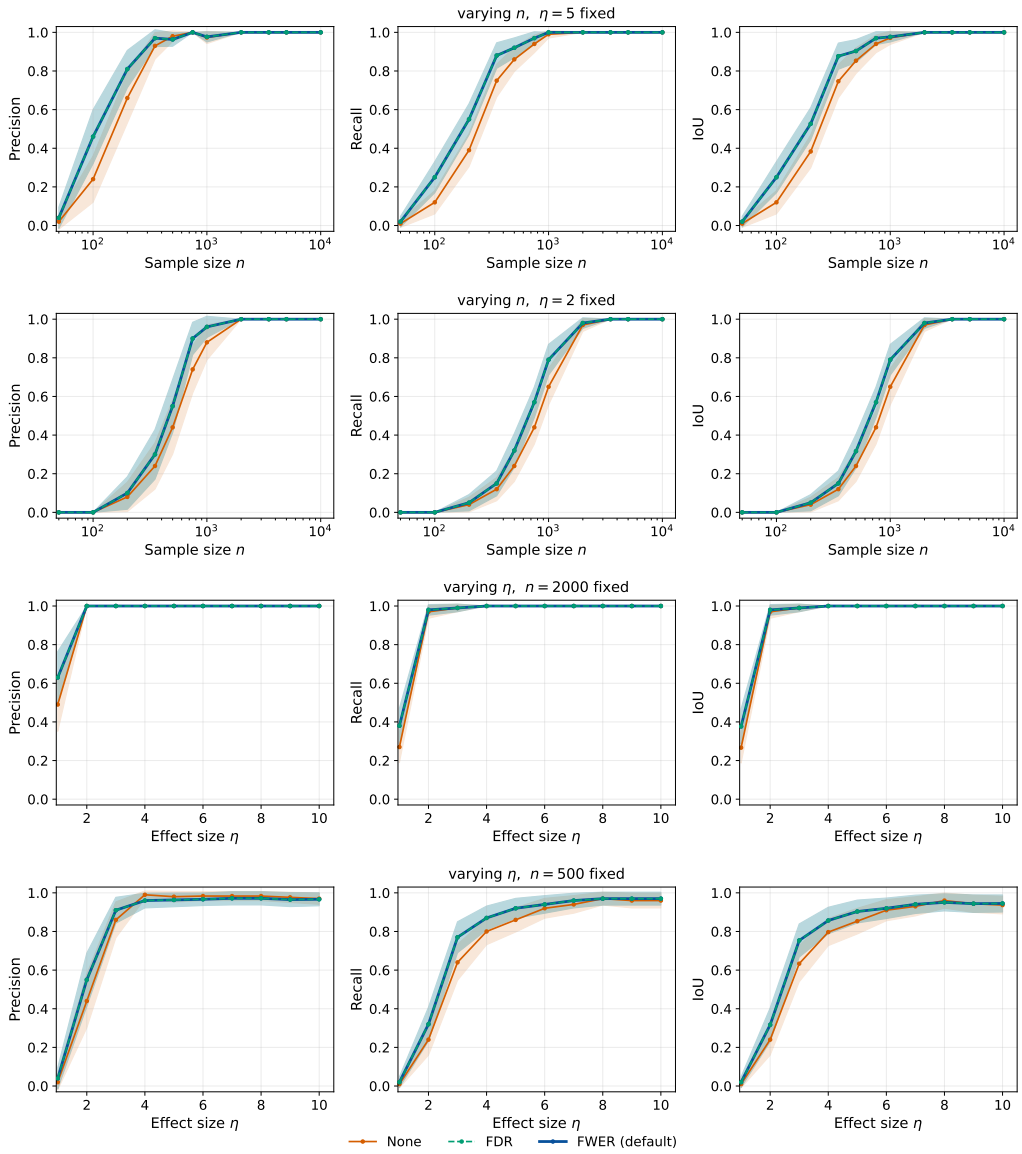


Figure 14: **Ablation: multiple-testing correction.** None, FDR (Benjamini–Hochberg), and FWER (Bonferroni, default), on the main setting. FWER and FDR are equivalent at this truth-set size; the unadjusted gate mainly delays recall. Reference: Figure 6.

principal coordinates as soon as the conditional gate is loose enough, hurting precision in the large-η and large-n regimes for at most a marginal recall gain. 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 $n = 2000$ row. The default $\rho = 0.5$ gives the best trade-off in standardized conditional signal strength (Appendix B.3).

Backward elimination. Figure 16 compares the forward step alone and the default (forward step followed by the terminal backward step), each with and without the interleaved backward step. The terminal backward step removes the residual companions left by the forward step: precision reaches 0.95 at $n = 350$ instead of $n = 2000$ at $\eta = 5$, and stays at or above 0.96 at $n = 500$ for $\eta \geq 4$ (forward step alone: 0.72 at $\eta = 10$), for a small delay in recall ($n = 750$ instead of $n = 500$ at $\eta = 5$; $\eta = 7$ instead of $\eta = 5$ at $n = 500$). The interleaved backward step changes no run. It matters

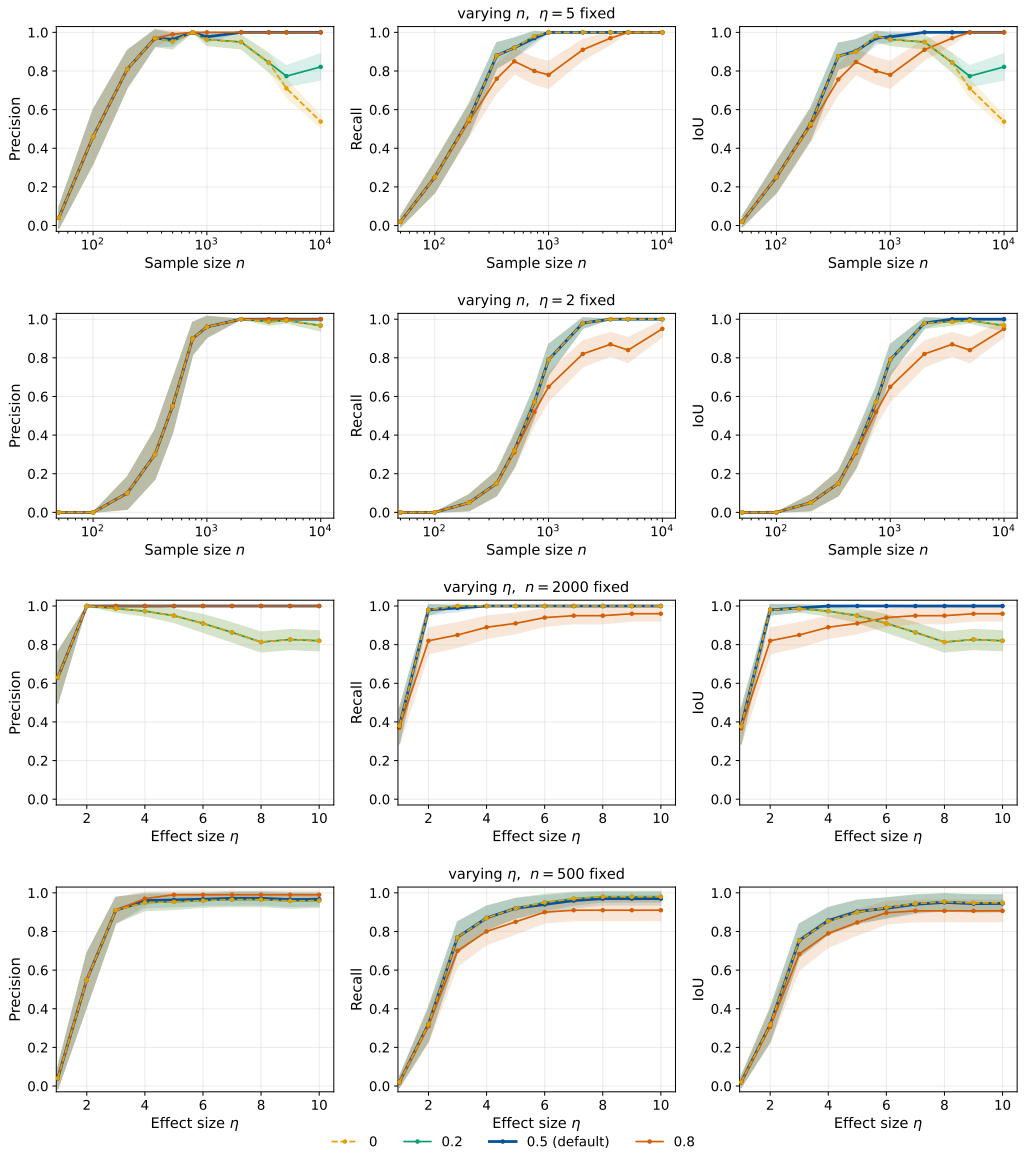


Figure 15: **Ablation: spectral gap ρ.** $\rho \in \{0, 0.2, 0.5, 0.8\}$, on the main setting. Low ρ admits correlated companions and erodes precision; high ρ blocks weaker principal coordinates and erodes recall. The default $\rho = 0.5$ jointly attains both. Reference: Figure 6.

only when a coordinate that mixes several modifiers enters before their principal coordinates and becomes redundant once they are in, which two well-aligned modifiers on sparse SAE codes never produce; we keep it as a variant for more entangled settings.

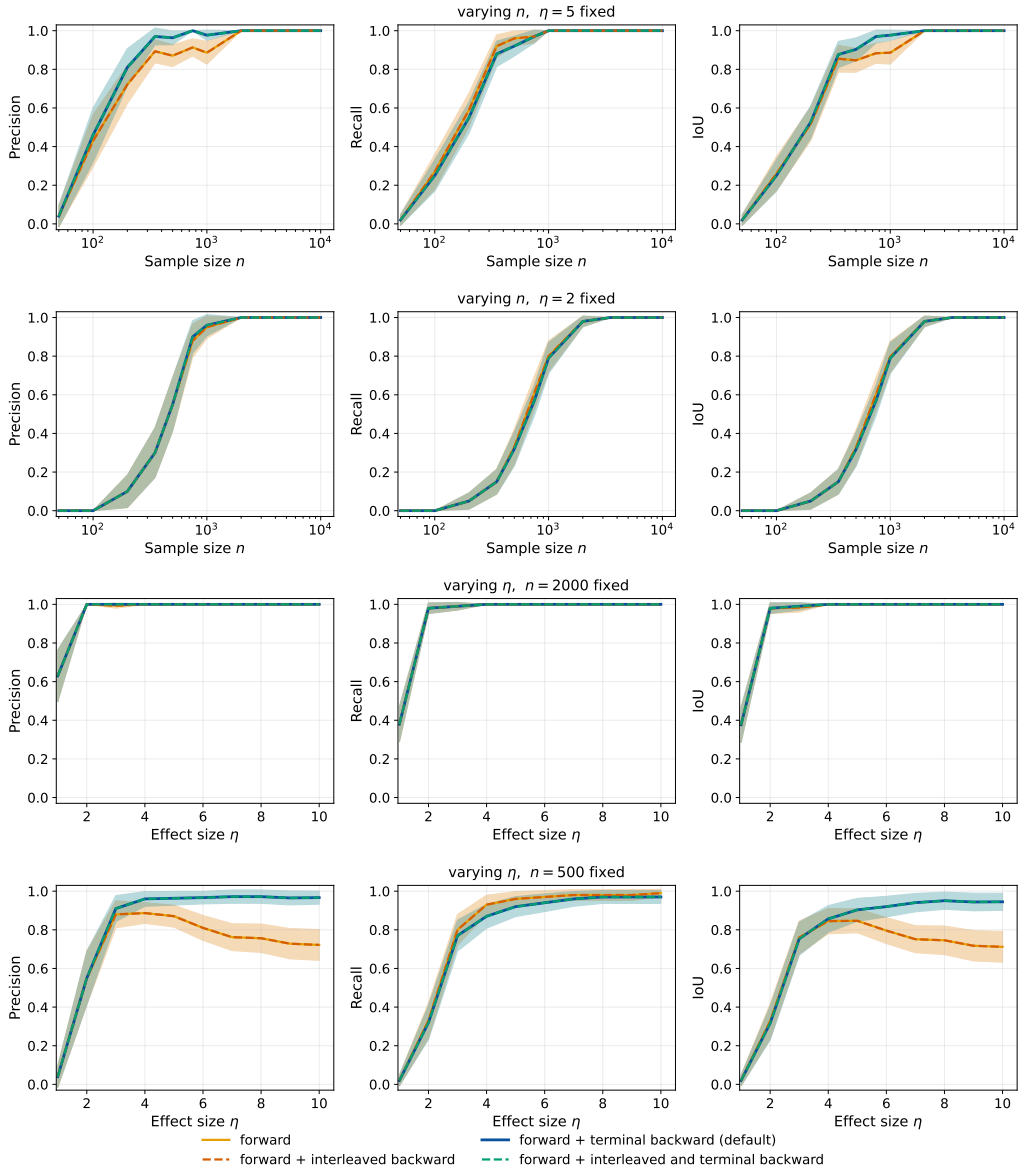


Figure 16: **Ablation: backward steps.** Forward step alone, with the interleaved backward step, with the terminal backward step (default), and with both, on the main setting. The terminal backward step raises precision for a small delay in recall; the interleaved backward step changes no run at $|\mathcal{S}^\star| = 2$ on a near-orthogonal $\boldsymbol{Z}$, so the dashed curves overlap the solid ones. Reference: Figure 6.

C.6 ADDITIONAL DATA GENERATING PROCESS ABLATIONS

We now hold the dictionary and NEXIS at the main setting and vary the number r of direct modifiers, with the same sweeps as in Appendix C.2. At $r = 1$, only WEARING_HAT modifies the effect: EYEGLASSES is still sampled with its prognostic effect β_{W_2}, but with no interaction, and $\mathcal{S}^\star = \{5348\}$. At $r = 3$, SIDEBURNS joins the two modifiers of Equation 33, with $(\gamma_{W_1}, \gamma_{W_2}, \gamma_{W_3}) = (+1, -1, +1)$ and $(\beta_{W_1}, \beta_{W_2}, \beta_{W_3}) = (0.3, -0.2, 0.3)$, the three attributes sampled independently at their CelebA prevalence, and $\mathcal{S}^\star = \{5348, 5537, 1683\}$. At $r = 0$, there is no effect modification: $\tau = \tau_0$ for every unit, the prognostic effects are those of the main setting, and $\mathcal{S}^\star = \emptyset$. Macro metrics average over the 42 cells of a figure.

No modifier ($r = 0$). With $\gamma = 0$, η multiplies nothing, and a larger constant effect τ_0 leaves every interaction statistic unchanged, so the four DGP conditions collapse to one; we sweep n with 200 seeds per value. Any selected coordinate is then a false discovery. Table 4 reports the share of runs with at least one. For NEXIS it is at most 0.015 at every n, and 0.003 pooled over the $2{,}200$ runs, below $\alpha = 0.05$, while uncorrected marginal testing makes a false discovery in 99.5% to 100% of the runs. At $r = 0$, NEXIS makes a false discovery in exactly the same runs as the Bonferroni marginal baseline, since with no modifier its first forward step is that test.

Table 4: **DGP ablation: no effect modification ($r = 0$).** Share of runs with at least one false discovery, $\alpha = 0.05$, 200 seeds per n.

	share of runs with any false discovery				mean false
n	NEXIS	Marginal	Marginal (FWER)	Marginal (FDR)	discoveries (NEXIS)
50	0.000	0.995	0.000	0.000	0.000
100	0.000	1.000	0.000	0.000	0.000
200	0.000	1.000	0.000	0.000	0.000
350	0.000	1.000	0.000	0.000	0.000
500	0.000	1.000	0.000	0.000	0.000
750	0.010	1.000	0.010	0.010	0.010
1000	0.015	1.000	0.015	0.015	0.015
2000	0.000	1.000	0.000	0.000	0.000
3500	0.005	1.000	0.005	0.005	0.005
5000	0.000	1.000	0.000	0.000	0.000
10000	0.000	1.000	0.000	0.000	0.000
all n	0.003	1.000	0.003	0.003	0.003

One modifier ($r = 1$). NEXIS recovers one modifier about as well as two (Figure 17): macro recall and IoU are 0.749 and 0.734, against 0.750 and 0.745 at $r = 2$, and the IoU at $n = 2000$, $\eta = 5$ is 1.000 in both; macro precision is lower, 0.734 against 0.808. The reason is that at $r = 2$ recovery is limited by WEARING_HAT, the harder of the two modifiers to recover. Its per-modifier macro recall, the share of runs whose selection contains its principal coordinate, is 0.713 at $r = 2$ against 0.787 for EYEGLASSES, and only 0.749 at $r = 1$, where it is the only modifier. The purely prognostic EYEGLASSES coordinate 5537 is selected in none of the $2{,}100$ NEXIS runs at $r = 1$.

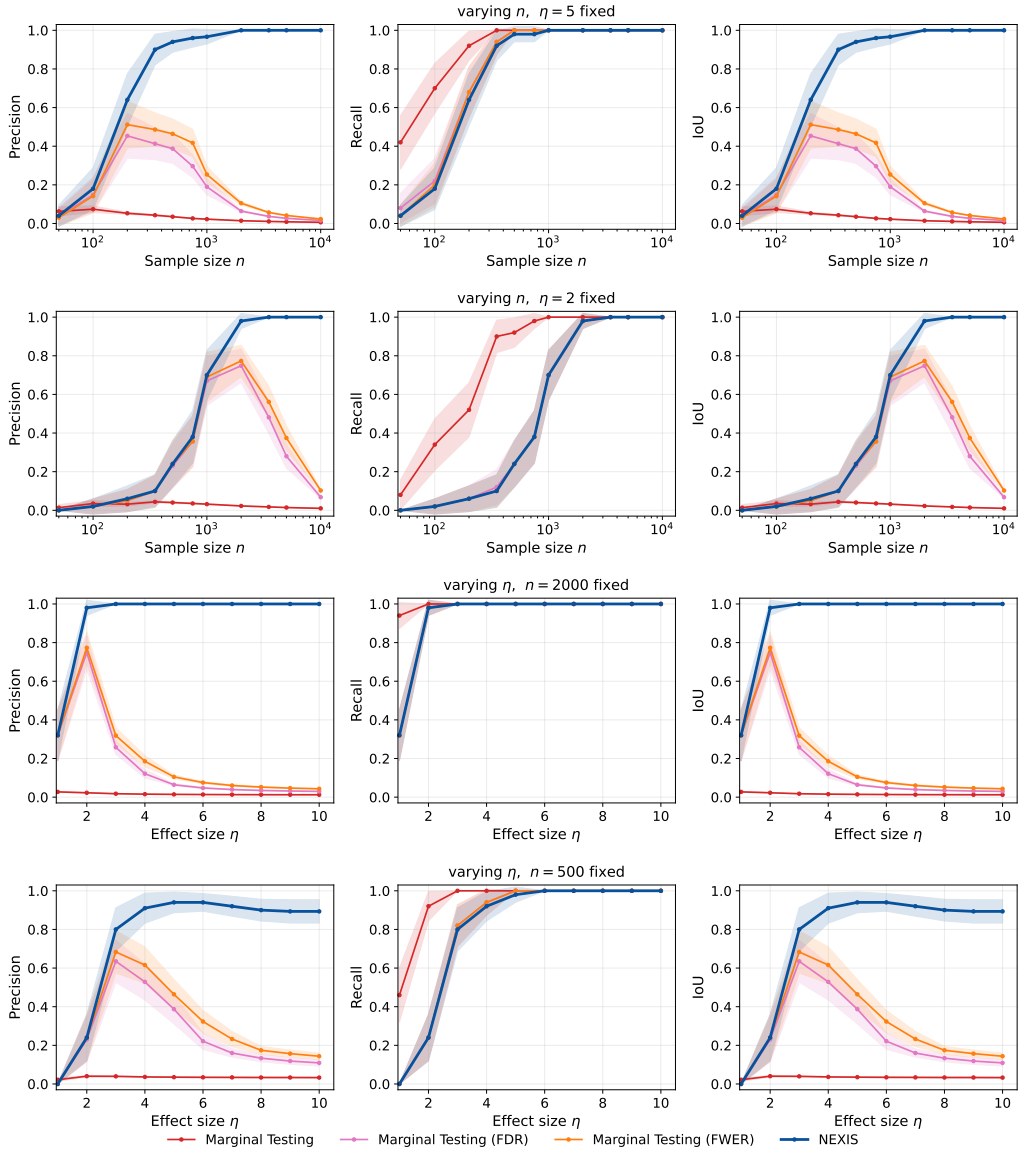


Figure 17: **DGP ablation: one direct modifier ($r = 1$).** Same layout as Figure 6, with only WEARING_HAT modifying the effect; EYEGLASSES keeps its prognostic effect. Recall and IoU are close to those with two modifiers. Reference: Figure 6.

Three modifiers ($r = 3$). The third modifier delays recall (Figure 18): recall crosses 0.95 at $n = 1000$ when $\eta = 5$ and at $n = 3500$ when $\eta = 2$, against $n = 750$ and $n = 2000$ at $r = 2$, and at $n = 500$ it stays below 0.95 up to $\eta = 10$. Macro precision drops less, from 0.808 to 0.761, and the IoU at $n = 2000$, $\eta = 5$ is 0.953; macro recall and IoU are 0.660 and 0.644, against 0.750 and 0.745. SIDEBURNS is the least aligned of the three concepts: its principal coordinate separates it with AUC 0.805 on $\boldsymbol{Z}$, against 0.969 and 0.956 for the other two (Appendix C.3). Consistently, its per-modifier macro recall is the lowest, 0.597, against 0.657 for WEARING_HAT and 0.726 for EYEGLASSES.

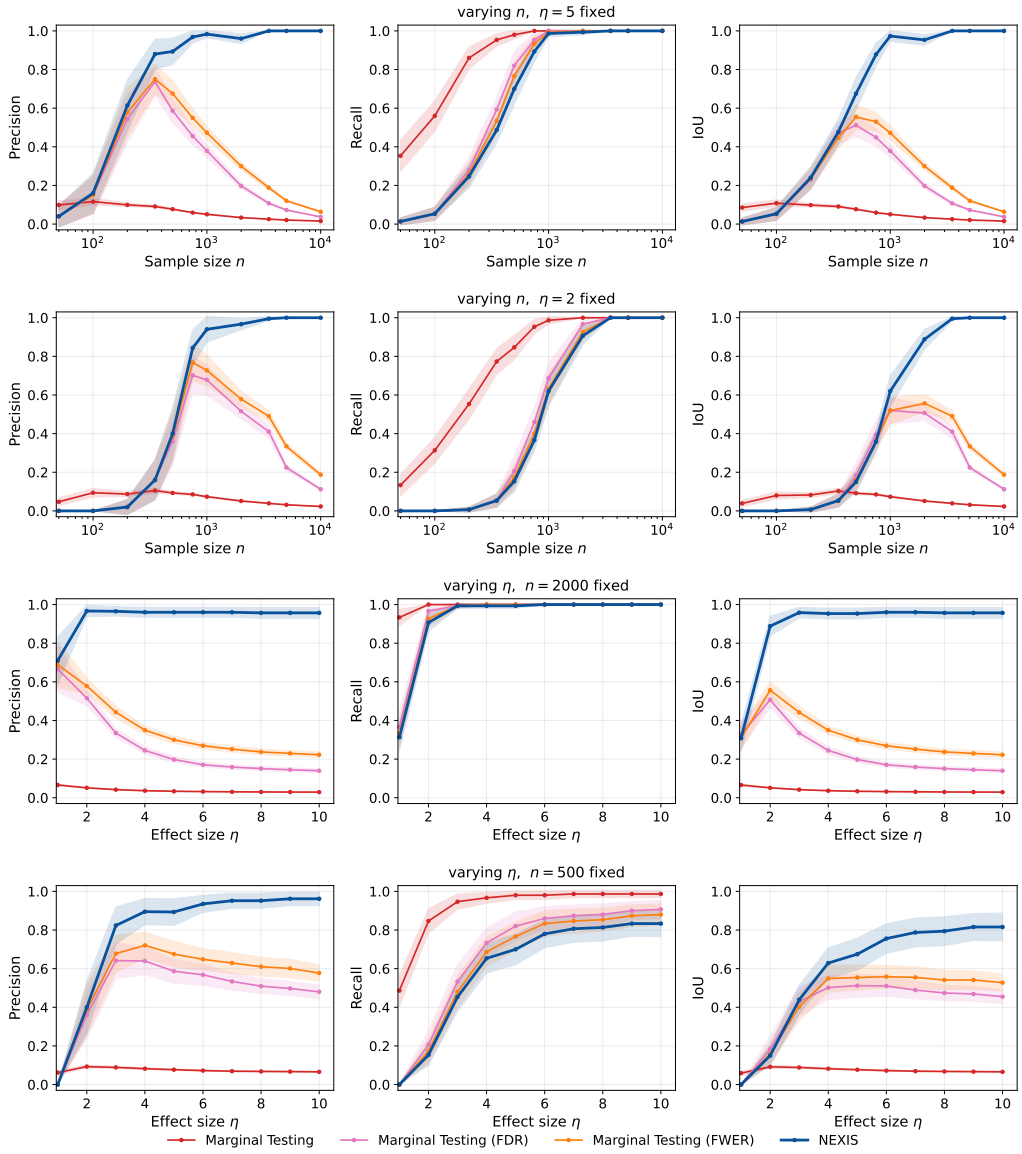


Figure 18: **DGP ablation: three direct modifiers ($r = 3$).** Same layout as Figure 6, with SIDEBURNS added as a third modifier. Recall is delayed, most for the less aligned SIDEBURNS. Reference: Figure 6.

C.7 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, $\rho = 0.5$, forward step followed by the terminal backward step. This is the configuration we run by default in the applied analyses (Appendix D, Appendix E), with two principled deviations:

- *Switch the test* when the conditional CATE cannot be assumed close to linear in $\boldsymbol{Z}$: to the GCM when the heterogeneity may be nonlinear but monotone, and to the PCM when non-monotone heterogeneity cannot be ruled out and enough observations are available. Both pay a power penalty on the semi-synthetic benchmark (Figure 13), and only the PCM covers non-monotone alternatives (Table 3, Appendix B.1). Both applied analyses use the linear test.
- *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 principal modifiers are expected to have similar standardized conditional signal strengths; in the absence of such prior, $\rho = 0.5$ is the safer choice.

A compact summary of the ablations is given in Table 5.

Table 5: Summary of CelebA ablations. Each row deviates from the main setting (k=20, $\boldsymbol{Z}$, FWER, ρ=0.5, linear test, terminal backward step) along a single axis.

Ablation (figure)	Varies	Finding
SAE sparsity (Fig. 11)	k=5 vs. 20	Too sparse ($k = 5$) leads to same recall, lower precision
Feature view (Fig. 12)	$\boldsymbol{Z}_{\text{pre}}$ vs. $\boldsymbol{Z}$	$\boldsymbol{Z}_{\text{pre}}$ needs more data
Conditional test (Fig. 13)	Linear vs. GCM/PCM	Linear best on a linear CATE; only PCM sees a U-shape
MHT correction (Fig. 14)	None / FDR / FWER	FDR $\equiv$ FWER; unadjusted delays recall
Spectral gap (Fig. 15)	$\rho \in \{0, 0.2, 0.5, 0.8\}$	ρ=0.5 jointly optimal
Backward steps (Fig. 16)	Interleaved / terminal	Terminal raises precision; interleaved inactive here

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.

C.8 COMPUTATIONAL BUDGET

Table 6: Compute budget for the CelebA semi-synthetic experiments. GPU refers to a single NVIDIA RTX 2080 Ti (the replica SAE of Appendix C.2 was trained on one H100).

Component	Hardware	Runtime
SigLIP 2 embedding (19,867 images)	GPU	$\sim$10 min
SAE training (main $k \in \{5, 20\}$, replica $k = 20$)	GPU	$\sim$1 h
Principal coordinates and Principal Alignment tests	CPU	minutes
All experiments (NEXIS, baselines)	CPU	$\sim$400 CPU-h ($\sim$10 h on 40 cores)

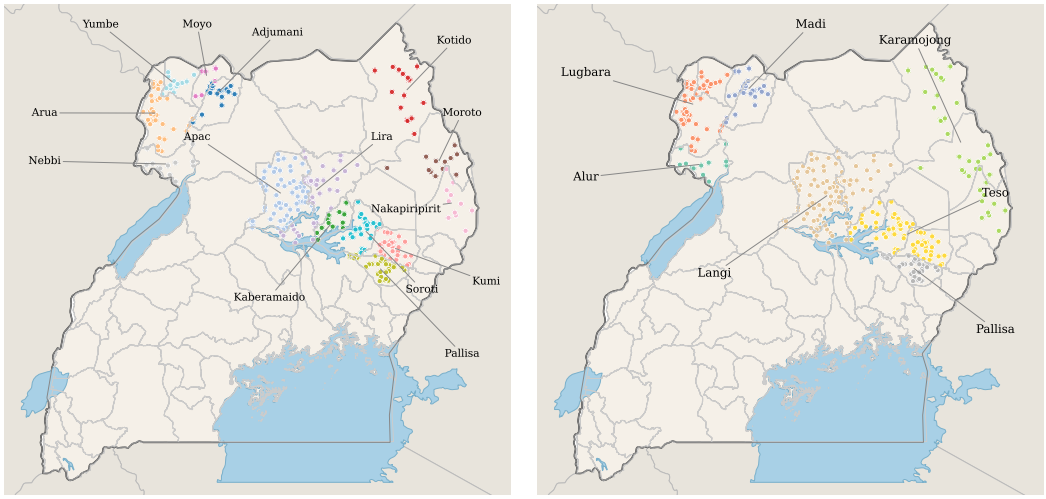
D Application 1: Youth Opportunities Program (Uganda)

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 (Blattman et al., 2014) 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 ∼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 327 distinct geographic communities, with a treatment rate of 39.6% (825 treated). Sites span Karamoja, Teso, Lango, and West Nile sub-regions. Figure 19 shows the geographic distribution of communities by district and by language group.



(a) Communities by district.

(b) Communities by language group.

Figure 19: Geographic distribution of the 331 YOP RCT sites (327 distinct 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, but we define T by grant receipt. Of the 265 groups assigned to the grant, 29 never received it, 21 for administrative reasons and 8 for theft or diversion (Blattman et al., 2014). We treat them as untreated; this exploratory as-treated analysis assumes non-receipt is as good as random. This may fail, since receiving groups were slightly more educated and wealthier, and is a limitation of our application. The ATE only summarizes whether the program worked on average; NEXIS asks *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, Karamojong, Pallisa[6]), the 5×5 km 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; 5×5 km tiles centred on each site's GPS centroid.
- *Bands.* Blue (B1), Green (B2), Red (B3), NIR (B4), SWIR1 (B5), SWIR2 (B7).
- *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** (Gao et al., 2024) 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, group composition (share female), 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. (2014) and reflects the dominant language of each district; it is not a modelling

[6]The language variable is inherited from Blattman et al. (2014) 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.

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, 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. NEXIS searches the 170 candidates in a single pass, in which the trial covariates and the satellite-derived features compete symmetrically.

Test. The default linear $T \times Z_j$ interaction t-test (Equation 27, Appendix B.1), conditioning on the already-selected set S. We keep the default for its power, which matters at $n = 2{,}082$ with 170 candidates: the GCM and PCM tests pay a measurable cost in power when the CATE is close to linear (Appendix B.1).

Correction. We run Algorithm 1 at $\alpha = 0.05$. The forward step admits the most significant candidate j if $p_j(S) \leq \alpha/|\bar{S}|$, with the spectral-gap gate at $\rho = 0.5$ (Appendix B.3): a candidate whose conditional t-statistic is less than half that of the weakest selected coordinate is not admitted. Once the forward step stops at $\widetilde{S}$, the backward step keeps $j \in \widetilde{S}$ only if $p_j(A) \leq \alpha/m$ for every $A \subseteq \widetilde{S} \setminus \{j\}$, with $m = 170$, i.e., $\alpha/m \approx 2.9 \times 10^{-4}$.

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

D.4 Discovered modifiers

Table 7 reports the coordinates selected by the forward step, in two tiers. *Certified* modifiers also survive the backward step: the Karamojong and Pallisa language groups and vegetation spatial heterogeneity (neuron 533) for skilled employment, and NDVI for log business assets. *Candidates* are selected by the forward step but not certified: the Lugbara language group and perennial river presence (neuron 339) for skilled employment, and structured agricultural landscape (neuron 820) for log business assets. A candidate carries a conditional signal, since it entered the selection given the coordinates selected before it, but at some subset of the other selected coordinates its p-value exceeds α/m. We read the candidates as hypotheses, not certified modifiers. For each coordinate, the table reports subgroup GATEs (active vs. inactive), the contrast Δ and two p-values. The *marginal* p-value tests the unconditional $T \times Z_j$ interaction and characterizes the raw signal of each coordinate. The *certification* p-value is the largest $p_j(A)$ over the subsets $A \subseteq \widetilde{S} \setminus \{j\}$ tested by the backward step, and the coordinate is certified if it is at most α/m. The original analysis of Blattman et al. (2014) reports significant positive average effects on both outcomes. The difference in means is also positive in our sample ($+0.32$ for skilled employment, $+0.61$ for log business assets), replicating the headline program-level finding.

Results interpretation. We add three observations. First, none of the individual-level demographics (age, sex, parental education) is selected on either outcome; 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. The two dampening groups (Karamojong and, as a candidate, Lugbara) sit at distinct national peripheries in the northeast and the northwest respectively, and were both conflict-affected or isolated at trial time (baseline 2008, endline 2010–2012), 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. Karamoja was under a government disarmament campaign through 2011, and West Nile was 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. 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 the selected coordinates: under the assumptions of Theorem 4.1

Table 7: Coordinates selected by NEXIS on YOP: *certified* modifiers survive the backward step, *candidates* are selected by the forward step only. GATE estimates for active vs. inactive subgroups (s.e. in parentheses), contrast Δ, *marginal* p-value of the unconditional linear $T \times Z_j$ interaction, and *certification* p-value, i.e., the largest $p_j(A)$ over the subsets tested by the backward step, compared with $\alpha/m = 0.05/170 \approx 2.9\times10^{-4}$. 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			p-value	
Modifier	Active	Inactive	Δ	Marginal	Certification
Panel A: Skilled Employment					
Certified					
Karamojong	-0.030 (0.060)	$+0.372$ (0.022)	-0.403 (0.063)	7.7×10^{-10}	7.6×10^{-8}
Pallisa	$+0.674$ (0.058)	$+0.288$ (0.022)	$+0.386$ (0.062)	6.3×10^{-8}	1.0×10^{-4}
Vegetation spatial heterogeneity	$+0.214$ (0.038)	$+0.373$ (0.025)	-0.159 (0.045)	6.7×10^{-5}	2.7×10^{-4}
Candidates					
Lugbara	$+0.092$ (0.061)	$+0.347$ (0.023)	-0.255 (0.065)	1.6×10^{-5}	3.2×10^{-4}
Perennial river presence	$+0.089$ (0.097)	$+0.330$ (0.021)	-0.242 (0.100)	2.1×10^{-4}	4.4×10^{-4}
Panel B: Log Business Assets					
Certified					
NDVI	$+0.668$ (0.061)	$+0.552$ (0.068)	$+0.115$ (0.092)	5.6×10^{-5}	5.6×10^{-5}
Candidates					
Structured agricultural landscape	$+0.368$ (0.094)	$+0.649$ (0.051)	-0.282 (0.107)	1.7×10^{-2}	1.7×10^{-2}

(Measurement and Representation Sufficiency, Principal Alignment, faithfulness, and validity and consistency of the CATE-equivalence test), NEXIS recovers the principal proxies of the direct effect modifiers with asymptotic probability at least $1 - \alpha$, 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. Certified modifiers and candidates alike should therefore be read as hypotheses, the certified ones being the strongest our pipeline can produce on this dataset (see also Section D.7), intended as inspirational starting points for the next iteration of the program rather than as established causal claims.

D.5 Marginal screening vs. NEXIS

To quantify the experimental power paradox of Section 3 on this dataset, we compare NEXIS to a pure marginal screen: each of the 170 coordinates tested individually for a $T \times Z_j$ interaction, without conditioning. Table 8 reports the result.

Table 8: Marginal screening vs. NEXIS on YOP. "Marginal" counts features for which the unconditional $T \times Z_j$ interaction is significant at level 0.05, uncorrected. "NEXIS" counts the coordinates selected by the forward step, split into those certified by the backward step and the candidates.

Outcome	Marginal	NEXIS
skilled employment	71 features	3 certified + 2 candidates
log business assets	45 features	1 certified + 1 candidate

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 redundancy among 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, and the backward step keeps only the coordinates that remain significant at level α/m given every subset of the others. The forward step selects 5 and 2 coordinates, of which the backward step certifies 3 and 1; under the assumptions of

Theorem 4.1, the certified set equals the set of principal proxies with asymptotic probability at least $1 - \alpha$.

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 261** $\rightarrow$ *perennial water presence* (open water bodies and wetlands; used in Appendix D.7).
- **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).

Figure 4 shows the two top- and two bottom-activating tiles of those shown to the VLM, plus the map of communities where each neuron is active.

D.7 LIMITATIONS

We flag three limitations of the present analysis.

Multilevel inference. Our tests treat individuals as independent. Individual outcomes are the unit of analysis, the guarantee of Theorem 4.1 is stated for i.i.d. units, and the individual-level test gives the power needed to generate hypotheses. Treatment, however, is assigned by group, and the satellite features vary by community. We therefore repeat the analysis with standard errors clustered by treatment-assignment group ($G = 439$) and district fixed effects, as in the original evaluation (Blattman et al., 2014), on the 128 candidates active in at least 5 groups per treatment arm. NEXIS then certifies only the Karamojong and Lugbara language groups for skilled employment, and one satellite atom, labelled *perennial water presence* by the VLM (neuron 261), for log business assets. The four environmental modifiers of Table 7 keep some evidence: given the other coordinates selected for the same outcome, a randomization test that replays the group lottery within district gives them p-values between 1.1×10^{-3} and 1.3×10^{-2}, short of α/m. The loss comes from the number of units, not from the effects: the effective sample shrinks from individuals to groups, and community-level atoms are active in few groups, e.g., the river atom in 4 treated groups. Clustering alone would leave the point estimates unchanged and only widen their uncertainty. The language groups vary at a coarser level still, but clustering at their own level (7 ethnolinguistic clusters) would ask whether the effect generalizes to other regions, which needs more sites. We therefore present the YOP modifiers as suggestive hypotheses.

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 5×5 km tile is centred on the community centroid; 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 PCM test of Appendix B.1 is a drop-in replacement that also detects non-monotone alternatives; we use the linear test here for its power at the present sample size.

D.8 Computational budget

Table 9: 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 (4 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 20. Treatment and comparison groups are constructed via a regression discontinuity design: eligibility is determined by a proxy means test (PMT) score cutoff, and a household is treated ($T = 1$) if its PMT score is at or below the cutoff, so that poorer households receive the transfer. Households just below and just above the threshold form the treatment and comparison groups respectively (Ghana LEAP 1000 Evaluation Team, 2018). Consistent with this design, 154 of 162 communities contain both treated and comparison households, with a median within-community treatment share of $\approx 50\%$.



Figure 20: 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 effect of eligibility for households near the PMT cutoff. We interpret this as the local average treatment effect on the treated (local ATT) for the marginal eligible population under the explicit assumption that noncompliance is ignorable and does not alter the eligibility contrast. This assumption and possible selection into the balanced panel are limitations.

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. Throughout this subsection, τ, $\mathcal{S}^\star$, and the structural assumptions are understood under the distribution of the marginal eligible population near the cutoff.

Asymptotic recovery 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 asymptotic exact-recovery guarantee (Equation 9) rests on Assumption 5: the effect heterogeneity-equivalence tests must be valid under the null and consistent under alternatives. For this application, we impose that the first-differenced untreated potential outcome is mean-independent of T given every fixed subset tested by NEXIS:

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

This is a subset-wise covariate-adjusted parallel trends condition (Callaway & Sant'Anna, 2021): the untreated counterfactual trend carries no residual mean dependence on treatment assignment at any conditioning set used by the algorithm. Together with the regularity conditions of the chosen test, Equation 34 supplies the application-specific validity required by Assumption 5(a); test consistency under alternatives remains Assumption 5(b). The recovery proof of Section A.2 then applies unchanged. 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 interpretation in the local population. Within a bandwidth around the PMT cutoff, local as-if-randomization makes units on either side comparable in expectation for variables varying continuously at the threshold (Lee & Lemieux, 2010). Together with Equation 34, this identifies the conditional local ATT from the observed treatment groups. It is not what upgrades conditioning on $\boldsymbol{W}^{\text{dir}}$ to intervening on it: once the local CATE is identified, $\tau(\boldsymbol{W}^{\text{dir}}) = \tau^{do}(\boldsymbol{W}^{\text{dir}})$ follows

from the definition of $\boldsymbol{W}^{\mathrm{dir}}$ as the complete direct-modifier set and the back-door argument of Section A.3. The identified interventional characterization is accordingly restricted to the marginal eligible population near the cutoff.

The subset-wise condition is a strong extension of standard conditional parallel trends because it must hold for every tested A; local as-if-randomization is standard. Both are precisely stateable and empirically supportable but not formally testable from the data alone. Together with the structural and testing assumptions of Theorem 4.1, they extend its guarantees to this quasi-experimental setting, with 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 features that operate at two levels of the data hierarchy: 131 learned landscape features from the SAE and 12 spectral indices computed from the same imagery (NDVI, NDWI, MNDWI, NDBI, EVI, BSI; mean and standard deviation per tile) at the community level, and 24 household-level survey covariates, for a total of 167 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 all of them 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.

Search. NEXIS searches the 167 candidates jointly in a single pass.

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 run Algorithm 1 at $\alpha = 0.05$, as in the Uganda application (Appendix D.3). The forward step admits the most significant candidate j if $p_j(S) \leq \alpha/|\bar{S}|$, with the spectral-gap gate at $\rho = 0.5$. Once it stops at $\widetilde{S}$, the backward step keeps $j \in \widetilde{S}$ only if $p_j(A) \leq \alpha/m$ for every $A \subseteq \widetilde{S} \setminus \{j\}$, with $m = 167$, i.e., $\alpha/m \approx 3.0 \times 10^{-4}$. An exploratory analysis without correction is reported in Section E.7.

E.5 DISCOVERED EFFECT MODIFIERS

NEXIS retains two satellite-derived landscape features, and no spectral index or household covariate (Table 10). The forward step selects no other coordinate and both pass the backward step, so both are certified and there are no candidates. Both carry high-confidence VLM semantic labels. The difference-in-differences estimate of the local ATT is $+7.35$ GH¢/month; in the few communities where these features are present, the estimated effect is 6–8$\times$ larger.

Table 10: NEXIS discoveries for LEAP 1000, both certified by the backward step. GATE estimates (in GH¢/month) for active vs. inactive subgroups (s.e. clustered by community in parentheses), contrast Δ, *marginal* p-value of the unconditional linear $T \times Z_j$ interaction, and *certification* p-value, i.e., the largest $p_j(A)$ over the subsets tested by the backward step, compared with $\alpha/m = 0.05/167 \approx 3.0\times10^{-4}$; both use standard errors clustered by community. Satellite-atom modifiers are reported under their VLM-assigned semantic label; activations are binarized at $z > 0$.

				p-value	
Modifier	GATE (active)	GATE (inactive)	Δ	Marginal	Certification
Ephemeral waterways	$+42.9$ (15.8)	$+6.0$ (3.9)	$+36.9$	3.2×10^{-8}	3.2×10^{-8}
Closed-canopy forest	$+56.2$ (19.7)	$+6.4$ (3.9)	$+49.8$	2.0×10^{-5}	2.0×10^{-5}

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 11): 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 11: 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. 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

Running NEXIS without multiple-testing correction on an earlier, smaller pool, which keeps the 72 neurons active in at least 10 communities together with the 24 survey covariates and 6 spectral indices, reveals additional patterns that are not statistically sufficient to certify discoveries but are informative for hypothesis generation. This pool excludes the two certified neurons, which are active in fewer communities. The most substantive pattern 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 is $+26.3$ GH¢/month in active communities vs. $+5.6$ in inactive ones; the conditional $p = 0.0085$ in this uncorrected run, but the GATE contrast is not significant ($p = 0.30$), and the signal does not survive FDR correction.

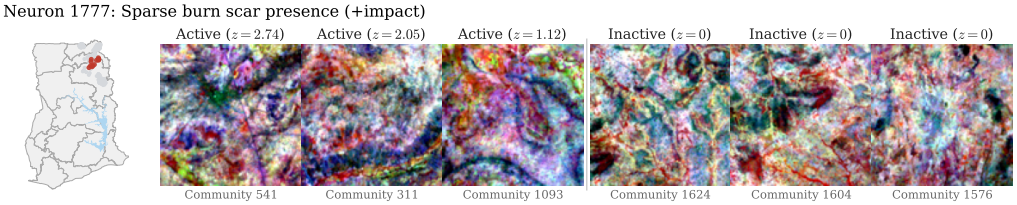


Figure 21: Top- and bottom-activating Landsat-8 tiles in false-color composite (NIR, Green, SWIR) for burn-scar, with the top 12 active communities concentrated in the East Mamprusi–Karaga cluster (see map). Active tiles show sparse irregular burn scars scattered across vegetation (VLM label: *sparse 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 (National Development Planning Commission, 2016) documents, under natural disasters and vulnerability (Section 4.4.10), that 151 metropolitan, municipal and district assemblies were affected by natural disasters in 2015, especially flooding and bush fires, with the bulk of incidents in the Eastern and Northern Regions. The burn scars in the

2015 imagery of the active communities are consistent with this regional record, 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. 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 (Ghana LEAP 1000 Evaluation Team, 2018) 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. Linearity is exact for the binary survey covariates; the other survey covariates (counts, ratios, head age) and the 12 continuous spectral indices carry the assumption non-trivially. The PCM test of Appendix B.1 is a drop-in replacement that also detects non-monotone alternatives, and is the recommended choice when enough data are available.

Active-community counts. The discovered features are active in only 6 (83 households) and 5 (42 households) communities respectively. Clustered inference with so few active clusters can overstate precision: a wild cluster bootstrap by community, testing each neuron given the other, gives $p = 0.015$ and $p = 0.056$ respectively, far above α/m. Community is also not the treatment-assignment unit (households are assigned by PMT score), so neither the clustered nor the bootstrap p-values are design-based. The magnitudes should therefore be treated with caution pending larger evaluations.

E.9 Computational budget

Table 12: 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 (167 candidates)	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