

A Related Work

Mechanistic interpretability aims to identify how internal components—such as attention heads and MLPs—contribute to model behavior. Early work explored circuit discovery in simplified settings [2], while recent methods adopt Pearl’s causal theory [3] and ablation-based interventions to reveal components essential to specific outputs.

Node-level patching These methods aim to identify the contribution of individual input features to a model’s prediction. Vig et al. [4] investigate gender bias through causal mediation analysis by selectively patching specific input features. Wang et al. [5] ablate features one by one to study their roles in indirect object identification, and further report the phenomenon of backup behavior, where other components compensate for ablated information. Meng et al. [6] restore specific token-level features to identify where factual knowledge is stored in the model when predicting objects in sentences, and demonstrate direct editing of such representations. Building on this, Meng et al. [7] extend the editing framework to perform large-scale knowledge editing across many memory locations. Heimersheim and Janiak [8] analyze individual input features to uncover the mechanisms behind repeated argument names in Python docstring generation. Zhang and Nanda [9] systematically examine the impact of various methodological choices in node-level patching, providing a comprehensive evaluation of this line of research.

Edge-level patching These methods examine the influence of neighboring feature pairs with direct computational dependencies. Conmy et al. [10] propose a greedy edge ablation algorithm to automatically abstract model behavior by iteratively removing dependencies and observing the impact on predictions. Syed et al. [12] accelerate edge-level analysis by introducing a gradient-based first-order approximation that estimates the effect of ablating each edge efficiently. Bhaskar et al. [11] replace missing edges with counterfactual activations obtained from corrupted examples, enabling gradient-based pruning without relying on discrete search or linear approximations.

Path-level patching These approaches investigate the contribution of distant feature pairs connected through multiple accumulated dependencies. Chan et al. [13] justify that ablations using zero or mean activations can induce misleading interventions, and instead propose hypothetical tests to explore decision paths. Hanna et al. [14] manually specify and patch internal paths in a language model to explain the circuit responsible for a year-span prediction task. Nanda et al. [15] manually ablate specific paths to investigate where factual knowledge related to sports is stored within the model.

When applying existing methods to trace paths for a given decision, node- and edge-level approaches are feasible under a backward chaining framework. However, they lack reliability as they perform chaining without causal referencing, making it difficult to guarantee the correctness of the identified paths. Path-level methods, on the other hand, often rely on hypothetical tests or require manual specification of paths, which renders full explanations infeasible due to the combinatorial complexity of possible paths. In contrast, our method enables reliable identification of causal paths while remaining computationally efficient, achieving tractable search on average in polynomial time.

B Sufficient Intervention Example

Example 2. First, DIRECT NOISE fails to satisfy Condition (5.a) because, in nodes with residual connections, it becomes ambiguous which component was actually intervened upon. That is, the same structural equation (e.g., $v_i + v_j$) can correspond to multiple graphical structures, making it impossible to reliably isolate the effect of the intervention. For instance, if a noise term is added to $v_i + v_j$ to intervene on v_i , we cannot distinguish whether the perturbation was intended for v_i or v_j .

Second, NOISE TOKEN fails to satisfy Condition (5.c) because, even if the noise added to a token embedding is sampled from a statistically well-behaved distribution (e.g., Gaussian), the scale of the resulting representation after forwarding through the model cannot be guaranteed to be stable. This instability makes it difficult to interpret the intervention effect in a controlled and consistent manner.

Third, ZERO MASK violates Condition (5.b) because it can break the causal edge property. Consider a simple model $y = Wx + b$: if x is zero-masked, then the output becomes entirely determined by the bias term b , regardless of the original input. This may result in no valid causal node set under our setting, as the decision no longer depends on any specific input node.

755 *In contrast, TOKEN RESAMPLING is suitable for our framework, as it (i) induces a distinct structural*
 756 *equation via alternative token embeddings, (ii) preserves the validity of Property 1, and (iii) provides*
 757 *a non-parametric intervention with in-distribution resampling.*

758 C Full Derivation of Unfolding Transformer Block

759 We begin with the following equation, Equation (1):

$$\begin{aligned}
 \left[[z_q^{(h)}]_{h=1}^H; [z_k^{(h)}]_{h=1}^H; [z_v^{(h)}]_{h=1}^H \right] &= \text{L}_{ia}(\text{LN}_1(z_{ib})), \\
 z_{oa} &= \text{L}_{oa}([\text{softmax}(z_q^{(h)} z_k^{(h)\top} / \sqrt{d_h}) z_v^{(h)}]_{h=1}^H), \\
 z_{im} &= z_{ib} + z_{oa}, \\
 z_{om} &= \text{L}_{om}(\phi(\text{L}_{im}(\text{LN}_2(z_{im})))), \\
 z_{ob} &= z_{im} + z_{om},
 \end{aligned} \tag{6}$$

760 where $z_{ib}, z_{oa}, z_{im}, z_{om}, z_{ob} \in \mathbb{R}^{T \times d_m}$ and $z_q^{(h)}, z_k^{(h)}, z_v^{(h)} \in \mathbb{R}^{T \times d_h}$, where T denotes the number
 761 of tokens, d_m the model dimension, H the number of heads, and $d_h = d_m/H$. Here, this can be
 762 unfolded into $2H + 2$ paths, as follows:

$$\begin{aligned}
 z_{ob} &= z_{im} + z_{om} = \underbrace{z_{ib}}_{\text{Residual Only "1 Path"}} + z_{oa} + z_{om}. \\
 z_{oa} &= \underbrace{\left(\sum_{h=1}^H z_q^{(h)} z_k^{(h)\top} \odot D_\alpha z_v^{(h)} W_{oa}^\top \right)}_{\text{Attention Only}} + \underbrace{b_{oa}}_{\text{Bias Terms from Attention} = H \cdot b_{\text{attn}} \text{ (i.e., Attention params. only)}}, \\
 &\quad \text{(i.e., Attention params. with the block input } z_{ib}, \text{ as contained in } z_q^{(h)}, z_k^{(h)}, z_v^{(h)}) \\
 &= \underbrace{\left(\sum_{h=1}^H z_q^{(h)} z_k^{(h)\top} \odot D_\alpha z_v^{(h)} W_{oa}^\top + b_{\text{attn}} \right)}_{\text{Attention Only "H Paths"}}, \\
 z_{om} &= \underbrace{z_{oa} W_{ln_2}^\top W_{im}^\top \odot D_\beta W_{om}^\top}_{\text{Attention+MLP}} \\
 &\quad \text{(i.e., MLP params. with the attention output } z_{oa}) \\
 &\quad + \underbrace{z_{ib} W_{ln_2}^\top W_{im}^\top D_\beta W_{om}^\top}_{\text{MLP Only}} + \underbrace{b_{ln_2} W_{im}^\top \odot D_\beta W_{om}^\top + b_{im} \odot D_\beta W_{om}^\top + b_{om}}_{\text{Bias Terms from MLP} = H \cdot b_{\text{attn+mlp}} + b_{\text{mlp}}}, \\
 &\quad \text{(i.e., MLP params. with the block input } z_{ib}) \quad \text{(i.e., MLP params. only)} \\
 &= \left(\left(\sum_{h=1}^H z_q^{(h)} z_k^{(h)\top} \odot D_\alpha z_v^{(h)} W_{oa}^\top \right) + b_{oa} \right) W_{ln_2}^\top W_{im}^\top \odot D_\beta W_{om}^\top \\
 &\quad + z_{ib} W_{ln_2}^\top W_{im}^\top D_\beta W_{om}^\top + H \cdot b_{\text{attn+mlp}} + b_{\text{mlp}}, \\
 &= \left(\sum_{h=1}^H z_q^{(h)} z_k^{(h)\top} \odot D_\alpha z_v^{(h)} W_{oa}^\top W_{ln_2}^\top W_{im}^\top \odot D_\beta W_{om}^\top \right) + b_{oa} W_{ln_2}^\top W_{im}^\top \odot D_\beta W_{om}^\top \\
 &\quad + z_{ib} W_{ln_2}^\top W_{im}^\top D_\beta W_{om}^\top + H \cdot b_{\text{attn+mlp}} + b_{\text{mlp}}, \\
 &= \underbrace{\left(\sum_{h=1}^H z_q^{(h)} z_k^{(h)\top} \odot D_\alpha z_v^{(h)} W_{oa}^\top W_{ln_2}^\top W_{im}^\top \odot D_\beta W_{om}^\top + \frac{b_{oa} W_{ln_2}^\top W_{im}^\top \odot D_\beta W_{om}^\top}{H} \right)}_{\text{Attention+MLP "H Paths"}}, \\
 &\quad + \underbrace{z_{ib} W_{ln_2}^\top W_{im}^\top D_\beta W_{om}^\top + b_{\text{mlp}}}_{\text{MLP Only "1 Path"}}. \tag{7}
 \end{aligned}$$

763 D Proofs

764 D.1 Proofs of Theorem 1

765 **Lemma 1** (Expected Evaluation Number for Theorem 1). *Consider a minimality-based subset search*
 766 *over n nodes, where each subset is independently selected as a causal node set with probability p .*
 767 *Then, the expected number of subset evaluations over all subsets is bounded by:*

$$n + (1 - p) \times \sum_{s=2}^n \max \left(0, \binom{n}{s} + \sum_{i=1}^{s-1} \sum_{m=1}^{\lfloor p \binom{n}{i} \rfloor} (-1)^m \binom{p \binom{n}{i}}{m} \binom{n - mi}{s - mi} \right).$$

768
 769 *Proof.* At step $s = 1$, there are $\binom{n}{1} = n$ singleton subsets, all of which must be evaluated, resulting
 770 in exactly n evaluations.

771 For each step $s \geq 2$, the number of candidate subsets of size s is $\binom{n}{s}$. A subset of size s is pruned if it
 772 contains any previously selected subset of size $i < s$. Thus, the expected number of subsets to be
 773 evaluated at step s is the total number of candidates minus the expected number of pruned subsets:

$$\binom{n}{s} - \mathbb{E}[\text{number of pruned subsets at step } s].$$

774 To estimate the pruning term, note that the number of subsets of size i is $\binom{n}{i}$, and each is independently
 775 selected with probability p . For any such subset of size $i < s$, the number of supersets of size s that
 776 include it is $\binom{n-i}{s-i}$, as the remaining $s - i$ nodes must be selected from the $n - i$ nodes not in the
 777 subset. Therefore, each selected subset of size i contributes in expectation $p \cdot \binom{n-i}{s-i}$ pruned subsets at
 778 step s . Multiplying by the total number $\binom{n}{i}$ of such subsets yields the expected number of pruned
 779 subsets due to size- i selections:

$$p \binom{n}{i} \binom{n-i}{s-i}.$$

780 Summing over all $1 \leq i < s$, the total expected number of pruned subsets at step s is:

$$\sum_{i=1}^{s-1} p \binom{n}{i} \binom{n-i}{s-i}. \quad (8)$$

781 Equation (8) assumes that pruning effects from different subsets are disjoint. However, a subset of
 782 size s may be pruned by multiple selected subsets of size i , which results in overcounting. To correct
 783 for this, we apply the inclusion–exclusion principle. Equation (8) corresponds to the first-order term
 784 of this expansion, and is fully incorporated into the more general form below.

785 For instance, when two subsets of size i are both selected and jointly included in a size- s subset, that
 786 subset is overcounted in the previous estimate and must be subtracted once. The number of such
 787 expected double-overlaps is:

$$- \sum_{i=1}^{s-1} \binom{p \binom{n}{i}}{2} \binom{n-2i}{s-2i}.$$

788 This logic generalizes to arbitrary overlap levels. When m mutually disjoint subsets of size i are
 789 selected and jointly included in a size- s subset, the corresponding correction term alternates in sign
 790 according to the inclusion–exclusion principle, and takes the form:

$$(-1)^m \binom{p \binom{n}{i}}{m} \binom{n-mi}{s-mi}.$$

791 Summing over all $m \geq 1$, the total expected number of pruned subsets at step s , with successive
 792 overcount corrections added and subtracted depending on the overlap level, is:

$$\sum_{i=1}^{s-1} \sum_{m=1}^{\lfloor p \binom{n}{i} \rfloor} (-1)^m \binom{p \binom{n}{i}}{m} \binom{n-mi}{s-mi}. \quad (9)$$

793 To obtain the total expected number of subset evaluations over all steps, we sum the unpruned subsets
 794 across all sizes. At step $s = 1$, all n singleton subsets are evaluated.

795 For $s \geq 2$, each subset of size s survives pruning with probability $1 - p$, and only if it has not been
 796 eliminated due to overlap with smaller selected subsets. Thus, the expected number of evaluations at
 797 step s is:

$$(1 - p) \cdot \left(\binom{n}{s} + \sum_{i=1}^{s-1} \sum_{m=1}^{\lfloor p \binom{n}{i} \rfloor} (-1)^m \binom{p \binom{n}{i}}{m} \binom{n - mi}{s - mi} \right).$$

798 The inner summation captures the pruning corrections from the inclusion–exclusion expansion in
 799 Equation (9). To ensure non-negativity of the expected number at each step, we take a maximum with
 800 zero.

801 Adding this to the evaluations from $s = 1$, the total expected number of subset evaluations is:

$$n + (1 - p) \sum_{s=2}^n \max \left(0, \binom{n}{s} + \sum_{i=1}^{s-1} \sum_{m=1}^{\lfloor p \binom{n}{i} \rfloor} (-1)^m \binom{p \binom{n}{i}}{m} \binom{n - mi}{s - mi} \right).$$

802

□

803 **Theorem 1** (Expected Time Complexity of Minimality-based Subset Search). *Consider a minimality-*
 804 *based subset search over n nodes, where each subset is independently selected as a causal node set*
 805 *with probability p . Then, the expected number of subset evaluations is upper-bounded as in Lemma 1.*
 806 *Based on this bound, the expected time complexity grows approximately as*

$$O \left(n^{\lfloor \log_2(\frac{1}{p} + 2) \rfloor} \right).$$

807 *Proof.* From Lemma 1, the expected number of subset evaluations is:

$$n + (1 - p) \sum_{s=2}^n \max \left(0, \binom{n}{s} + \sum_{i=1}^{s-1} \sum_{m=1}^{\lfloor p \binom{n}{i} \rfloor} (-1)^m \binom{p \binom{n}{i}}{m} \binom{n - mi}{s - mi} \right). \quad (10)$$

808 We approximate the pruning term by keeping only the $m = 1$ intersection term in the inclusion-
 809 exclusion sum, discarding the rest ($m \geq 2$), which largely cancel out due to alternating signs.

$$\binom{n}{s} - \sum_{i=1}^{s-1} p \binom{n}{i} \binom{n - i}{s - i}.$$

810 Using the fact that $\binom{n}{i} \binom{n - i}{s - i} = \binom{n}{s} \binom{s}{i}$ and $\sum_{i=1}^{s-1} \binom{s}{i} = 2^s - 2$, we simplify the above as:

$$\binom{n}{s} (1 - p(2^s - 2)).$$

811 Substituting this into Equation (10), we obtain the following:

$$n + (1 - p) \sum_{s=2}^n \max \left(0, \binom{n}{s} (1 - p(2^s - 2)) \right). \quad (11)$$

812 A subset of size s contributes to the sum ($\sum_{s=2}^n$) only if the term with pruning factor is positive, i.e.,
 813 the term $(1 - p(2^s - 2))$ applied to the positive binomial term $\binom{n}{s}$ is greater than zero:

$$1 - p(2^s - 2) > 0 \quad \Leftrightarrow \quad s < \log_2 \left(\frac{1}{p} + 2 \right).$$

814 Here, let $s^* := \left\lfloor \log_2 \left(\frac{1}{p} + 2 \right) \right\rfloor$ be the largest such s that satisfies this condition. Then, Equation (11)
 815 reduces to:

$$n + (1 - p) \sum_{s=2}^{s^*} \binom{n}{s} (1 - p(2^s - 2)).$$

816 Since each binomial coefficient satisfies $\binom{n}{s} = O(n^s)$ and the term $(1 - p(2^s - 2))$ acts as a constant
 817 for each fixed s , the summation is upper-bounded as

$$\sum_{s=2}^{s^*} \binom{n}{s} (1 - p(2^s - 2)) = O(n^2 + \dots + n^{s^*}) = O(n^{s^*}).$$

818 Thus, the total number of evaluations grows approximately as

$$n + (1 - p) \cdot O(n^{s^*}) = O(n^{s^*}).$$

819

□

820 **Remark 2.** (This remark extends Remark 1.) Since the underlying problem is NP-complete, the time
 821 complexity approaches $O(2^n)$ as p becomes very small. This occurs when the term with pruning
 822 factor $1 - p(2^s - 2)$ remains positive up to $s = n$, which requires $p < \frac{1}{2^n - 2}$. However, such cases
 823 occur only infrequently in practice, especially considering that n corresponds to the number of path
 824 nodes in a single transformer block, typically $2H + 2$.

825 D.2 Proof of Theorem 2

826 **Theorem 2** (Causal Union Reference Reliability). Consider a minimality-based subset search over n
 827 nodes, where each subset is independently selected as a causal node set with probability p . Suppose
 828 that a collection of such sets, $V_{out}^{(j+1)} = \{V_i^{(j+1)}\}_{i=1}^k$, is identified from the $(j+1)$ -th block, i.e., the
 829 one directly downstream. Their union, denoted as $P = \bigcup_{i=1}^k V_i^{(j+1)}$, serves as the causal subpath
 830 reference for the minimality-based subset search in the j -th block. Let s_{avg} denote the average size of
 831 the k causal node sets in $V_{out}^{(j+1)}$. Then, the reliability of the resulting causal node set obtained using
 832 P is given by:

$$p + (1 - p) \left(1 - \left(1 - \frac{s_{avg}}{n} \right)^k \right)^n \rightarrow 1 \quad (12)$$

833 *Proof.* To evaluate the reliability of using the union $\bigcup_{i=1}^k V_i^{(j+1)}$ as the causal subpath reference P
 834 during the search in the j -th block, we consider the following two cases under which this strategy
 835 ensures a reliable outcome.

836 First, P itself could have been selected as a causal node set during the search in the $(j+1)$ -th block,
 837 independently with probability p , even though it was not explicitly selected because its subsets had
 838 already been included.

839 Second, if P was not selected in this way, it must equal the full set of path nodes in the j -th block. In
 840 this case, P must at least correspond to the case in Property 1, exemplified by Example 1.

841 The total probability of success across these two cases is given by:

$$p + (1 - p) \cdot \hat{p}, \quad (13)$$

842 where \hat{p} denotes the probability of the second case.

843 To compute \hat{p} , note that the probability that a single causal node set in $V_{out}^{(j+1)}$, with average size s_{avg} ,
 844 contains a given node among the n path nodes in the j -th block is:

$$\frac{s_{avg}}{n}.$$

845 Accordingly, the probability that at least one of the k causal node sets contains a given node is:

$$1 - \left(1 - \frac{s_{avg}}{n} \right)^k.$$

Extending this, the probability that the k causal node sets collectively cover all n path nodes, i.e., that P equals the full set of path nodes in the j -th block, is:

$$\hat{p} = \left(1 - \left(1 - \frac{s_{\text{avg}}}{n}\right)^k\right)^n.$$

Substituting this into Equation (13), the reliability can be expressed as follows:

$$p + (1 - p) \cdot \left(1 - \left(1 - \frac{s_{\text{avg}}}{n}\right)^k\right)^n.$$

To analyze convergence, define $x = \frac{s_{\text{avg}}}{n}$. Using the inequality:

$$(1 - x) \leq e^{-x},$$

we obtain:

$$(1 - (1 - x)^k)^n \geq (1 - e^{-xk})^n.$$

Substituting back $x = \frac{s_{\text{avg}}}{n}$, the lower bound for \hat{p} becomes:

$$\hat{p} \geq \left(1 - e^{-k \cdot \frac{s_{\text{avg}}}{n}}\right)^n.$$

Here, k is equal to the number of subset evaluations selected with probability p from the minimality-based subset search. As n grows large, k asymptotically follows the evaluation bound shown in Lemma 1, which converges to n^{s^*} . Therefore, k can be approximated as:

$$k \approx p \cdot n^{s^*} = p \cdot n^{\lfloor \log_2(\frac{1}{p} + 2) \rfloor}.$$

Thus, the reliability is lower bounded as:

$$\begin{aligned} p + (1 - p) \cdot \left(1 - \left(1 - \frac{s_{\text{avg}}}{n}\right)^k\right)^n &\geq p + (1 - p) \cdot \left(1 - e^{-k \cdot \frac{s_{\text{avg}}}{n}}\right)^n \\ &\approx p + (1 - p) \cdot \left(1 - e^{-p \cdot n^{s^*} \cdot \frac{s_{\text{avg}}}{n}}\right)^n \end{aligned}$$

We now analyze the asymptotic behavior of this lower bound as n grows large, focusing on how the reliability converges under varying values of p .

Specifically, as $p \rightarrow 1$, i.e., when most combinations of path nodes are selected as causal node sets with probability close to 1, the minimality-based subset search algorithm tends to converge at step 1. That is, both s^* and s_{avg} converge to 1, and the expression simplifies as follows:

$$p + (1 - p) \cdot (1 - e^{-p})^n \rightarrow 1$$

This is because, as $p \rightarrow 1$, the factor $(1 - p)$ approaches zero, while $(1 - e^{-p})^n$ vanishes as n increases. As a result, only the p term remains, and the expression converges to 1.

Conversely, as $p \rightarrow 0$, i.e., when the probability of selecting any combination of path nodes as a causal node set becomes vanishingly zero, the minimality-based subset search algorithm tends to converge at step n ; thus, $s^* \rightarrow n$. In this case, s_{avg} converges to $\frac{n}{2}$ for the following reason:

$$s_{\text{avg}} = \frac{\sum_{i=1}^{s^*} i \cdot \binom{n}{i}}{\sum_{i=1}^{s^*} \binom{n}{i}} \approx \frac{n \cdot 2^{n-1}}{2^n} = \frac{n}{2}$$

Therefore, the lower bound of the reliability simplifies as follows:

$$p + (1 - p) \cdot \left(1 - e^{-p \cdot n^{n \cdot \frac{1}{2}}}\right)^n \rightarrow 1$$

This is because the term $\left(1 - e^{-p \cdot n^{n \cdot \frac{1}{2}}}\right)^n$ converges to 1 as n grows large, while p goes to zero. Hence, the entire expression converges to 1.

Note that in our context, n refers to the number of path nodes extracted from a single block. In typical transformer architectures, even with a single attention head, the number of path nodes per block is at

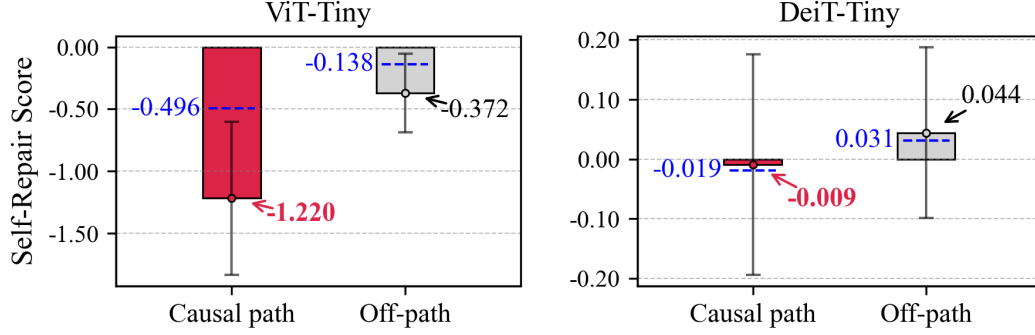


Figure A: **Self-repair scores on causal path vs. off-path components in vision models.** Each bar shows the mean (dot with arrow) and standard deviation (error bar); medians are shown as blue dashed lines. Lower scores indicate less self-repair. Results are averaged over IMAGENET and OFFICEHOME.

least 4 (i.e., $2H + 2$). Empirically substituting such moderate values of n into the expression confirms that the lower bound still converges to 1, even when n is not particularly large.

Furthermore, since the reliability lower bound is monotonic in p , it converges to 1 regardless of the specific value of $p \in (0, 1)$. In addition, because reliability represents a probability bounded between 0 and 1, the entire expression is also upper bounded by 1. Accordingly, the reliability satisfies the following tight bound:

$$p + (1 - p) \cdot \left(1 - \left(1 - \frac{s_{\text{avg}}}{n}\right)^k\right)^n \rightarrow 1$$

In conclusion, using the union of causal node sets from the previous block as the causal subpath reference in the current block yields a reliability that consistently converges to 1. \square

E Implementation Details and Additional Results

E.1 Common Experimental Setup

The code is attached as part of the supplementary material and is also accessible through an anonymized link¹. All experiments were conducted using an RTX Quadro 6000 24GB GPU with PyTorch 2.1.0 and CUDA 11.8. For three language models (GPT2-xs², Pythia-14m³, and Pythia-1b⁴) and two vision models (ViT-tiny and DeiT-tiny⁵), the internal decision paths were traced while keeping all pretrained parameters frozen. Note that for the OFFICEHOME dataset, only the classifier was fine-tuned to match the number of classes. In the case of Pythia, since it adopts a parallel residual structure where the block input is directly fed into both the attention and MLP components, unfolding is adapted accordingly to reflect this architecture. Since a decision path is considered meaningful only if the model produces the correct output for a given input, we sampled subset of the evaluation set that were correctly predicted; but the same evaluation set was used across all methods to ensure a fair comparison.

When using our proposed method with TOKEN RESAMPLING, to prevent incorrect interventions caused by coincidentally selecting token embeddings similar to the original input, we sampled 100 different resampling batches and computed the causal evaluation as the average output across these interventions. For D_α and D_β in each block, we computed their values using the original block input when implementing the unfolded path nodes. Following the identification of causal paths, we utilized their union for further analysis, as justified by Theorem 2.

¹<https://tinyurl.com/neurips15566>

²<https://huggingface.co/AlgorithmicResearchGroup/gpt2-xs>

³<https://huggingface.co/EleutherAI/pythia-14m>

⁴<https://huggingface.co/EleutherAI/pythia-1b>

⁵<https://github.com/huggingface/pytorch-image-models>

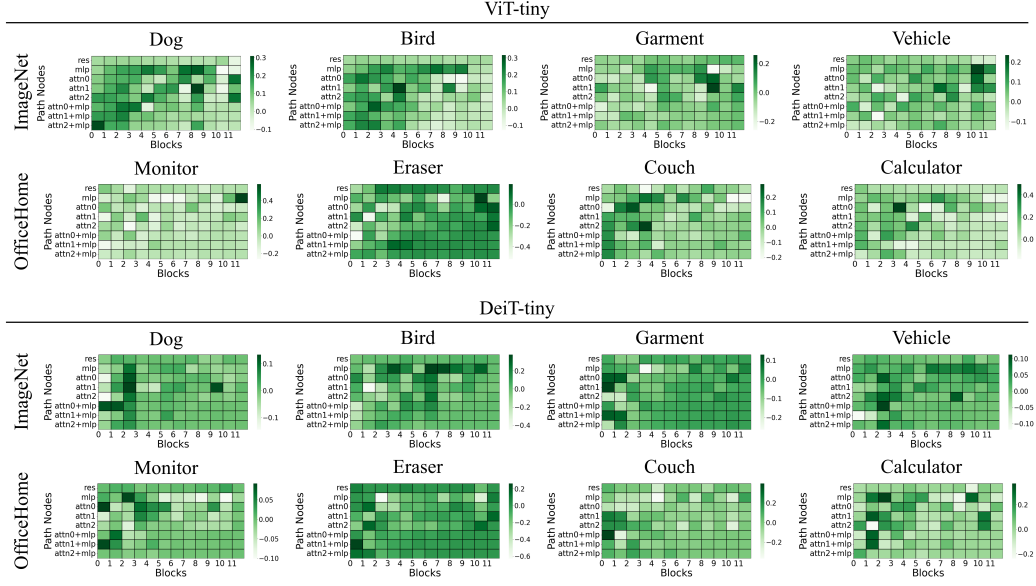


Figure B: **Class-specific causal path patterns.** The class-wise difference in average causal path ratios between a target class and all classes, highlighting class-specific paths. Here, res, mlp, and attn# indicate residual, MLP, and attention paths from head #, respectively.

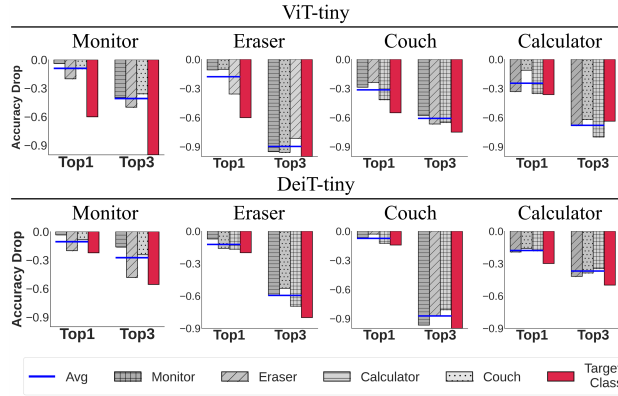


Figure C: **Causal paths uniquely activated for specific classes on OFFICEHOME.** Accuracy drop when ablating the most class-specific path, showing selective reliance by each class.

898 E.2 Self-repair Analysis in Vision Models

899 **Setup** Following prior work [16], we measure the self-repair score of each attention head across
900 layers. A head is considered to be on the causal path if it is included in any causal path at least once;
901 otherwise, it is classified as off-path.

902 **Results** Extending the findings from the main script on language models, we examine whether the
903 same pattern holds in vision models. As shown in Figure A, although self-repair behavior is inherently
904 noisy, as previously reported, we still observe consistently lower self-repair scores on components
905 included in the causal path. This suggests that, even in vision models, the identified causal paths
906 correspond to components that play an essential and non-redundant role in the decision-making
907 process of the model.

908 E.3 Class-specific Path Discovery

909 **Setup** We conducted experiments by removing the largest activated causal subpath (Top-1 in
910 Figures 4 and B) or the top three largest ones (Top-3 in Figures 4 and B) for a given super-class. For

	Model	KNOWN51000			T-REx		
		Hit. (\uparrow)	Faith. (\uparrow)	Spars. (\downarrow)	Hit. (\uparrow)	Faith. (\uparrow)	Spars. (\downarrow)
NT ₁	GPT2-xs	0.0000	0.0004	0.8244	0.0000	0.0005	0.8261
	Pythia-1b	0.0000	0.0001	0.2197	0.0000	0.0001	0.2197
	Pythia-14m	0.0000	0.0009	0.8964	0.0000	0.0012	0.8964
NT _{10%}	GPT2-xs	0.0000	0.0001	0.6960	0.0000	0.0001	0.6870
	Pythia-1b	0.0000	0.0001	<u>0.1503</u>	0.0000	0.0001	<u>0.1503</u>
	Pythia-14m	0.0000	0.0002	0.8526	0.0000	0.0010	0.8526
ET _{all}	GPT2-xs	0.1111	0.1442	0.9868	<u>0.6666</u>	<u>1.0643</u>	0.9868
	Pythia-1b	0.4699	0.2016	0.9998	0.0000	0.0002	0.9629
	Pythia-14m	0.0000	0.0007	0.9737	0.0000	0.0016	0.9737
ET _{cls}	GPT2-xs	0.5556	0.3922	0.9868	0.6078	0.9543	0.9853
	Pythia-1b	<u>0.5904</u>	<u>0.4988</u>	0.9964	0.4686	0.3459	0.9903
	Pythia-14m	0.1429	0.1639	0.9901	0.5195	0.4855	0.9967
CPT (ours)	GPT2-xs	1.0000	0.3742	0.9004	0.9903	0.7260	0.8929
	Pythia-1b	0.9518	0.2593	0.7246	0.9843	0.6192	0.6966
	Pythia-14m	0.9821	0.3857	0.9848	0.9870	0.9218	0.9851

Table A: **Quantitative results (language)**. Detailed results three models on two datasets. The average performance of each method across models and datasets in this table is averaged in Table 2. Hit. and Spars. range between 0 and 1, while Faith. takes non-negative values (i.e., 0 or greater). A Faith. score greater than 1 indicates that the logit for the original decision was amplified by that factor.

	Model	IMAGENET			OFFICEHOME		
		Hit. (\uparrow)	Faith. (\uparrow)	Spars. (\downarrow)	Hit. (\uparrow)	Faith. (\uparrow)	Spars. (\downarrow)
NT ₁	ViT-tiny	0.0014	0.0016	0.7618	0.0140	0.0252	0.7943
	DeiT-tiny	0.0000	0.0019	0.6608	0.0265	0.0258	0.6933
NT _{10%}	ViT-tiny	0.0000	0.0036	0.0873	0.0202	0.0281	0.0721
	DeiT-tiny	0.0000	0.0009	<u>0.0844</u>	0.0109	0.0206	<u>0.0759</u>
ET _{all}	ViT-tiny	0.2816	0.1693	0.9997	0.4022	0.3776	1.0000
	DeiT-tiny	<u>0.5749</u>	<u>0.2833</u>	1.0000	<u>0.5229</u>	<u>0.4361</u>	1.0000
ET _{cls}	ViT-tiny	0.0695	0.0435	0.9573	0.2147	0.1889	0.9291
	DeiT-tiny	0.4561	0.2541	0.9928	0.3105	0.2464	0.9810
CPT (ours)	ViT-tiny	0.9743	0.2099	0.8114	0.9506	0.4835	0.7218
	DeiT-tiny	0.9675	0.1175	0.6812	0.9630	0.3855	0.6975

Table B: **Quantitative results (vision)**. Detailed results two models on two datasets. The average performance of each method across models and datasets in this table is averaged in Table 3. Hit. and Spars. range between 0 and 1, while Faith. takes non-negative values (i.e., 0 or greater). A Faith. score greater than 1 indicates that the logit for the original decision was amplified by that factor.

911 IMAGENET, we selected a subset of 1,000 ImageNet classes and grouped them into four super-classes
912 based on WordNet semantic similarity. The defined superclasses are as follows: **dog**, which includes
913 Maltese, beagle, Saluki, Siberian husky, and golden retriever; **bird (oscines)**, consisting of brambling,
914 robin, jay, and chickadee; **garment**, comprising kimono, cardigan, T-shirt, and sweatshirt; and **vehicle**,
915 which includes cab, minibus, moving van, police van, and school bus. In contrast, in OFFICEHOME,
916 the visual dissimilarity between classes is too large to form meaningful super-classes, so each class
917 was treated as its own super-class.

Results The activation ratio patterns of the causal paths for each super-class can be seen in Figure B. The stronger the color contrast, the more exclusively the corresponding causal subpath is activated for that specific super-class. Furthermore, similar to the main script results, Figure C shows that in OFFICEHOME, ablating the most activated causal subpath for each super-class leads to a significantly larger accuracy drop for that super-class compared to the average drop across others. This indicates the existence of causal subpaths that are specifically responsible for class-specific decisions. However, when more causal subpaths are ablated, the performance drop becomes more uniform, suggesting that, since each class is treated as its own super-class, the model does not rely heavily on a large number of unique subpaths per class. In other words, critical components for fine-grained decisions appear to be more sparsely encoded within the model.

E.4 Detailed Quantitative Comparisons

Setup To adapt the baselines to a decision-focused path tracing framework, we explored the model internals in a backward chaining manner and assigned each discovered decision path to one of the following types: residual-only, attention-only, MLP-only, or attention+MLP-only.

Let y with decision c^* denote the original model output, and let y' with decision c'^* be the output when all components not on the selected path (as determined by each method) are pruned. For **Hit.** (hit rate), we measured the proportion of the dataset where $c^* = c'^*$. For **Faith.** (faithfulness), we measured the average ratio $y'(c^*)/y(c^*)$ across the dataset. For **Spars.** (sparsity), we computed the average ratio of FLOPs used by the components along the selected decision path to the total FLOPs of the original model. Note that even if an attention-only path is not selected, if an attention+MLP path is chosen, the attention layer’s FLOPs are included in the calculation.

Results The results across models and datasets—those averaged in Tables 2 and 3 of the main script—are presented in Tables A and B. As previously mentioned, our method uniquely achieves a near-perfect hit rate, reliably identifying the causal path responsible for the original decision. While some individual cases show relatively higher faithfulness scores under other methods, the consistently low hit rates point to a critical limitation: even if a method increases the logit value for the original decision index, the fact that the final prediction changes implies that it has increased the logits for other indices even more. This means that the identified path is not truly associated with the intended decision, but rather explains a different one—leading to a misattribution. In fact, when examining the sparsity scores, these methods often utilize nearly the entire internal structure of the original model, suggesting minimal pruning and limited explanatory focus.

For instance, ET_{all} applied to GPT2-xs on T-REX marginally increased the logit of the original decision index (1.0643). However, the hit rate remained at 0.6666, indicating that the method amplified other logits even more and ultimately changed the decision. Similarly, for DeiT-tiny on IMAGENET, both ET_{all} and ET_{cls} may appear to better recover the original logit value compared to ours. However, while our method explicitly selects a decision path that preserves the original decision (achieving a hit rate close to 1), their selected paths maintain the original decision in only about half the cases. Notably, these results occur despite their extremely high sparsity scores, indicating that almost the entire model structure was utilized.