DISTRIBUTIONAL CAUSAL PROBING OF VISION MODELS VIA SELF-CONTROLLED INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Occlusion studies often summarize model responses by average performance changes, obscuring heterogeneous effects across images and their relationship to representation changes. We introduce Self-Controlled Intervention, a paired design for probing frozen vision models with deterministic interventions at prespecified doses and directions. Building on this design, we establish a statistical framework that characterizes heterogeneous decision effects, representation changes, and their dependence. We conduct extensive numerical experiments in image classification and object detection to evaluate the framework. The results reveal substantial effect heterogeneity and task-dependent directional vulnerability. These findings challenge robustness assessments based on average performance and caution against inferring decision-level vulnerability from representation drift alone.

1 INTRODUCTION

Understanding how vision models respond to partial loss of visual input is important for assessing robustness and interpreting model predictions. Occlusion provides a direct experimental approach in which selected image regions are masked and the resulting predictions are compared with those for the original image (Zeiler & Fergus, 2014; Naseer et al., 2021). By systematically manipulating the location and extent of occlusion, researchers can evaluate model behavior under different spatial patterns and degrees of visual disruption.

The most relevant literature evaluates vision models from three perspectives, ranging from end-to-end performance to internal representations. Occlusion benchmarks measure task performance under varying degrees of visual degradation (Modi et al., 2023; Wei et al., 2025). Visual attribution methods aim to explain individual predictions using gradients, perturbations, token transformations, or concepts that link model outputs to input regions (Selvaraju et al., 2017; Sundararajan et al., 2017; Fong & Vedaldi, 2017; Petsiuk et al., 2018; Wu et al., 2024; Wang et al., 2024). Some perturbation-based methods further assess feature necessity and sufficiency (Chen et al., 2024). Representation analyses compare learned features across layers, models, or conditions (Raghu et al., 2017; Kornblith et al., 2019; Jiang et al., 2025), while intervention-based representation learning isolates semantic factors for controllable image editing (Huang et al., 2025). These approaches address task-level robustness, input-level dependencies, and internal model representations from different perspectives.

Standard evaluation paradigms typically rely on aggregate performance metrics, which provide an incomplete account of an intervention. For instance, a single average effect can conceal substantial heterogeneity. It may reflect either modest shifts across images or severe losses within vulnerable subgroups, and this may further depend on occlusion direction. Furthermore, tracking representation drift in isolation fails to establish its relationship to decision changes. These issues motivate a distributional analysis that connects decision and representation responses within a common intervention design. Interpreting these responses presents a fundamental difficulty, as masking may remove task-relevant information while simultaneously shifting inputs away from the underlying data distribution (Zhu et al., 2026). Crucially, the total observed response alone cannot disentangle the effect of information loss from that of distribution shift.

To address these issues, we introduce Self-Controlled Intervention (SCI), a paired design for probing frozen vision models through deterministic interventions at prespecified doses and directions. Exact replication allows each image to serve as its own control, making its total intervention effect directly

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.

observable. Statistical inference then concerns the distribution of these effects across the image population. Building on this design, we develop a framework that jointly characterizes decision effects, representation changes, and their statistical dependence within fixed intervention settings. For classification, a placebo extension formalizes the conditions under which comparisons with background occlusion separate the mean contributions of information loss and intervention artifacts.

We evaluate the framework through extensive experiments in image classification and object detection. The results reveal substantial heterogeneity across images and show that decision and representation responses can follow distinct dose trajectories despite statistical dependence. These findings show how average performance can conceal vulnerability and why representation drift alone can be insufficient to characterize decision effects.

Contributions. Our key contributions are as follows: (i) We introduce SCI, a paired design for frozen vision models that makes image-level total intervention effects directly observable and enables distributional analysis across prespecified doses and directions. (ii) We develop a statistical framework linking decision effects, representation changes, and their dependence, and formalize conditions for separating the mean effects of information loss and intervention artifacts using placebo comparisons. (iii) We evaluate SCI through extensive experiments with ten classifiers and two detectors, revealing heterogeneous effects and task-dependent directional vulnerability. Divergent decision and representation dose trajectories caution against inferring decision vulnerability from representation drift alone.

Roadmap. Section 2 develops the design, Section 3 applies it to classification and detection, and Section 4 discusses the implications for occlusion-based evaluation. The appendix provides the complete operator specification, formal results, per-cell estimates, robustness checks, and limitations.

2 METHOD

We first introduce the SCI design and instantiate it with structured occlusion (Section 2.1). The resulting paired observations allow us to characterize decision effects, representation changes, and their dependence (Section 2.2). To interpret the mean decision effect, we introduce a placebo comparison that separates signal and artifact contributions under explicit assumptions (Section 2.3).

2.1 THE SELF-CONTROLLED INTERVENTION DESIGN

SCI uses each image as its own control, comparing a frozen model's responses before and after a specified image intervention. Let $X_1, \ldots, X_n$ be images from a population P, with their class labels or object annotations held fixed throughout evaluation. We include these annotations in the observation X_i and suppress them in the notation below. Let $\mathcal{I}$ denote a deterministic operator that modifies the pixels of image X at dose $t \in [0, 1]$ and direction $d \in \mathcal{D}$, where $\mathcal{D}$ is a finite prespecified set. The clean image satisfies $\mathcal{I}(X, 0, d) = X$ for every direction. Throughout this work, dose is defined geometrically: it specifies the fraction of support pixels replaced, while direction determines their order. Because direction controls both the starting location and the sequence of replacement, the same nominal dose can affect different visual content across images. This inherent variation motivates the distributional analysis that follows.

We record both the task response and an internal representation to examine the intervention at two levels. Let g_θ return the chosen D-dimensional representation and let f return the scalar task response, evaluated using the unchanged annotations. Define

$$m_i(t, d) = g_\theta[\mathcal{I}(X_i, t, d)], \quad Y_i(t, d) = f[\mathcal{I}(X_i, t, d)], \quad \tau_i(t, d) = Y_i(t, d) - Y_i(0). \tag{1}$$

The baseline quantities $m_i(0)$ and $Y_i(0)$ are shared across directions. Classification uses the logit for the ground-truth class, and detection uses an image-level soft recall that combines confidence and localization (equation 9, Section 3.2).

Each image is evaluated at every point of the prespecified intervention grid. Exact replication, together with a deterministic operator and evaluation pipeline, makes both responses in the paired contrast observable (Holland, 1986). The difference $\tau_i(t, d)$ therefore directly measures the total effect of the specified intervention on that image's response.

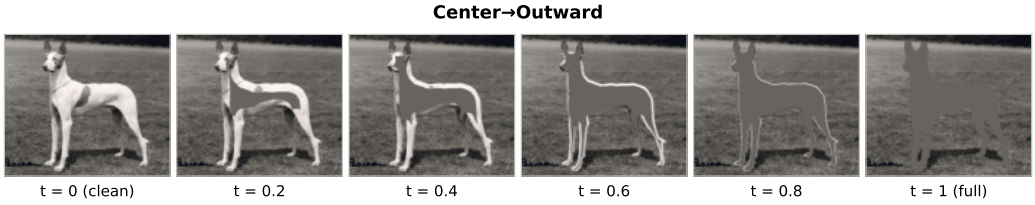


Figure 1: **Center→Outward occlusion at six doses.** An increasing fraction of pixels is replaced within the foreground support, beginning with those farthest from its boundary. The support stays fixed, and the replacement color is computed from the original image.

To examine how responses depend on the amount and spatial pattern of occlusion, we replace pixels within a foreground support. Dose t specifies the fraction of support pixels replaced, and direction d specifies their order. For each image, the support mask, pixel ordering, and replacement color are fixed prior to evaluating the dose grid. This process is illustrated in Figure 1 using Center→Outward occlusion as an example.

The support is obtained by dilating the foreground mask. Center→Outward orders pixels by decreasing distance from its boundary, while horizontal and vertical directions order them by image coordinates. At dose t, the operator replaces a prefix containing the prescribed fraction of support pixels. These prefixes are nested, so increasing the dose retains every previously occluded pixel. Replaced pixels receive a weighted average of colors sampled along the support boundary, computed once from the original image. We apply this construction to both tasks. Classification uses a three-pixel dilation and four directions (Appendix B.2), and detection uses instance-relative supports, scale-dependent dilation, and five directions (Appendix D.3). In both settings, the operator preserves image coordinates and changes pixel values within the chosen support.

2.2 Three Levels of Statistical Inference

With the model and operator fixed, equation 1 applies equally to a random image $X \sim P$. Let $Y(t,d) = f[\mathcal{I}(X,t,d)]$ and $\tau(X;t,d) = Y(t,d) - Y(0)$ be the population version. Then the realized effect $\tau_i(t,d) = \tau(X_i;t,d)$ varies across images drawn from P. The mean alone can conceal large changes concentrated in a subset of images, so we begin by characterizing the distribution of these paired effects.

Let $F_{\tau(t,d)}$ denote the distribution function of $\tau(X;t,d)$ under P. The average treatment effect (ATE) and quantile intervention effects (QIEs) are defined as

$$\tau(t,d) = \mathbb{E}_P[\tau(X;t,d)], \qquad \tau_q(t,d) = F^{-1}_{\tau(t,d)}(q), \quad q \in (0,1). \tag{2}$$

We estimate the ATE by $\hat{\tau}_n(t,d) = n^{-1}\sum_{i=1}^n \tau_i(t,d)$ and the QIEs by empirical quantiles of $\{\tau_i(t,d)\}_{i=1}^n$. Low quantiles describe the largest response decreases, whereas high quantiles reflect mild decreases or potential increases in the target logit. The QIE is defined directly as a quantile of paired individual-level effects. In contrast, the conventional quantile treatment effect merely compares marginal quantiles, $F^{-1}_{Y(t,d)}(q) - F^{-1}_{Y(0)}(q)$ (Firpo, 2007). Because these marginal distributions cannot uniquely identify the distribution of individual treatment effects (Heckman et al., 1997; Caughey et al., 2023), SCI preserves the image-level pairing necessary to estimate this distribution.

The paired design also allows us to examine how the model's internal representation changes under the same intervention. To account for the pairing, we evaluate population representation drift using a paired U-statistic for the squared maximum mean discrepancy (MMD^2) (Gretton et al., 2012). For each image i, we define the feature change as

$$\Delta m_i(t,d) = m_i(t,d) - m_i(0), \tag{3}$$

and quantify its vector shift using the Euclidean displacement magnitude $\delta_i(t,d) = \|\Delta m_i(t,d)\|_2$ alongside the angular deviation $c_i(t,d) = 1 - \cos[m_i(0), m_i(t,d)]$. Descriptive MMD^2 trajectories use a fixed clean-reference bandwidth within each model. Appendix B.1 specifies the separate swap-invariant bandwidth used for randomization calibration. We further evaluate subspace departure and

local intrinsic dimension using principal component analysis (PCA) residuals and the Two-Nearest Neighbors (Two-NN) estimator (Facco et al., 2017). Complete details on feature preprocessing and estimators appear in Appendix D.2.

Having characterized the decision and representation responses separately, we next examine how they are related across images. For each (t, d), the Hilbert–Schmidt Independence Criterion (HSIC) (Gretton et al., 2005; 2008) evaluates dependence between the representation-change vectors $\Delta m_i(t, d)$ and decision effects $\tau_i(t, d)$, using a common baseline feature basis. We also report normalized HSIC to describe association strength. Keeping dose and direction fixed separates within-setting dependence from association induced by a shared dose trend. The kernel, projection, and testing procedures are detailed in Appendices D.2 and B.1.

Association at a fixed dose leaves open how the two response levels evolve as the dose increases. To compare their trajectories, we divide $|\hat{\tau}_n(t, d)|$ and estimated MMD^2 by their respective nonzero values at full occlusion ($t = 1$). The area under the normalized curve (AUC) summarizes its shape over $[0, 1]$, while the half-dose t_{50} is the first interpolated dose at which the curve reaches 0.5. A linear trajectory has an AUC of 0.5. For monotone trajectories, a larger AUC indicates an earlier response, corresponding to a smaller t. We call a trajectory *brittle* when its AUC is below 0.5, indicating that most of the change occurs at higher doses, and *elastic* when its AUC is above 0.5, indicating an earlier response followed by slower change. These terms describe trajectory shape only and do not imply physical reversibility.

Deterministic evaluation directly yields individual effects $\tau_i(t, d)$, from which population properties are consistently estimated under standard sampling inference. Propositions B.1 and B.2 establish population identification and asymptotic normality for the sample mean effect. In Corollary B.1, we quantify the precision gain achieved by pairing under positive covariance. To evaluate this uncertainty in practice, we compute confidence intervals via bootstrap resampling, sampling images in classification and city clusters in detection, while keeping paired observations intact within each unit. Table B.1 summarizes all estimands, with full details on uncertainty quantification procedures and hypothesis testing protocols provided in Appendix B.1.

2.3 A Placebo-Based Decomposition of Mean Effects

While the preceding analyses quantify the total occlusion response, this aggregate measure conflates genuine information loss with sensitivity to input distribution shifts (Hooker et al., 2019; Zhu et al., 2026). Disentangling these two factors requires a baseline that isolates pure distributional artifacts. To this end, we evaluate a reference perturbation that preserves task-relevant information, denoting its mean artifact response as $\tau_{\mathrm{ood}}(t, d)$. Subtracting this artifact baseline from the total response isolates the net signal component:

$$\tau_{\mathrm{sig}}(t, d) := \tau(t, d) - \tau_{\mathrm{ood}}(t, d). \tag{4}$$

While conceptually intuitive, this additive decomposition is fundamentally ill-posed without additional structural constraints. We formalize this non-identifiability result below.

Proposition 1 (Non-identification of the mean decomposition under SCI alone)**.** *The paired responses identify the total effect $\tau(t, d)$ but do not separately identify $\tau_{\mathrm{sig}}(t, d)$ and $\tau_{\mathrm{ood}}(t, d)$. Without restrictions on the artifact reference, every pair summing to $\tau(t, d)$ is compatible with the observed total response.*

To assess the artifact contribution in classification, we introduce a background-occlusion placebo. Its operator $\mathcal{I}_0(X, t, d)$ uses the same directional ordering, fill rule, and support-relative dose as the foreground intervention. Its mean effect is $\tau^0(t, d) = \mathbb{E}_P\{f[\mathcal{I}_0(X, t, d)] - f(X)\}$. The shared clean baseline cancels in the image-level contrast,

$$\tilde{\tau}_i(t, d) = f[\mathcal{I}(X_i, t, d)] - f[\mathcal{I}_0(X_i, t, d)], \tag{5}$$

whose population mean is $\tilde{\tau}(t, d) = \tau(t, d) - \tau^0(t, d)$. Empirical quantiles of $\{\tilde{\tau}_i(t, d)\}$ describe heterogeneity in the observed placebo contrasts. To interpret the mean contrast $\tilde{\tau}(t, d)$ as a signal component, write the analogous placebo decomposition as $\tau^0 = \tau^0_{\mathrm{sig}} + \tau^0_{\mathrm{ood}}$ and impose the following conditions on the two arms.

Assumption 1 (Zero mean signal effect of the placebo)**.** *At the specified (t, d), the placebo has zero mean signal component $\tau^0_{\mathrm{sig}}(t, d) = 0$, so $\tau^0(t, d) = \tau^0_{\mathrm{ood}}(t, d)$.*

Assumption 2 (Common mean artifact response)**.** *At the specified (t, d), the intervention and placebo have equal mean artifact responses, that is, $\tau_{\mathrm{ood}}(t, d) = \tau^0_{\mathrm{ood}}(t, d)$.*

These conditions ensure that the mean placebo effect is equal to the intervention's mean artifact component.

Proposition 2 (Placebo identification of the mean signal effect)**.** *Under Assumptions 1–2, we have*

$$\tau_{\mathrm{sig}}(t, d) = \tilde{\tau}(t, d) = \tau(t, d) - \tau^0(t, d). \tag{6}$$

This identification holds at the population mean level, leaving the unit-level decomposition unspecified.

By Proposition 2, $\tau_{\mathrm{sig}}(t, d)$ can be consistently estimated by the plug-in difference $\hat{\tau}_n(t, d) - \hat{\tau}^0_n(t, d)$.

Remark 1 (Sensitivity to reference assumptions)**.** *When the placebo assumptions are relaxed, the bias of the placebo-identified estimand relative to the true signal decomposes as*

$$\tilde{\tau}(t, d) - \tau_{\mathrm{sig}}(t, d) = \left\{\tau_{\mathrm{ood}}(t, d) - \tau^0_{\mathrm{ood}}(t, d)\right\} - \tau^0_{\mathrm{sig}}(t, d). \tag{7}$$

The two terms distinguish artifact mismatch from information loss under the placebo. Their signs determine the direction of bias.

To assess the empirical comparability of the two arms, we evaluate their departure from the same clean-reference feature distribution. For each image, the atypicality score $\rho \in [0, 1]$ is the clean-reference percentile of its feature-space nearest-neighbor distance, with larger values indicating greater departure from the clean representation distribution. Let $\bar{\rho}(t, d)$ and $\bar{\rho}_0(t, d)$ denote the mean scores in the intervention and placebo arms, respectively. We summarize their discrepancy by

$$\Delta_\rho(t, d) = \left|\bar{\rho}(t, d) - \bar{\rho}_0(t, d)\right|. \tag{8}$$

A smaller $\Delta_\rho(t, d)$ indicates that the two operators induce similar average departures from the clean feature distribution. We use $c = 0.08$ as a descriptive threshold and highlight cells with $\Delta_\rho(t, d) \leq c$. Appendix C.3 reports the complete dose-wise score and gap trajectories.

As a complementary check, model and head randomization evaluate how directional patterns depend on learned parameters (Adebayo et al., 2018), as detailed in Appendix C.2.

3 Experiments on Heterogeneous Vulnerabilities

We evaluate our method on two target tasks: classification and detection. We start with classification to assess how well the model isolates signal from artifacts, and then move to detection to examine its transferability. We construct cluster-bootstrap percentile intervals over images for classification and over cities for detection. The confidence intervals are pointwise for most estimands. The headline QIE fan uses a simultaneous confidence band over the dose–quantile grid, while the placebo-contrast fans in Figure C.2 use simultaneous bands over the dose grid. All directional testing strictly adheres to a fixed protocol, with full details provided in Appendix B.1.

3.1 Image Classification

We select ten classes from the mini-ImageNet validation split (Vinyals et al., 2016) to cover objects with different spatial arrangements of distinctive parts. We use SAM-3 (Carion et al., 2025) to obtain foreground masks and retain the 2,858 images for which it produces a usable mask. Because the excluded images have lower baseline performance, our results describe the subset with usable masks. All the details are reported in Appendix D.1. Each retained image is evaluated at the clean baseline and at ten nonzero doses from $t = 0.02$ to $t = 1$. We evaluate four intervention directions, encompassing Center→Outward and three peripheral paths (Left→Right, Bottom→Top, and Top→Bottom). The same images are used throughout the grid, so all comparisons remain paired across intervention settings. We evaluate ten architectures pretrained on ImageNet (Deng et al., 2009), including convolutional networks and vision Transformers (He et al., 2016; Simonyan & Zisserman, 2015; Szegedy et al., 2015; Tan & Le, 2019; Howard et al., 2019; Liu et al., 2022; Dosovitskiy et al., 2021; Liu et al., 2021b). We conduct the representation-level analyses on four representative architectures: MobileNet-V3-L, ResNet-50, ConvNeXt-Base, and ViT-B/16. We measure

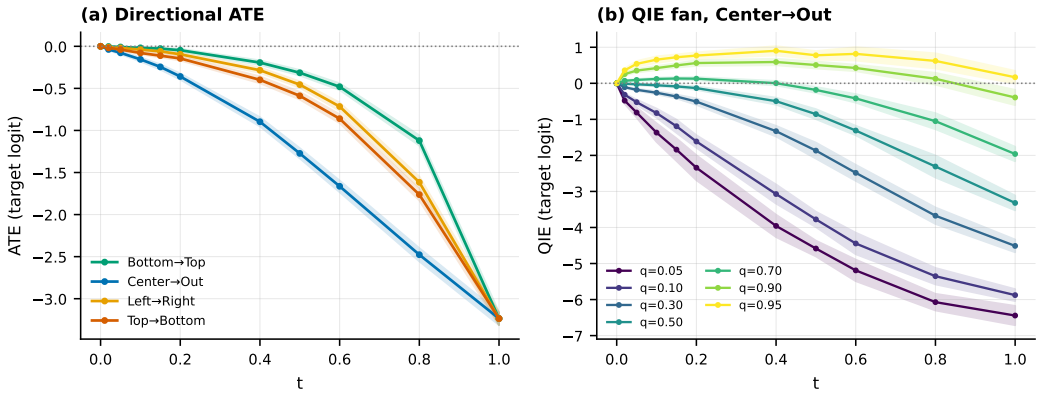


Figure 2: **Decision-level dose–response of ResNet-50.** (a) Directional ATE trajectories: a stable hierarchy emerges with dose (curves coincide at $t = 1$ by construction). (b) QIE fan, Center→Outward: the quantile spread widens monotonically with dose, reflecting growing cross-image heterogeneity. Bands are simultaneous 95% sup-t bands over the dose–quantile grid (Montiel Olea & Plagborg-Møller, 2019); uncertainty procedures are detailed in Appendix B.1.

the model's output using the logit of the correct class as our primary metric, while also reporting top-1 accuracy for reference.

We first examine how the ground-truth-class logit changes under occlusion. Figure 2a shows the ResNet-50 ATE trajectories, and results for all ten architectures appear in Table D.2 of Appendix D.1. These results indicate that the ATE remains strictly negative at all non-zero doses across all models and directions, with its magnitude increasing monotonically with dose. Beyond the average effect, the QIE fans in Figure 2b highlight substantial image-level heterogeneity. While the lower quantiles decline steeply as dose increases, the upper quantiles remain positive at low-to-intermediate dose levels under the simultaneous confidence band (Appendix B.1). This indicates that logits can actually increase for certain images even when the mean effect is negative. Upon full occlusion, the upper tail converges toward zero, whereas the lower tail stays markedly negative.

To interpret the mean response, we compare foreground occlusion against the background placebo from Section 2.3, using the paired contrast $\tilde{\tau}_i(t, d)$ in equation 5 at the same nominal dose and direction. Figure C.3 presents the mean atypicality scores across both arms, with full evaluations detailed in Appendix C.3. Across all evaluated models and directions, the atypicality gap remains strictly below 0.08 up to $t = 0.05$. As dose increases, placebo scores plateau near the clean median, while foreground scores continue to rise. Consequently, higher doses increasingly fail the mean-atypicality diagnostic introduced in Section 2.3.

Figure 3 illustrates the corresponding treatment contrasts. At low doses under Center→Outward occlusion, the mean foreground–background contrast stays near zero or positive despite a negative total foreground effect. If Assumptions 1–2 are satisfied, this pattern suggests a substantial artifact contribution to the mean response. Nevertheless, the lower decile of paired contrasts remains negative, showing that certain images experience a larger logit drop under foreground occlusion than under background occlusion even when the mean difference is small. At higher doses, these contrasts drop further into negative territory as the atypicality gap widens. Following Remark 1, we report these results as sensitivity contrasts between the two operators.

We next compare the dose trajectories of the total foreground effect across directions. Following Section 2.2, nearly all architectures exhibit brittle decision-level responses under Center→Outward occlusion, decaying primarily at higher dose levels. GoogLeNet and VGG-16-BN are the two elastic exceptions, dropping earlier and more gradually. Across the three peripheral directions, brittle behavior holds for 29 of the 30 architecture–direction trajectories. Here, the sole exception is VGG-16-BN under Top→Bottom, whose nearly linear trajectory falls just above the AUC $= 0.5$ threshold and is therefore classified as elastic (AUC $= 0.53$; Table D.3).

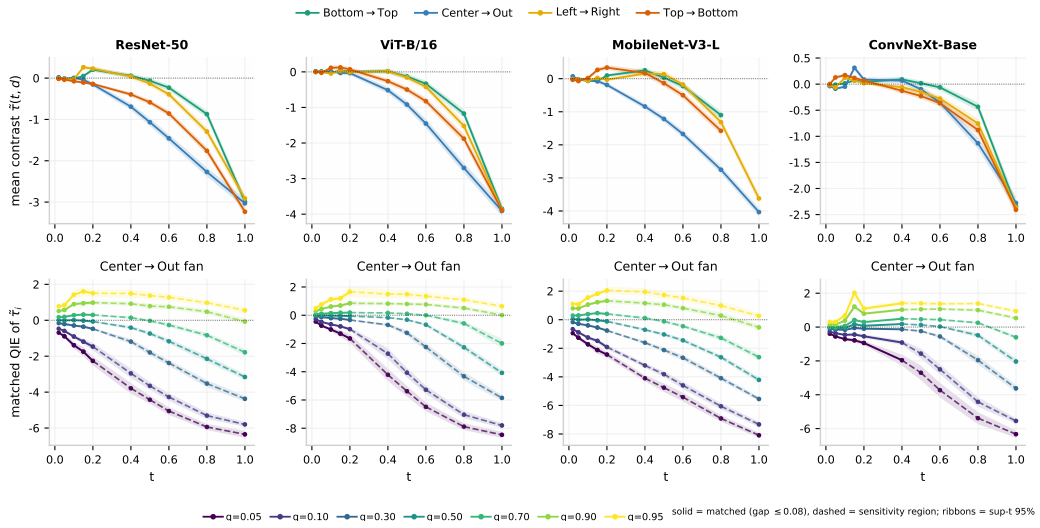


Figure 3: **Paired placebo contrasts, four architectures.** *Top*: mean contrast $\tilde{\tau}(t, d)$ by dose and direction. *Bottom*: empirical quantiles (q_{05}–q_{95}) of $\tilde{\tau}_i(t, d)$ under Center→Outward (all directions: Appendix C.3). Solid segments denote cells with a mean atypicality gap at most 0.08; dashed segments denote the remaining sensitivity region. Bands are 95% cluster-bootstrap intervals.

We next compare directional vulnerability by the magnitude of the mean logit decrease. As illustrated in Figure 2a, Center→Outward emerges as the most damaging configuration for ResNet-50 across nearly all dose levels prior to full occlusion. At $t = 0.8$, all ten architectures exhibit the same damage hierarchy, ordered from largest to smallest effect as Center→Outward, Top→Bottom, Left→Right, and Bottom→Top (Table D.2). Category-level differences in center-sensitivity align with the spatial distribution of distinctive object features, with the center-concentration ratios reported in Table D.1 and the corresponding category-level dose trajectories shown in Figure D.1. In contrast, the detection experiments in Section 3.2 exhibit a distinctly different ranking, indicating that directional vulnerability is task-dependent.

To examine the source of this directional structure, we apply random occlusion configurations that replace the same number of pixels within the same foreground support and use the same fill rule. As illustrated in Figure D.2, variation across directional ATEs substantially exceeds that across random configurations in nearly all model–dose settings, even though the directional effects remain within the random-control range. We further evaluate whether these differences depend on learned parameters by randomizing the model backbone or classification head. As shown in Table C.3, randomization reduces baseline accuracy to chance level, substantially diminishes full-occlusion effects, and collapses directional variance across both architectures. Together, these control experiments support a contribution from learned parameters to the observed directional differences, beyond the raw number of pixels replaced.

Finally, we examine representation drift in the four selected architectures and relate it to the decision effects. Using the within-image swap procedure described in Appendix B.1, we reject the exchangeability null in all 148 distinct intervention settings after counting each model's direction-independent full-occlusion endpoint once. To compare trajectory shapes, we normalize each model's representation drift by its value at full occlusion following Section 2.2. As illustrated in Figure D.3, ViT-B/16 is the sole architecture exhibiting an elastic representation trajectory under Center→Outward occlusion, whereas all four architectures display brittle behavior under peripheral directions. Crucially, these representation trajectories can decouple from decision-level responses. As summarized in Table D.4, ViT-B/16 exhibits elastic representation drift alongside a brittle logit response, whereas ConvNeXt-Base remains brittle at both levels.

Figure D.5 characterizes the geometry of the representation drift. PCA residuals reveal that intervened representations move systematically away from the principal subspace fitted on clean features. At full occlusion, the normalized residual reaches $2.80\times$ for ResNet-50, $2.16\times$ for ViT-B/16, $2.14\times$ for ConvNeXt-Base, and $1.67\times$ for MobileNet-V3-L. Two-NN estimates indicate no sys-

tematic contraction in local intrinsic dimension, with moderate expansion for some architectures. To assess the relation to decision effects, we compute HSIC between representation-change vectors $\Delta m_i(t, d)$ and paired logit effects $\tau_i(t, d)$ at fixed doses and directions. The permutation test rejects independence in all 148 distinct intervention settings after within-model BH adjustment, though normalized HSIC trajectories vary across architectures (Figure D.4). Together, these findings confirm that representation drifts and decision effects are associated across images within fixed intervention settings, though this association alone does not establish that representation drift fully accounts for the decision response.

3.2 Object Detection

We next examine whether the directional rankings and dose–response shapes observed in classification carry over to object detection. We use the combined training and validation splits of Cityscapes (Cordts et al., 2016), retaining 3,457 paired images containing at least one of the seven object classes shared with COCO. These images span 21 city clusters. The two detectors are Mask R-CNN, a two-stage convolutional model (He et al., 2017), and Mask DINO, a query-based transformer (Li et al., 2023). Both are evaluated zero-shot, without adaptation to Cityscapes. The five directions comprise the four used in classification plus Right→Left, and the five doses are $t \in \{0.2, 0.4, 0.5, 0.6, 0.8\}$. All the details are described in Appendix D.3.

Our primary response is soft recall, which combines detection confidence and localization. For each image, let G_i be its nonempty set of annotated objects and let $\hat{b}_j$ denote a detection returned on $\mathcal{I}(X_i, t, d)$. Write cls for the class label, conf for detection confidence, and IoU for intersection-over-union overlap. For each annotated object $g \in G_i$, we define a score $s(g; t, d)$ and average these scores to obtain the image-level response:

$$Y_i(t, d) = \frac{1}{|G_i|} \sum_{g \in G_i} s(g; t, d), \qquad s(g; t, d) = \max_{j:\, \mathrm{cls}(\hat{b}_j) = \mathrm{cls}(g)} \mathrm{conf}(\hat{b}_j)\, \mathrm{IoU}(\hat{b}_j, g). \tag{9}$$

Each annotated object receives the largest confidence–overlap product among detections of the same class, or zero if none is returned. We apply no additional confidence threshold to the returned detections, and the same detection may contribute to more than one object. We also check the results using metrics with thresholds and one-to-one matching, together with dataset-level mean average precision (mAP; Appendix D.3).

Figure 4 shows that the ATE remains negative across every evaluated model, direction, and dose level, with losses generally increasing in magnitude as dose increases. The directional ranking differs from that in classification. Horizontal directions dominate from $t = 0.4$ onward in both detectors. At the lowest dose, Bottom→Top produces the largest mean loss for Mask R-CNN, whereas Left→Right is marginally largest for Mask DINO. Center→Outward never leads. Table D.7 reports the complete ATE grid for both detectors.

Table D.8 summarizes the corresponding QIE patterns across object categories. For Car, the upper effect decile is negative in 46 of the 50 evaluated cells, indicating damage across nearly the entire image population. Bicycle, Motorcycle, and Bus retain a zero or positive upper tail, indicating a resistant subpopulation. Person is intermediate between these patterns, with a negative upper decile in 28 cells and a zero or positive upper decile in the remaining 22. Train has a near-zero median ATE but a positive upper decile, whereas Truck has zero estimated effect in all 50 cells.

Table D.9 reports the directional comparisons. Nearly all pairwise comparisons remain significant after adjustment for multiple testing. Under the primary soft-recall response, Bottom→Top produces a larger loss than Top→Bottom in all ten detector–dose settings. To check whether the ranking depends on the response measure, we compare soft recall with mAP computed under the COCO evaluation protocol. Table D.11 shows that the mAP results reproduce three main patterns: Bottom→Top leads at the lowest dose, horizontal directions dominate at intermediate and higher doses, and Center→Outward never leads. The matched-coverage regression in Appendix D.3 shows that directional differences remain after adjustment for the fraction of pixels actually occluded. The exact ordering at higher doses nevertheless depends on the metric.

Table D.5 compares the directional results under macro-F1 and micro-F1. The directional leader depends on the aggregation rule and confidence threshold. Directional vulnerability therefore depends on the task and response measure. The directional reversal may stem from both task and dataset

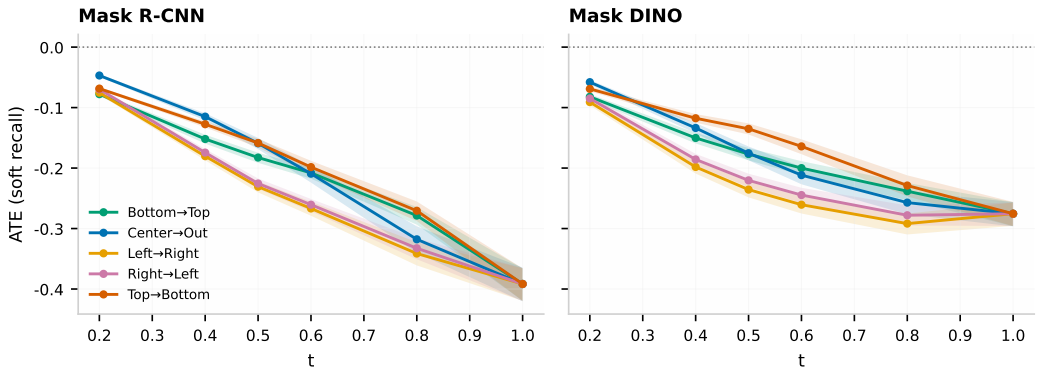


Figure 4: **Detection dose–response under the soft recall.** Directional ATE trajectories, Mask R-CNN (left) and Mask DINO (right): horizontal directions lead from mid-dose onward, while Bottom→Top leads at the lowest dose (Mask R-CNN). Bands: 95% city-cluster bootstrap CIs.

characteristics. Cityscapes driving scenes have strong spatial structure, with road surfaces concentrated near the bottom of the image and vehicles often extending horizontally. These spatial patterns align with the directions that show larger performance drops. Object detection also lacks a placebo control arm that would allow these spatial factors to be separated from the effects of the task. These directional effects therefore combine information loss and sensitivity to distribution shift.

The two detectors also differ in the shapes of their dose–response trajectories. Table D.6 reports the corresponding decision-level AUC and t_{50} values. Mask R-CNN is brittle under Center→Outward and the two vertical directions, but elastic under the two horizontal directions. Mask DINO is elastic in all five directions. The horizontal directions have the largest AUCs and the smallest t_{50} values in both detectors, indicating an earlier decision-level response.

Finally, we relate representation changes to soft-recall effects. Figure D.6 shows that, for Mask R-CNN, the representation-displacement leader shifts from Bottom→Top at low doses to Center→Outward at higher doses. For Mask DINO, Center→Outward produces the largest displacement from $t = 0.4$ onward. These orderings do not reproduce the soft-recall hierarchy in Figure 4. Cluster-aware HSIC tests at fixed dose and direction reject independence in all evaluated settings. Table D.10 reports the corresponding cluster-aware statistics. Despite this statistical dependence, the descriptive nHSIC magnitudes remain small, indicating weak association. We therefore do not classify these representation trajectories as brittle or elastic, given that the detection protocol lacks the full-occlusion endpoint used for normalization in Section 2.2.

4 Discussion and Conclusion

The experiments reveal three features of occlusion-based evaluation. First, mean foreground and background responses can be close even when the foreground response is strongly negative. Under the placebo reference assumptions, this suggests a substantial artifact contribution to the average low-dose center response. The distribution of paired contrasts nevertheless remains heterogeneous. Second, directional rankings differ between the studied classification and detection settings, so a robustness ranking must specify its task, model, and intervention. Third, decision and representation trajectories can have different shapes despite statistical dependence. For example, ViT-B/16 has an elastic representation trajectory and a brittle decision trajectory, and the measured association does not by itself explain this difference.

SCI connects these findings through a common paired design. It makes image-level total effects observable and their population distribution an inferential target. The placebo extension conditionally interprets the mean response, while Remark 1 makes its reference assumptions explicit. Limitations and extensions appear in Appendix E.

Reproducibility Statement

An anonymized repository with all code, pipelines, stored predictions, and analysis scripts accompanies the submission. Every figure and table can be regenerated from the stored artifacts, and all protocols are specified in Section 2 and the appendix.

AI Use Statement

Generative AI tools were used to improve the clarity and readability of the manuscript, assist in discovering potentially relevant literature, and debug research code. All suggested references were independently verified against the original sources. All AI-assisted code was reviewed, tested, and validated by the authors. The tools were not used to formulate the research questions or hypotheses, design the methodology or experiments, generate data, prove mathematical claims, or interpret the experimental results. The authors take full responsibility for the content, results, and conclusions of this work.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9505–9515, 2018.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6111–6122, 2019.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts, 2025. arXiv preprint arXiv:2511.16719.

Devin Caughey, Allan Dafoe, Xinran Li, and Luke Miratrix. Randomisation inference beyond the sharp null: Bounded null hypotheses and quantiles of individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1471–1491, 2023. doi: 10.1093/jrsssb/qkad080.

Xuexin Chen, Ruichu Cai, Zhengting Huang, Yuxuan Zhu, Julien Horwood, Zhifeng Hao, Zijian Li, and Jos'e Miguel Hern'andez-Lobato. Feature attribution with necessity and sufficiency via dual-stage perturbation test for causal explanation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6486–6502. PMLR, 2024.

Chi-Ning Chou, Hang Le, Yichen Wang, and SueYeon Chung. Feature learning beyond the lazy-rich dichotomy: Insights from representational geometry. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10700–10740. PMLR, 2025.

Chi-Ning Chou, Artem Kirsanov, Yao-Yuan Yang, and SueYeon Chung. Diagnosing generalization failures from representational geometry markers. In *International Conference on Learning Representations (ICLR)*, 2026.

Kacper Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/4e382cb49370f64415df2672b19fb1f2-Abstract.html.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017.

Sergio Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75 (1):259–276, 2007.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3429–3437, 2017.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning (ICML)*, pp. 2376–2384, 2019.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Sch"olkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory (ALT)*, pp. 63–77, 2005.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Sch"olkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 585–592, 2008.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Sch"olkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Doll'ar, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, 2017.

James J. Heckman, Jeffrey Smith, and Nancy Clements. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535, 1997.

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

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9737–9748, 2019.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019.

Shanshan Huang, Haoxuan Li, Chunyuan Zheng, Lei Wang, Guorui Liao, Zhili Gong, Huayi Yang, and Li Liu. Visual representation learning through causal intervention for controllable image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23484–23493, 2025.

Jiachen Jiang, Jinxin Zhou, and Zhihui Zhu. Tracing representation progression: Analyzing and enhancing layer-wise similarity. In *International Conference on Learning Representations (ICLR)*, 2025.

Kaleb Kassaw, Francesco Luzi, Leslie M. Collins, and Jordan M. Malof. Are deep learning models robust to partial object occlusion in visual recognition tasks? *Pattern Recognition*, 171:112215, 2026. doi: 10.1016/j.patcog.2025.112215.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Sch"utt, Sven D"ahne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer, 2019.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, pp. 3519–3529, 2019.

Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 777–784, 2004.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask DINO: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3041–3050, 2023.

Xiutian Li, Siqi Sun, and Rui Feng. Causal representation learning via counterfactual intervention. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 3234–3242, 2024.

Hongjiao Liu, Anna M. Plantinga, Yunhua Xiang, and Michael C. Wu. A kernel-based test of independence for cluster-correlated data. In *Advances in Neural Information Processing Systems*, volume 34, 2021a. URL `https://proceedings.neurips.cc/paper/2021/hash/51be2fed6c55f5aa0c16ff14c140b187-Abstract.html`.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.

Rajat Modi, Vibhav Vineet, and Yogesh Rawat. On occlusions in video action detection: Benchmark datasets and training recipes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 57306–57335, 2023.

Jos'e Luis Montiel Olea and Mikkel Plagborg-Møller. Simultaneous confidence bands: Theory, implementation, and an application to SVARs. *Journal of Applied Econometrics*, 34(1):1–17, 2019.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 23296–23308, 2021.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6076–6085, 2017.

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.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, pp. 3319–3328, 2017.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3630–3638, 2016.

Hengyi Wang, Shiwei Tan, and Hao Wang. Probabilistic conceptual explainers: Trustworthy conceptual explanations for vision foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51502–51522. PMLR, 2024.

Wenbo Wei, Jun Wang, and Abhir Bhalerao. COCO-OLAC: A benchmark for occluded panoptic segmentation and image understanding. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Junyi Wu, Bin Duan, Weitai Kang, Hao Tang, and Yan Yan. Token transformation matters: Towards faithful post-hoc explanation for vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10926–10935, 2024.

You Wu, Xucheng Wang, Xiangyang Yang, Mengyuan Liu, Dan Zeng, Hengzhou Ye, and Shuiwang Li. Learning occlusion-robust vision transformers for real-time uav tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17103–17113, 2025.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pp. 818–833, 2014.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13001–13008, 2020. doi: 10.1609/aaai.v34i07.7000.

Zhiyu Zhu, Zhibo Jin, Jiayu Zhang, Bartlomiej Sobieski, Przemyslaw Biecek, Fang Chen, and Jianlong Zhou. Faithfulness under the distribution: A new look at attribution evaluation. In *International Conference on Learning Representations (ICLR)*, 2026.

APPENDIX

This appendix is organized as follows. Appendix A provides extended related work. Appendix B presents the theoretical properties of the self-controlled design, the detailed specification of the occlusion operator, and the placebo identification argument. Appendix C reports the placebo and control experiments. Appendix D provides the complete classification, representation, and detection analyses, together with robustness checks. Finally, Appendix E discusses limitations and directions for future work.

A EXTENDED RELATED WORK

This section reviews four areas related to SCI: post-hoc attribution and perturbation-based explanation, representation geometry and distribution shift, occlusion robustness evaluation, and intervention or causal representation learning.

Post-hoc attribution and perturbation-based explanation. Gradient-based attribution methods such as Grad-CAM (Selvaraju et al., 2017) and Integrated Gradients (Sundararajan et al., 2017) assign importance scores to image features for an individual prediction. Perturbation-based methods instead modify the input and measure the resulting prediction change. Meaningful Perturbation (Fong & Vedaldi, 2017) learns an image-specific mask, while RISE (Petsiuk et al., 2018) aggregates predictions over randomly sampled masks.

Recent work has extended attribution analysis beyond pixel-level importance. TokenTM (Wu et al., 2024) incorporates changes in token representations when explaining Vision Transformers. PACE (Wang et al., 2024) develops probabilistic conceptual explanations for vision foundation models, considering properties such as faithfulness, stability, sparsity, and parsimony. FANS (Chen et al., 2024) formulates perturbation-based attribution in terms of necessity and sufficiency.

The reliability of attribution evaluation also depends on the model and the perturbation procedure. Parameter-randomization and implementation sanity checks have revealed that some attribution results may be insensitive to learned model parameters or affected by implementation choices (Adebayo et al., 2018; Kindermans et al., 2019). Insertion and deletion evaluations may additionally move masked images away from the data distribution. FUD (Zhu et al., 2026) addresses this issue by using score-based diffusion models to reconstruct masked regions more realistically.

These studies motivate careful specification of both the perturbation and the response being measured. SCI uses predetermined directions and doses across eligible images and summarizes the resulting paired response distribution, including its mean and quantiles. Its purpose is model evaluation rather than the construction of an attribution map for an individual prediction.

Representation geometry, similarity, and distribution shift. The manifold view of representation learning motivates studying the structure of learned representations beyond task accuracy (Bengio et al., 2013). SVCCA (Raghu et al., 2017) and CKA (Kornblith et al., 2019) compare representations across layers, networks, or training conditions. Intrinsic-dimension estimators provide complementary descriptions of local representation geometry (Levina & Bickel, 2004; Facco et al., 2017; Ansuini et al., 2019). Recent work has continued to develop similarity and geometric analyses for modern architectures. For example, Jiang et al. (2025) study layer-wise similarity in Transformers and relate sample-wise cosine similarity to patterns captured by CKA. Other work uses geometric quantities to characterize feature-learning regimes and learned manifolds (Chou et al., 2025; 2026).

Input perturbations can also induce distribution shift, making it useful to examine how intervened representations depart from the clean reference distribution. In SCI, representation drift is summarized using MMD, per-image displacement, departure from a clean principal subspace, and local intrinsic dimension. An atypicality score measures departure from the clean feature distribution and is used as a diagnostic when comparing foreground and background interventions. Dose-stratified HSIC further describes the statistical dependence between representation displacement and decision change.

Occlusion robustness and controlled perturbation benchmarks. Occlusion has long been used both to evaluate visual recognition and to construct training-time augmentation schemes. Random

erasing and related methods introduce masked regions during training to improve robustness (DeVries & Taylor, 2017; Zhong et al., 2020). Controlled occlusion benchmarks evaluate how performance changes with occlusion severity. Modi et al. (2023) study controlled and realistic occlusion in video action detection, while COCO-OLAC (Wei et al., 2025) provides manually annotated occlusion levels for large-scale image understanding. Other recent work compares CNN and Transformer robustness under controlled and natural occlusion (Kassaw et al., 2026). Training approaches such as occlusion-robust Transformers seek representations that remain stable under masking (Wu et al., 2025).

This literature commonly reports aggregate performance measures such as accuracy, AP, or PQ across occlusion conditions. SCI retains the paired response for each image and treats direction and dose as design variables. This supports distributional summaries of heterogeneous responses while remaining compatible with conventional aggregate robustness measures.

Intervention, ablation, and causal analysis. Several methods use intervention-like operations to study input features or learned representations. ROAR (Hooker et al., 2019) removes pixels identified as important by an attribution method and retrains the model, evaluating whether the attribution ranking selects features that remain important after retraining. Counterfactual explanation methods instead construct plausible modifications that alter the prediction of an individual instance (Goyal et al., 2019).

Causal representation learning studies how representations may reflect causal variables or how interventions can help identify latent factors (Schölkopf et al., 2021; Li et al., 2024). Recent visual work uses causal interventions to separate semantic attributes from nuisance variation when learning representations (Huang et al., 2025).

SCI uses interventions to evaluate frozen models over a common direction–dose grid. It does not assume a structural causal model over latent visual factors. The representation–decision analysis is therefore interpreted as statistical dependence between two responses to the same intervention, rather than as identification of a causal pathway between them.

B Design and Theoretical Properties

This section collects the formal and implementation details underlying SCI. Appendix B.1 establishes the sampling properties of the paired estimators and specifies the testing procedures. Appendix B.2 defines the structured occlusion operator and its nesting properties. Appendix B.3 provides the placebo decomposition and the proofs supporting its mean-level interpretation.

B.1 Theoretical Properties of the Self-Controlled Design

This appendix develops statistical inference for the paired effects measured by SCI. The model, intervention operator, and evaluation pipeline are fixed. Each image provides both clean and intervened responses. The only uncertainty comes from sampling the target image. The results below assume independent images. Detection uses the city as the resampling unit, as described in Appendix D.3.

The QIE estimand. Let $F_{\tau(t,d)}(x) = \Pr_{X\sim P}[\tau(X;t,d) \le x]$ denote the cumulative distribution function of the unit-level effect, with generalized inverse $F^{-1}_{\tau(t,d)}(q) := \inf\{x \in \mathbb{R} : F_{\tau(t,d)}(x) \ge q\}$. The quantile intervention effect surface is $\tau_q(t,d) = F^{-1}_{\tau(t,d)}(q)$ for $q \in (0,1)$, estimated by the empirical quantile of $\{\tau_i(t,d)\}_{i=1}^n$.

Proposition B.1 (Direct observation and population identification)**.** *Under exact replication with a deterministic operator and evaluation pipeline (Section 2.1), the individual effect $\tau_i(t,d)$ in equation 1 is directly observed. For fixed f and $\mathcal{I}$, its population distribution and mean $\tau(t,d)$ are identified as functionals of the image law P.*

Proof. Both $f[\mathcal{I}(X_i,t,d)]$ and $f[\mathcal{I}(X_i,0,d)]$ are computed, so $\tau_i(t,d)$ is observed for every i. Taking expectations over P gives the estimand as a functional of P. □

Remark B.1 (Scope). *The identified quantity is the population response to the specified intervention on the fixed model. The distribution of the observed paired effects determines this quantity; interpretation of its signal and artifact components requires the additional assumptions in Section 2.3.*

Proposition B.2 (Sampling inference). *Fix (t, d) and let $X_i \overset{\text{i.i.d.}}{\sim} P$. If $\mathbb{E}\big[\tau_i(t,d)^2\big] < \infty$ and $V(t,d) = \mathrm{Var}\big(\tau_i(t,d)\big) > 0$, then we have*

$$\hat{\tau}_n(t,d) \xrightarrow{p} \tau(t,d), \qquad \text{and} \qquad \sqrt{n}\,\big(\hat{\tau}_n(t,d) - \tau(t,d)\big) \xrightarrow{d} \mathcal{N}\big(0,\, V(t,d)\big), \tag{B.1}$$

with the plug-in variance estimator $\hat{V}(t,d) = \frac{1}{n}\sum_i \big(\tau_i(t,d) - \hat{\tau}_n(t,d)\big)^2$. Writing $Y(t) = f[\mathcal{I}(X,t,d)]$,

$$V(t,d) \;=\; \mathrm{Var}\big(Y(t)\big) + \mathrm{Var}\big(Y(0)\big) - 2\,\mathrm{Cov}\big(Y(t), Y(0)\big). \tag{B.2}$$

Proof. $\tau_i(t,d) = \tau(X_i; t, d)$ are i.i.d. with finite second moment, so the weak law of large numbers and the central limit theorem apply directly to the paired sample mean. □

Remark B.2 (Sampling uncertainty). *For the fixed evaluation pipeline, repeated evaluations of $\mathcal{I}(X_i, t, d)$ return the same response. Uncertainty in equation B.1 concerns variation across image samples from P.*

Corollary B.1 (Variance comparison under pairing). *Relative to an unpaired two-sample design that estimates $\tau(t,d)$ from different images in the intervention and baseline arms, with variance*

$$n^{-1}\big[\mathrm{Var}(Y(t)) + \mathrm{Var}(Y(0))\big],$$

the self-controlled paired estimator has variance

$$n^{-1}\mathrm{Var}\big(Y(t) - Y(0)\big).$$

The difference between the unpaired and paired variances is

$$2n^{-1}\mathrm{Cov}\big(Y(t), Y(0)\big).$$

Consequently, pairing is more efficient when this covariance is positive, equally efficient when it is zero, and less efficient when it is negative. The sign of the covariance is not implied by the fact that both responses are evaluated on the same image. Pairing therefore affects the precision of the mean-effect estimator through the covariance term but does not alter the underlying unit-level effects. Observing both responses for each image additionally makes τ_i and its heterogeneity $\mathrm{Var}(\tau_i)$ directly estimable.

Proof. Immediate from equation B.2. □

Distributional note. Because paired differences are observed directly, the empirical distribution $\hat{F}_{t,d}$ of $\{\tau_i(t,d)\}_{i=1}^n$ directly estimates $F_{\tau(t,d)}$ in the target image population. For classification, under independent image sampling, the empirical distribution is uniformly consistent by the Glivenko–Cantelli theorem and converges to a Gaussian process at the $\sqrt{n}$ rate by Donsker's theorem. Detection inference instead treats cities as the independent sampling units and does not invoke an image-level i.i.d. empirical-process argument.

Asymptotically Gaussian inference for a targeted quantile additionally requires $F_{\tau(t,d)}$ to be continuous with strictly positive density in a neighborhood of that quantile, ensuring Hadamard differentiability of the quantile map. We invoke this regularity condition for the classification logit quantiles. Detection responses can contain point masses, particularly at zero. When a targeted detection quantile is affected by such a point mass, the smooth quantile-process approximation does not apply; the corresponding generalized-inverse empirical quantile and city-cluster bootstrap interval are reported as descriptive distributional summaries rather than as asymptotically Gaussian confidence intervals.

Unless otherwise stated, the reported bands provide pointwise (q,t) coverage. Two analyses use simultaneous bands. For the headline decision-level fan (ResNet-50, Center→Outward), we construct a simultaneous sup-t band over the dose–quantile grid. For the placebo-contrast fans in Figure C.2, we construct simultaneous sup-t bands over the dose grid. In both cases, the critical value is estimated as the 95th percentile of the bootstrap studentized maximum statistic using $B = 2{,}000$ image-level resamples. Extreme quantiles beyond q_{05} and q_{95} are not reported, as tail bootstrap intervals become unstable below roughly 140 observations per tail.

Testing protocol. Directional comparisons use a Friedman omnibus test followed by two-sided Wilcoxon signed-rank pairwise comparisons, with Benjamini–Hochberg adjustment within each model–dose family. For classification, images are the paired sampling units.

For detection, let

$$\bar{\tau}_g(t,d) = \frac{1}{n_g} \sum_{i:\, c(i)=g} \tau_i(t,d)$$

denote the mean intervention effect in city g, where $g = 1, \ldots, 21$. At each model–dose combination, the Friedman statistic Q is computed from the 21×5 matrix $\{\bar{\tau}_g(t,d)\}_{g,d}$, treating cities as paired blocks and directions as treatments.

Significance is calibrated by independently permuting the five direction labels within each city. With $B_{\text{perm}} = 200{,}000$ permutations, the inclusive Monte Carlo value is

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^{B_{\text{perm}}} \mathbf{1}\{Q^{*(b)} \geq Q_{\text{obs}}\}}{B_{\text{perm}} + 1}.$$

For all ten detector–dose omnibus tests, no permuted statistic reached the observed value, giving $p_{\text{perm}} = 1/200001$.

For each pair of directions (d, d'), the two-sided Wilcoxon signed-rank test is applied to the 21 city-level paired contrasts

$$\{\bar{\tau}_g(t,d) - \bar{\tau}_g(t,d')\}_{g=1}^{21}.$$

The ten resulting pairwise values are adjusted by Benjamini–Hochberg within each model–dose family. Where a distributional statement is needed, we report empirical quantiles with the corresponding city-cluster bootstrap intervals.

Permutation schemes. The reported MMD procedure independently swaps each image's clean and intervened feature vectors with probability $1/2$ (1,000 permutations). Its reference assumption is exchangeability of the two arms within an image, which is stronger than equality of their marginal distributions. For randomization calibration, the test bandwidth is selected within each stratum from a pair-subsampled pooled union containing both arms of every selected image pair. This pooled set is unchanged by any allowed within-pair swap. The clean-reference bandwidth used for the descriptive MMD trajectories is estimated separately once per model and held fixed across the intervention grid. The HSIC procedure permutes image indices within each classification dose–direction stratum (1,000 permutations), breaking the pairing between representation changes and decision effects. This reference distribution requires exchangeable sampling units under independence. For detection, where images are clustered within 21 cities, we instead use the cross-city statistic in Appendix D.3 with a city-level Rademacher multiplier bootstrap (20,000 draws). This cluster-aware calibration avoids treating images from the same city as independent units (Liu et al., 2021a; Chwialkowski et al., 2014).

Table B.1: **Estimand map.** Estimands, estimators, and uncertainty quantification for the three levels of Section 2.2 and the placebo decomposition of Section 2.3. Intervals use an image-level bootstrap for classification and a city-cluster bootstrap for detection. Unless otherwise stated, intervals are pointwise. The headline QIE fan (Figure 2b) uses a simultaneous sup-t band over the dose–quantile grid, while the placebo-contrast fans (Figure C.2) use simultaneous sup-t bands over the dose grid.

Estimand	Level	Estimator	Uncertainty
ATE $\tau(t,d)$	decision	paired sample mean	cluster bootstrap
QIE $\tau_q(t,d)$	decision	empirical quantile of $\{\tau_i\}$	cluster bootstrap
Signal ATE $\tau_{\text{sig}}(t,d)$	decision	placebo DiD of paired means	cluster bootstrap; reference assumptions
Contrast quantiles	decision	empirical quantile of the per-image contrast	cluster bootstrap
MMD^2 drift	representation	paired U-statistic estimator	cluster bootstrap; swap-perm. test
Displacement δ_i, angle c_i	representation	paired feature changes	cluster bootstrap
PCA residual r_i	representation	distance to baseline top-k subspace	cluster bootstrap
Two-NN intrinsic dimension	representation	point estimate	bootstrap CI
HSIC / nHSIC; $T_\times$ in detection	coupling	dose-stratified, shared PC-256 basis	image permutation in classification; city-multiplier bootstrap in detection; BH within model
Distance correlation	coupling	robustness check of nHSIC	permutation test

B.2 Detailed Specification of the Structured Occlusion Operator

This appendix specifies the classification occlusion operator. Let $X \in \mathbb{R}^{H\times W\times C}$ denote the pixel array of an annotated image and let $M \in \{0,1\}^{H\times W}$ be its foreground mask. The operator first constructs a fixed foreground support, orders its pixels according to the chosen direction, and then replaces a dose-dependent prefix of that ordering.

Definition B.1 (Structured occlusion operator)**.** *The operator $\mathcal{I}(X,t,d)$ is constructed via the following three core definitions.*

Foreground support. *The binary foreground mask is dilated with a circular structuring element of radius three pixels:*

$$G = \mathrm{supp}(M \oplus \mathcal{K}_3), \tag{B.3}$$

where $\oplus$ denotes morphological dilation and $\mathcal{K}_3$ is the 7×7 circular kernel used in the implementation. The same support G is used at every dose and in every direction.

Directional ordering. *For a pixel $p = (y(p), x(p)) \in G$, the horizontal coordinate x increases from left to right and the vertical coordinate y increases from top to bottom. Define the distance transform*

$$D_G(p) = \min_{q\notin G} \|p - q\|_2. \tag{B.4}$$

Each direction assigns a score to every pixel:

$$s_d(p) = \begin{cases} D_G(p), & d = \text{Center} \to \text{Outward},\\ -x(p), & d = \text{Left} \to \text{Right},\\ y(p), & d = \text{Bottom} \to \text{Top},\\ -y(p), & d = \text{Top} \to \text{Bottom}. \end{cases} \tag{B.5}$$

When G covers the entire grid, D_G has no minimizing q. The distance transform addresses this by treating the array border as background, keeping the ordering well-defined. No evaluated mask reaches this case. Let $\pi_{G,d}$ be the ordering of the pixels in G obtained by sorting $s_d(p)$ from largest to smallest. Ties are resolved by the fixed deterministic ordering returned by the implementation, and the same ordering is used at every dose.

Let $N_G = |G|$. The number of replaced pixels is

$$k(t,G) = \begin{cases} N_G, & t \geq 0.99,\\ \lfloor tN_G \rfloor, & t < 0.99. \end{cases} \tag{B.6}$$

The implementation uses the first case to represent the full-occlusion endpoint. No evaluated dose lies in $[0.99, 1)$, so this convention affects only $t = 1$ in the experiments. The set of replaced pixels is

$$\mathcal{O}_G(t,d) = \{\pi_{G,d}(1), \ldots, \pi_{G,d}(k(t,G))\}. \tag{B.7}$$

Replacement color. *Let*

$$B_G = (G \oplus \mathcal{K}_1) \setminus (G \ominus \mathcal{K}_1) \tag{B.8}$$

be the boundary band obtained with a square 3×3 all-ones structuring element $\mathcal{K}_1$, where $\ominus$ denotes morphological erosion. Let

$$\mu_G = \frac{1}{|G|}\sum_{p\in G} p \tag{B.9}$$

be the centroid of the dilated support. For $q \in B_G$, define

$$w(q) = \exp\left(-\frac{\|q - \mu_G\|_2^2}{2\sigma^2}\right), \qquad \sigma = 5. \tag{B.10}$$

The replacement color is the weighted average

$$c(X,G) = \frac{\sum_{q\in B_G} w(q)X(q)}{\sum_{q\in B_G} w(q)}. \tag{B.11}$$

If the boundary band is empty, the implementation uses the mean color within G. With $\sigma = 5$ pixels the weights decay quickly for large supports, so the fill is dominated by the boundary pixels nearest the centroid and in the limit reduces to a nearest-boundary color sample. Distances are computed in double precision, and no evaluated configuration produced a zero weight sum. The directional rankings are unchanged under four alternative fill rules, including fills without centroid weighting (Appendix D.4). A weighted average may produce an RGB value that is not equal to any individual boundary pixel; its role is to derive the fill from the image rather than impose a fixed color such as black or gray.

The intervened image is

$$\mathcal{I}(X, t, d)(p) = \begin{cases} c(X, G), & p \in \mathcal{O}_G(t, d), \\ X(p), & p \notin \mathcal{O}_G(t, d). \end{cases} \tag{B.12}$$

If the foreground mask is empty, the operator returns the original image.

Proposition B.3 (Dose control and nesting)**.** *For a fixed image and direction, we have*

$$|\mathcal{O}_G(t, d)| = k(t, G). \tag{B.13}$$

If $t_1 < t_2$, then

$$\mathcal{O}_G(t_1, d) \subseteq \mathcal{O}_G(t_2, d). \tag{B.14}$$

The inclusion is strict exactly when $k(t_1, G) < k(t_2, G)$.

Proof. The operator selects the first $k(t, G)$ elements of the fixed ordering $\pi_{G,d}$. Since $k(t, G)$ is nondecreasing in t, every pixel selected at dose t_1 remains selected at dose $t_2 > t_1$. Two doses can produce the same set when the floor operation gives the same value of $k(t, G)$. □

Proposition B.4 (Determinism and fixed coordinates)**.** *For fixed X, M, t, and d, the operator produces one intervened image. It does not move, rotate, or rescale any image region, and every pixel outside $\mathcal{O}_G(t, d)$ retains its original value. The support G and replacement color $c(X, G)$ do not vary with dose or direction.*

Proof. The dilated support, directional scores, pixel ordering, and replacement color are deterministic functions of the image and its foreground mask. The replacement rule in equation B.12 changes pixel values only on $\mathcal{O}_G(t, d)$ and does not alter pixel coordinates. □

Remark B.3 (Interpretation)**.** *The operator performs structured replacement of pixels within a foreground support. It does not remove the object from the image, and it does not produce a pure shape stimulus. At full dose, the support is filled with a color computed from its original boundary, while the surrounding image remains unchanged.*

Direction specifies which pixels are replaced first. The three side-to-side directions are defined by image coordinates, so their contrasts also reflect the location within an object at which replacement begins. Center→Outward instead follows distance to the boundary of the foreground support.

Increasing the dose removes more of the original foreground pixels, but it can also move the image farther from the model's training distribution. The total response to $\mathcal{I}$ is directly measurable, whereas an interpretation in terms of information loss requires the placebo reference assumptions described in Section 2.3. Replacing $\mathcal{I}$ with another deterministic operator would leave the statistical procedure applicable, but it would generally change the numerical estimands.

B.3 The Placebo Decomposition of Signal and Artifact in Detail

This appendix proves the mean-level statements in Section 2.3 and describes the reference assumptions behind the classification placebo.

Proof of Proposition 1. SCI determines the paired total responses and their population mean $\tau(t, d)$. Without restrictions on the artifact reference, any c gives a decomposition $\tau_{\text{sig}}(t, d) = c$ and $\tau_{\text{ood}}(t, d) = \tau(t, d) - c$ with the same observed total response. Thus the two components are not separately identified. □

Proof of Proposition 2. Assumption 1 gives $\tau^0(t,d) = \tau^0_{\text{ood}}(t,d)$, and Assumption 2 gives $\tau^0_{\text{ood}}(t,d) = \tau_{\text{ood}}(t,d)$. Substitution in equation 4 yields $\tau_{\text{sig}}(t,d) = \tau(t,d) - \tau^0(t,d)$. The shared baseline cancels in each image-level placebo contrast, so its sample mean estimates this difference. The empirical quantiles estimate the distribution of $\tilde{\tau}_i(t,d)$, whose unit-level signal and artifact components remain unspecified by the two mean-level assumptions. □

The reference design. The placebo replaces background pixels using the same directional ordering, Gaussian boundary fill, and support-relative dose rule as the foreground intervention.

Background operator. Let $\mathcal{D}(X)$ be the dilated foreground support of Appendix B.2 and Ω the image grid. The background support is the complement $B(X) = \Omega \setminus \mathcal{D}(X)$. Pixels of $B(X)$ are ordered by the same score s_d as the foreground arm (Appendix B.2). For Center→Outward, the score is the Euclidean distance transform to the foreground boundary $\text{dist}(p, \partial\mathcal{D})$ in ascending order, replacing background pixels nearest the object first. Pixel ordering operates independently of a centroid. Under the support-relative dose rule, the placebo replaces $\lfloor t\,|B(X)| \rfloor$ pixels, using the same Gaussian boundary fill band as the foreground arm. Background context can contribute to a prediction, so zero mean signal loss is a substantive reference assumption. High placebo doses, especially $t \geq 0.9$, remove more context and serve as sensitivity references. Moreover, equal support-relative doses need not replace equal numbers of pixels in the two arms (Appendix C.1).

Sensitivity of the mean contrast. Writing $\tau^0 = \tau^0_{\text{sig}} + \tau^0_{\text{ood}}$ and subtracting this identity from $\tau = \tau_{\text{sig}} + \tau_{\text{ood}}$ yields equation 7. When mean artifact responses agree and $\tau^0_{\text{sig}} \leq 0$, the observed contrast satisfies $\tilde{\tau} \geq \tau_{\text{sig}}$. This is an upper bound on the signed signal effect, or a lower bound on the signed loss $-\tau_{\text{sig}}$ after reversing signs. A general artifact mismatch has no fixed bias direction.

Within-dose dependence. Representation drift and decision effects can both vary with dose. Their dependence is therefore assessed within fixed dose–direction strata, following Section 2.2.

C Placebo Arms and Controls

This section reports the control experiments used in the classification analysis. Appendix C.1 gives the complete same-dose foreground–background contrasts. Appendix C.2 reports the model-randomization sanity check. Appendix C.3 presents the atypicality and reference-distribution diagnostics.

C.1 The Same-Dose Difference-in-Differences

Pipeline. The background operator $\mathcal{I}_0$ uses the same directional ordering and Gaussian boundary fill as the foreground operator. We evaluate the support-relative doses $t \in \{0.02, 0.05, 0.10, 0.15, 0.2, 0.5, 0.8, 0.9, 1.0\}$. Equal nominal doses specify equal fractions of the respective supports, allowing the absolute numbers of replaced pixels to vary. To evaluate potential bias in the placebo contrast, we assess mean atypicality gaps using the diagnostic in Appendix C.3. Inference uses an image-level cluster bootstrap ($B = 2{,}000$) with deterministic per-cell seeds, retaining the paired responses of each image.

Same-dose contrasts. Table C.1 reports the mean paired placebo contrast $\tilde{\tau}(t,d)$ across the shared three-dose grid. The contrast is significantly negative in 47 of 48 cells. The sole exception occurs for ConvNeXt-Base under Center→Outward at $t = 0.2$, where the contrast is significantly positive. This is because background occlusion reduces the response more than foreground occlusion. Correspondingly, the ratio ATE^0/ATE can exceed one. At $t = 0.8$, this ratio varies substantially from 0.11 to 0.48 across models and directions, showing that the relative magnitude of the placebo response depends on both architecture and intervention direction. Under Assumptions 1–2, the ratio can be interpreted as the signed artifact-to-total response ratio, with the numerator representing the mean placebo artifact response and the denominator representing the mean total intervention response.

Table C.1: **Same-dose placebo contrasts.** Mean contrast $\tilde{\tau}(t, d)$ and placebo-to-intervention ATE ratio $\mathrm{ATE}^0/\mathrm{ATE}$ (contrast / ratio), on the shared dose grid for four architectures.

Model	Direction	$t = 0.2$	$t = 0.5$	$t = 0.8$
ConvNeXt-Base	Bottom→Top	−0.08 / 0.61	−0.14 / 0.65	−0.43 / 0.48
ConvNeXt-Base	Center→Out	0.23 / 1.93	−0.23 / 0.59	−1.27 / 0.21
ConvNeXt-Base	Left→Right	−0.16 / 0.34	−0.28 / 0.48	−0.76 / 0.34
ConvNeXt-Base	Top→Bottom	−0.16 / 0.41	−0.35 / 0.39	−0.88 / 0.28
MobileNet-V3-L	Bottom→Top	−0.13 / 0.19	−0.40 / 0.39	−1.22 / 0.32
MobileNet-V3-L	Center→Out	−0.19 / 0.68	−1.15 / 0.28	−2.47 / 0.21
MobileNet-V3-L	Left→Right	−0.14 / 0.06	−0.55 / 0.18	−1.52 / 0.28
MobileNet-V3-L	Top→Bottom	−0.26 / 0.11	−0.87 / 0.16	−2.00 / 0.19
ResNet-50	Bottom→Top	−0.03 / 0.32	−0.24 / 0.25	−0.87 / 0.22
ResNet-50	Center→Out	−0.15 / 0.57	−0.95 / 0.26	−2.08 / 0.16
ResNet-50	Left→Right	−0.08 / 0.13	−0.38 / 0.16	−1.29 / 0.20
ResNet-50	Top→Bottom	−0.12 / 0.15	−0.49 / 0.18	−1.50 / 0.15
ViT-B/16	Bottom→Top	−0.09 / 0.34	−0.34 / 0.29	−1.23 / 0.19
ViT-B/16	Center→Out	−0.10 / 0.69	−0.92 / 0.24	−2.60 / 0.13
ViT-B/16	Left→Right	−0.09 / 0.28	−0.36 / 0.26	−1.52 / 0.18
ViT-B/16	Top→Bottom	−0.20 / 0.16	−0.74 / 0.08	−1.94 / 0.11

Contrast quantiles. Table C.2 reports empirical quantiles of the paired placebo contrast under Center→Outward. Its lower decile remains negative in every architecture. At $t = 0.8$, its magnitude is $2.5\times$ to $4.3\times$ that of the placebo effect's lower decile, while the upper decile remains near or above zero through moderate doses. These distributions show substantial heterogeneity in relative responses to foreground and background occlusion. The per-model fans (Figure C.2, with mean-score diagnostics in Appendix C.3) display the full contrast trajectories.

Table C.2: **Quantiles of the paired placebo contrast under Center→Outward.** Quantiles of $\tilde{\tau}_i$ at q_{10} [95% CI] and q_{90}, four architectures.

Model	$t = 0.2$	$t = 0.5$	$t = 0.8$	$t = 1.0$
ResNet-50	−1.46 [-1.65,-1.37] / +0.98	−3.59 [-3.74,-3.41] / +1.02	−5.21 [-5.32,-5.08] / +0.69	−5.82 [-5.90,-5.72] / +0.19
ViT-B/16	−0.99 [-1.08,-0.91] / +0.70	−4.06 [-4.24,-3.82] / +0.80	−6.99 [-7.22,-6.82] / +0.76	−7.85 [-7.98,-7.71] / +0.14
MobileNet-V3-L	−1.91 [-2.01,-1.77] / +1.32	−3.85 [-4.00,-3.71] / +1.26	−6.03 [-6.18,-5.93] / +0.79	−7.20 [-7.36,-7.06] / +0.42
ConvNeXt-Base	−0.39 [-0.45,-0.36] / +1.01	−1.76 [-1.95,-1.57] / +0.97	−4.50 [-4.63,-4.32] / +0.95	−5.68 [-5.83,-5.57] / +0.76

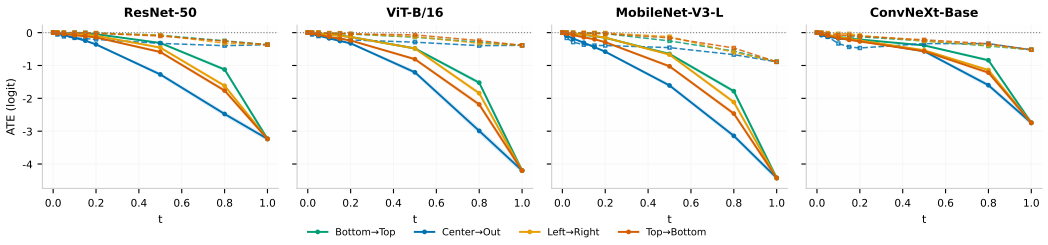
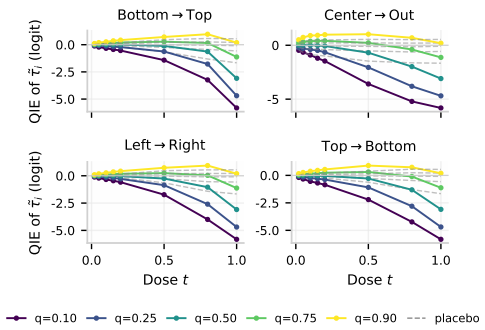


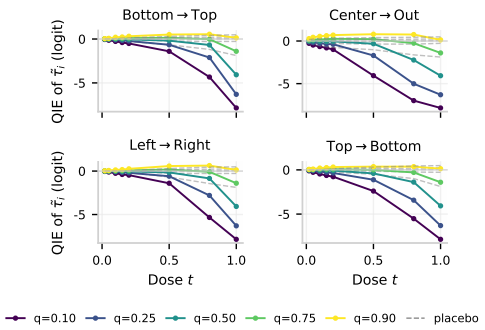
Figure C.1: **Same-dose foreground and background responses.** Directional ATE trajectories for the intervention arm (solid) and background placebo (dashed), with 95% cluster-bootstrap intervals on the intervention arm. Their difference is the mean paired placebo contrast. The accompanying atypicality diagnostic is reported in Appendix C.3.

C.2 Model-Randomization Sanity Check

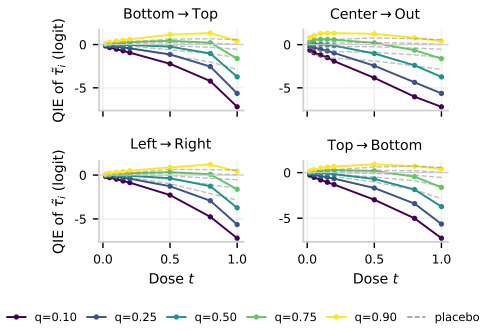
Following the C1 control of Section 2.3 (Adebayo et al., 2018), we re-run the directional SCI pipeline on a randomly-initialized backbone (three seeds) and on a randomized classification head, for ResNet-50 and ViT-B/16. Three questions settle whether the findings are properties of learned representations.



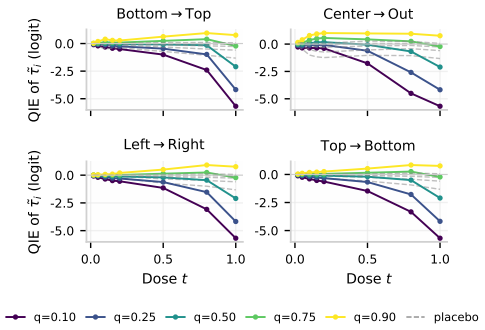
(a) ResNet-50



(b) ViT-B/16



(c) MobileNet-V3-L



(d) ConvNeXt-Base

Figure C.2: **Quantiles of the paired placebo contrast, four architectures.** Solid curves give quantiles of $\tilde{\tau}_i$ and dashed curves give quantiles of the placebo effect. At high doses, the lower contrast decile has a larger negative magnitude than the placebo decile; the pattern is shared across architectures. Shaded bands are 95% simultaneous (sup-t) intervals over the dose grid, computed as in Appendix B.1.

Q1: Does baseline performance collapse to chance? Yes. Pretrained baselines are 94.7% (ResNet-50) and 96.8% (ViT-B/16) top-1 on the curated ten-class subset. Randomly initialized backbones score 0.0%, and the randomized head scores $\leq 0.6\%$.

Q2: Does the effect collapse? Yes. The Center→Outward ATE at $t = 1.0$ collapses from -3.24 (ResNet-50) and -4.20 (ViT-B/16) to small values with reversed sign under random initialization ($+0.24$ to $+0.59$ for ResNet-50 across seeds, exactly 0.000 for ViT-B/16) and to ≤ 0.12 under the randomized head.

Q3: Does the direction hierarchy vanish? Yes. The spread of directional ATEs at $t = 0.8$ collapses from 1.36 (ResNet-50) and 1.47 (ViT-B/16) to 0.07–0.08 under random initialization and 0.03–0.04 under the randomized head.

Table C.3: **C1 model-randomization sanity check.** Baseline top-1, Center→Outward ATE at $t = 1.0$, and directional ATE spread at $t = 0.8$ under weight randomization. The effect requires end-to-end trained mappings.

Condition	Model	Baseline top-1	ATE (center, t=1)	Spread (t=0.8)
pretrained	ResNet-50	94.7%	-3.24	1.36
pretrained	ViT-B/16	96.8%	-4.20	1.47
random init (s0)	ResNet-50	0.0%	$+0.31$	0.07
random init (s1)	ResNet-50	0.0%	$+0.24$	0.08
random init (s2)	ResNet-50	0.0%	$+0.59$	0.07
random init (s0)	ViT-B/16	0.0%	0.000	0.00
random head (s0)	ResNet-50	0.6%	$+0.01$	0.03
random head (s0)	ViT-B/16	0.0%	$+0.12$	0.04

Mechanism note. For randomly initialized ViT-B/16, exact zeros stem from constant-output collapse. The untrained attention layers map all inputs to a single representation, leaving target logits invariant across doses with zero logit dispersion at $t = 0$. The randomized-head control exonerates the pipeline itself, since it keeps the trained backbone and shows that the effect requires the trained end-to-end mapping rather than the architecture or the operator alone. The residual positive monotone trend of randomly initialized ResNet-50 is the reversed-sign footprint of an untrained convolutional stack, roughly an order of magnitude below the trained effect (4–13$\times$ smaller across seeds) and with no stable direction hierarchy.

C.3 Reference-Distribution Diagnostics

Atypicality score. For each (t, d), the image-level score $\rho_i(t, d)$ is based on the mean distance from the intervened representation to its $k = 10$ nearest neighbors in the clean feature set. Distances are computed in the leading-256-PC basis fitted to clean features and mapped through the empirical CDF of clean leave-one-out distances. Thus $\rho \in [0, 1]$, with larger values indicating greater departure from the clean reference set and 0.5 corresponding to its median distance. The same scoring rule is used for the placebo arm. Scores are comparable across intervention settings within a model. Raw distances are reported alongside them.

Score trajectories. Figure C.3 shows that foreground mean scores separate by direction, with center occlusion producing the largest departure. The tested background placebos attain lower mean scores than foreground occlusion at high doses, even when the whole background support is replaced. Some placebo trajectories turn downward at the highest doses, and foreground scores lie slightly below the clean median at the smallest doses. These are properties of the chosen feature-distance diagnostic.

Association across the intervention grid. Regressing cell-mean responses $\hat{\tau}_n(t, d)$ on cell-mean atypicality $\bar{\rho}(t, d)$ summarizes their association across intervention settings (Table C.4). The score alone accounts for $R^2 = 0.76$–$R^2 = 0.87$, and direction indicators add about 0.01. Both variables change with dose, so the pooled association also reflects this shared trend. This regression describes

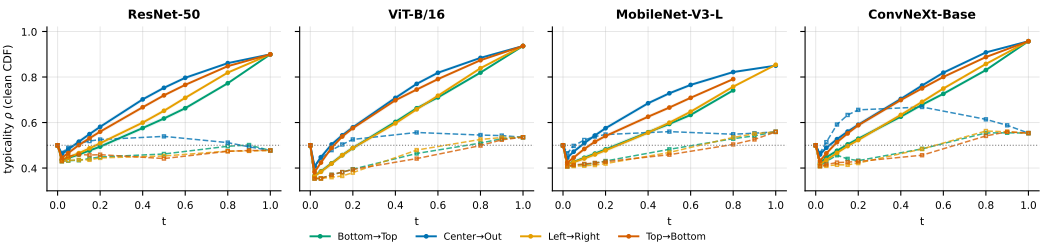


Figure C.3: **Mean atypicality under foreground and background occlusion.** Mean scores $\bar{\rho}(t, d)$ and $\bar{\rho}_0(t, d)$, based on clean-reference percentiles of feature-space nearest-neighbor distances. Foreground curves (solid) reach 0.85–0.96, while background curves (dashed) plateau at 0.48–0.56, including at $t = 1.0$. The mean-score gap is the diagnostic used in Section 2.3.

the observed response surface; the mean signal/artifact interpretation uses the separate assumptions in Section 2.3.

Table C.4: **Association of intervention responses with atypicality.** R^2 of cell-mean ATE on cell-mean atypicality (generic) and the increment from direction dummies (direction-specific), with 95% image-level bootstrap CIs.

Model	Generic R^2	Direction-specific ΔR^2
MobileNet-V3-L	0.866 [0.857, 0.874]	0.012 [0.009, 0.015]
ResNet-50	0.848 [0.840, 0.856]	0.014 [0.011, 0.018]
ConvNeXt-Base	0.758 [0.745, 0.771]	0.012 [0.009, 0.014]
ViT-B/16	0.801 [0.792, 0.810]	0.010 [0.008, 0.012]

Same-dose contrasts and the mean-score diagnostic. We summarize $\tilde{\tau}_i(t, d)$ by its mean and quantiles (q_{05} to q_{95}), with image-level cluster-bootstrap intervals. The diagnostic compares mean scores through $|\bar{\rho}_0(t, d) - \bar{\rho}(t, d)|$ and flags cells with a gap above 0.08 for sensitivity interpretation. Every model–direction cell has a gap at most 0.07 through $t = 0.05$. The fraction passing the 0.08 threshold is 88% at $t = 0.10$, 69% at $t = 0.15$, and 50% at $t = 0.2$; no cell passes at the shared doses from $t = 0.5$ onward. Passing this mean-score screen does not by itself establish matching of the score distributions or a common artifact response.

Figure C.5 shows the contrast quantiles for each architecture and direction, with line styles indicating the mean-score diagnostic. Figure C.4 shows the low-dose detail and its uncertainty intervals.

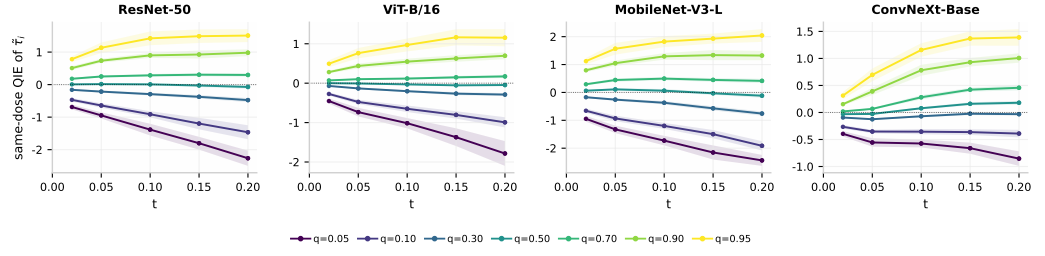
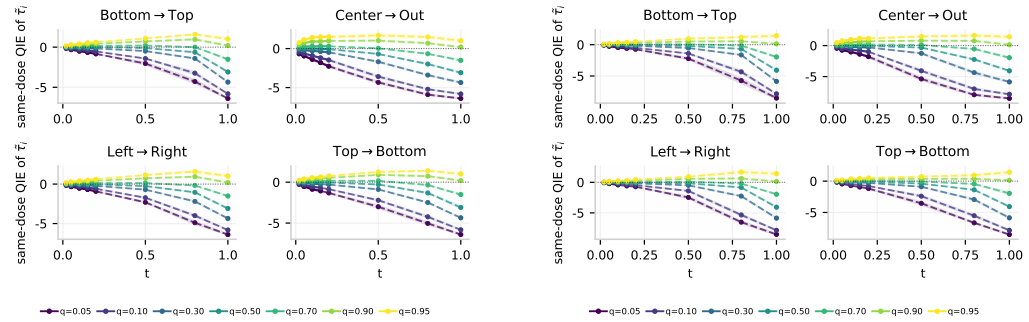
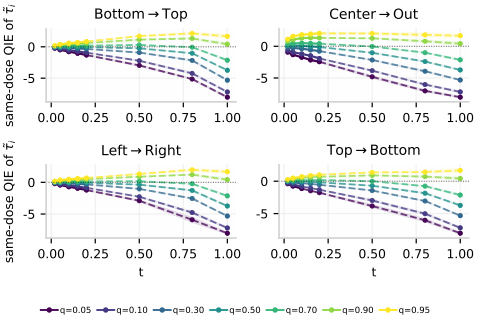


Figure C.4: **Low-dose paired placebo contrasts.** Quantiles of $\tilde{\tau}_i(t, d)$ under Center→Outward, with 95% cluster-bootstrap intervals. The mean and median are small relative to the tails and vary in sign across cells. Negative lower deciles and positive upper deciles coexist throughout the plotted grid, showing heterogeneous relative responses to the two operators.

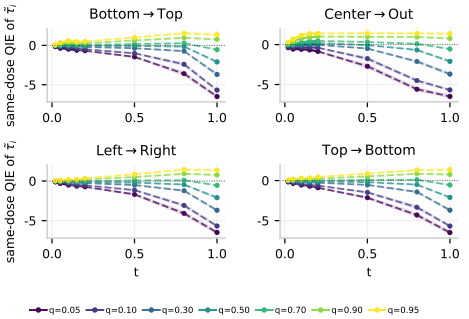


(a) ResNet-50

(b) ViT-B/16



(c) MobileNet-V3-L



(d) ConvNeXt-Base

Figure C.5: **Paired placebo contrast quantiles, four architectures.** Solid segments and points mark cells with mean atypicality gap at most 0.08; dashed segments mark larger gaps. Quantile levels range from q_{05} to q_{95}, with 95% cluster-bootstrap intervals.

D EXPERIMENTAL DETAILS

This section provides the complete experimental results and audits. Appendix D.1 reports the classification sample, category-level analyses, full architecture results, and area-matched random controls. Appendix D.2 presents the representation-level estimators and geometric analyses. Appendix D.3 reports the detection setup, statistical audits, and complete result tables. Appendix D.4 collects the additional robustness checks.

D.1 CLASSIFICATION DETAILS

Category selection. The ten mini-ImageNet categories are goose, meerkat, cannon, Ibizan hound, poncho, rhinoceros beetle, white wolf, catamaran, garbage truck, and iPod, selected to be mutually distant in the cosine-distance matrix of ImageNet-pretrained ResNet-50 class prototypes and to span distinct part-anchor topologies. After removing images without a successful SAM-3 detection, $n = 2{,}858$ images remain (95.3% of the curated pool, per-class n from 258 to 299). Analyses are conditional on successful masking.

Selection on masking. The 142 images without a successful SAM-3 detection are not a random subset. Their baseline top-1 accuracy is 64.8% (ResNet-50) and 71.8% (ViT-B/16), against 94.7% and 96.8% on the kept set, with target-logit means 1.0–1.1 lower (Mann–Whitney $p < 10^{-14}$ for both metrics and both models; bootstrap CIs exclude zero). The excluded set concentrates in fine-grained or weak-contour categories (Ibizan hound 42, iPod 28, rhinoceros beetle 19 of 142). All classification estimands are therefore conditional on the maskable subpopulation, which is easier than the raw pool, and we do not extrapolate them beyond it.

Category-level response and part-anchor geometry. Disaggregating by category, the center-concentration ratio C (center over mean peripheral damage at $t = 0.8$, averaged over architectures) tracks the part-anchor topology, highest for single central anchors and lowest for distributed-feature categories (Table D.1).

Table D.1: **Center-concentration ratio C by category.** Center over mean peripheral ATE at $t = 0.8$, averaged over the ten architectures. Categories with a single central anchor top the list.

Category	C
goose	2.16
meerkat	1.96
cannon	1.61
Ibizan hound	1.54
poncho	1.53
rhinoceros beetle	1.46
white wolf	1.43
catamaran	1.33
garbage truck	1.32
iPod	1.16

Full ten-architecture decision table. The main text reports ResNet-50 as the representative decision-level dose–response. Table D.2 presents all ten architectures at $t = 0.2$ and $t = 0.8$. The ordering Center→Outward > Top→Bottom > Left→Right > Bottom→Top holds for every architecture individually at $t = 0.8$, with Kendall's $W = 1.00$.

Area-matched random controls. We compared the four prespecified directions with five random control configurations at $t \in \{0.2, 0.5, 0.8\}$. Each control uses the same dilated foreground support, replaces the same number of pixels, and uses the same Gaussian boundary fill as the corresponding directional intervention. The controls differ only in pixel ordering. A control either applies a linear sweep at a sampled angle or a radial ordering with a sampled center and inward or outward progression.

The random seed depends on the control index and dose, but not on image identity. The five controls are therefore treated as fixed sampled configurations at each dose rather than as independent random

Table D.2: **Decision-level ATE for all ten architectures** at $t = 0.2$ and $t = 0.8$ (target-class logit, $n = 2{,}858$). At $t = 0.8$ the ordering center $>$ top $>$ left $>$ bottom holds for every architecture individually.

	$t = 0.2$				$t = 0.8$			
Model	center	top	left	bottom	center	top	left	bottom
ConvNeXt-Base	-0.24	-0.27	-0.24	-0.21	-1.60	-1.22	-1.14	-0.84
EfficientNet-B0	-0.53	-0.38	-0.16	-0.18	-3.67	-3.11	-2.56	-2.13
GoogLeNet	-1.90	-1.17	-0.57	-0.68	-5.89	-5.37	-4.90	-4.32
MobileNet-V3-L	-0.58	-0.29	-0.15	-0.16	-3.14	-2.47	-2.12	-1.78
ResNet-101	-0.34	-0.17	-0.09	-0.07	-2.54	-1.71	-1.54	-1.30
ResNet-50	-0.36	-0.14	-0.09	-0.05	-2.48	-1.76	-1.62	-1.12
Swin-T	-0.48	-0.30	-0.22	-0.23	-2.83	-2.03	-1.70	-1.49
VGG-16-BN	-3.62	-3.67	-1.73	-1.71	-12.85	-11.91	-10.80	-9.50
ViT-B/16	-0.33	-0.24	-0.13	-0.13	-2.99	-2.19	-1.85	-1.52
ViT-L/16	-0.16	-0.19	-0.08	-0.12	-2.75	-1.81	-1.41	-1.20

Table D.3: **Decision-level trajectory shape across all ten architectures.** AUC of the dose-normalized total foreground-effect trajectory within each direction. AUC below 0.5 denotes a brittle trajectory, whereas AUC above 0.5 denotes an elastic trajectory; elastic cases are bolded.

Model	Center→Out	Top→Bottom	Left→Right	Bottom→Top
ConvNeXt-Base	0.32	0.29	0.28	0.23
EfficientNet-B0	0.37	0.32	0.26	0.24
GoogLeNet	**0.58**	0.48	0.40	0.38
MobileNet-V3-L	0.42	0.32	0.27	0.25
ResNet-101	0.41	0.28	0.25	0.23
ResNet-50	0.43	0.29	0.27	0.21
Swin-T	0.38	0.30	0.26	0.24
VGG-16-BN	**0.56**	**0.53**	0.42	0.40
ViT-B/16	0.38	0.30	0.24	0.23
ViT-L/16	0.34	0.27	0.21	0.22

draws for every image. The configuration associated with a given control index may also change across doses, so these controls are not interpreted as dose–response trajectories.

For each model and dose, we compare the standard deviation of the four directional ATEs with the standard deviation across the five control ATEs:

$$R(t) = \frac{\mathrm{sd}_d\{\widehat{\tau}(t,d)\}}{\mathrm{sd}_k\{\widehat{\tau}_k^{\mathrm{rand}}(t)\}}. \tag{D.1}$$

With only five control configurations, the random-control dispersion is a coarse estimate, so we treat this comparison descriptively rather than as a formal test. Directional dispersion exceeded random-control dispersion in 28 of the 30 model–dose cells; the two exceptions were ConvNeXt-Base at $t = 0.2$ and $t = 0.5$. The median ratio was 2.745, with a range of 0.649–6.626.

We also located each directional effect within the five same-image random-control effects. Across models, doses, and directions, these percentiles ranged from 0.391 to 0.730. Their averages were 0.62 for Center→Outward, 0.56 for Top→Bottom, 0.49 for Left→Right, and 0.47 for Bottom→Top. Thus, matching the number of replaced pixels accounts for much of the overall response level but does not reproduce the directional ordering.

The ranking of the four directions was also consistent across the ten classification models. Kendall's coefficient of concordance was 0.816, 0.928, and 1.000 at doses 0.2, 0.5, and 0.8, respectively, with $p < 2 \times 10^{-5}$ at every dose.

D.2 Representation-Level Details

Feature preprocessing and geometric summaries. For each model, PCA is fitted once to the clean representations. Let $\bar{m}(0)$ be their mean and Π_k the orthogonal projector onto the lead-

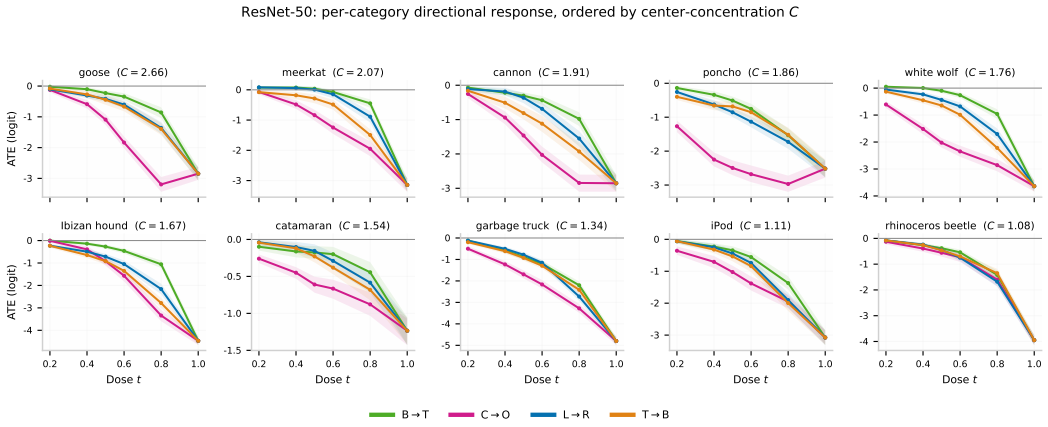


Figure D.1: **Category-disaggregated individual response.** Per-category dose–response functions for ResNet-50, stratified by part-anchor topology.

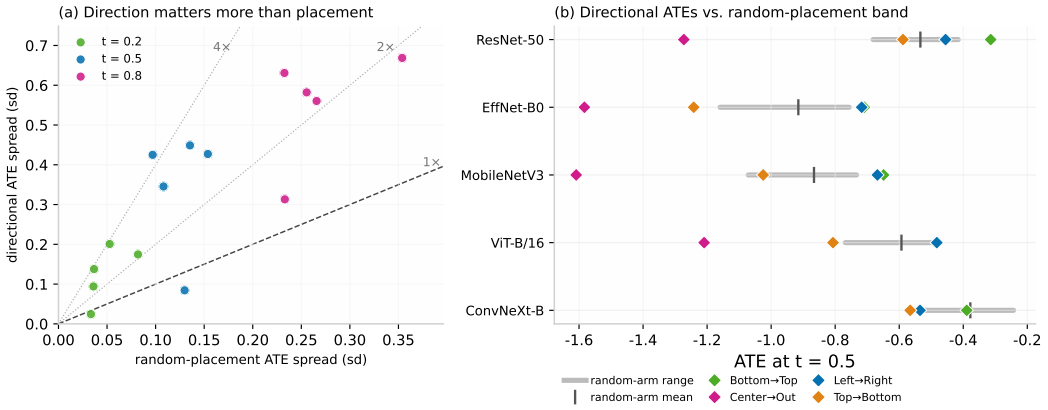


Figure D.2: **Area-matched random controls.** Directional ATE dispersion is compared with dispersion across five random control configurations at the same dose. Each control uses the same foreground support, replaced area, and fill rule as the directional intervention.

ing components explaining 90% of baseline variance. The projection residual is $r_i(t,d) = \|(I_D - \Pi_k)[m_i(t,d) - \bar{m}(0)]\|_2$. The displacement and angular summaries in Section 2.2 compare each intervened representation with its own clean counterpart. Two-NN estimates local intrinsic dimension across doses.

For HSIC, representation differences are projected onto the leading 256 baseline principal components when $D > 256$. The same basis is used at every dose and direction. Kernel bandwidths for representation changes and decision effects are estimated once per model using samples pooled across its intervention grid and are then held fixed.

MMD. Let $m_i^{(t)} = m_i(t,d)$. Drift under (t,d) is measured by the paired U-statistic estimator (Gretton et al., 2012) with Gaussian kernel $k(u,v) = \exp[-\|u-v\|^2/2\sigma^2]$:

$$\widehat{\mathrm{MMD}}_u^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k\big(m_i^{(0)}, m_j^{(0)}\big) + \frac{1}{n(n-1)} \sum_{i \neq j} k\big(m_i^{(t)}, m_j^{(t)}\big) - \frac{2}{n(n-1)} \sum_{i \neq j} k\big(m_i^{(0)}, m_j^{(t)}\big). \tag{D.2}$$

Table D.4: **Decision- vs. representation-level descriptors** under Center→Outward occlusion, with dose-stratified coupling. The elastic case is bolded. ViT-B/16 shows the dissociation: elastic in representation, brittle in decision, weakest coupling.

	Decision		Representation		
Model	AUC	t_{50}	AUC	t_{50}	nHSIC$_{t=0.8}$
MobileNet-V3-L	0.42	0.63	0.48	0.51	0.16
ResNet-50	0.43	0.59	0.43	0.58	0.15
ConvNeXt-Base	0.32	0.74	0.32	0.71	0.20
ViT-B/16	0.38	0.66	**0.53**	0.48	0.09

For the descriptive MMD trajectories, σ is the median pairwise ℓ_2 distance among baseline features, estimated once per model and held fixed across directions and doses. For the within-image swap test, a separate bandwidth is selected within each stratum from a pair-subsampled pooled union of the clean and intervened features. Because both arms of every selected pair enter this pooled union, the test bandwidth is invariant to the allowed swaps. The cross term uses different images ($i \neq j$), whose representations are independent under image-level sampling. The within-image pairs are reserved for describing individual changes.

HSIC. Let K be the Gaussian kernel matrix for the representation-change vectors, projected as described above when $D > 256$, and L the Gaussian kernel matrix for the decision effects $\{\tau_i(t,d)\}_{i=1}^n$. With centering matrix $H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, we have

$$\widehat{\mathrm{HSIC}} \;=\; \frac{1}{n^2}\,\mathrm{tr}(KHLH). \tag{D.3}$$

For comparisons of association strength, we report the normalized statistic

$$\widehat{\mathrm{nHSIC}} = \frac{\widehat{\mathrm{HSIC}}(K,L)}{\sqrt{\widehat{\mathrm{HSIC}}(K,K)\,\widehat{\mathrm{HSIC}}(L,L)}}.$$

For detection, the coupling variable is the scalar representation displacement $\delta_i(t,d) = \|\Delta m_i(t,d)\|_2$, and images are clustered within cities. Let $K_c = HKH$, $L_c = HLH$, and $A = K_c \circ L_c$. If $c(i)$ denotes the city of image i, define

$$Q_{gh} = \sum_{i:c(i)=g} \sum_{j:c(j)=h} A_{ij}.$$

The detection test uses the cross-city statistic

$$T_\times = \frac{\sum_{g\neq h} Q_{gh}}{n^2 - \sum_g n_g^2}. \tag{D.4}$$

For bootstrap draw b, one Rademacher multiplier $w_g^{(b)}$ is assigned to each city and

$$T_\times^{*(b)} = \frac{\sum_{g\neq h} w_g^{(b)} w_h^{(b)} Q_{gh}}{n^2 - \sum_g n_g^2}.$$

The inclusive Monte Carlo value is computed from 20,000 draws and adjusted by Benjamini–Hochberg across the 25 cells within each detector. This construction follows the cluster-aware motivation of Liu et al. (2021a) and the multiplier-bootstrap treatment of dependent kernel statistics in Chwialkowski et al. (2014). We retain nHSIC as a descriptive measure of coupling magnitude.

Figure D.3 compares the normalized decision- and representation-level trajectories for the four representatives over ten doses, Figure D.4 shows the corresponding dose-stratified nHSIC trajectories, and Figure D.5 reports the geometric diagnostics. The nominal grid is $4 \times 4 \times 10 = 160$ model–direction–dose display positions. Because $t = 1.0$ is the shared, direction-independent full-occlusion endpoint, the analysis contains 148 distinct intervention settings after counting this endpoint once per model. The within-image swap test rejects the arm-exchangeability null in every

setting ($B = 1{,}000$; $p = 1/1001$ throughout). At $t = 1$, normalized PCA residuals reach $2.80\times$ for ResNet-50, $2.16\times$ for ViT-B/16, $2.14\times$ for ConvNeXt-Base, and $1.67\times$ for MobileNet-V3-L. Two-NN estimates span 20.8–29.1 against clean baselines of 20.6–24.9, showing no systematic contraction in local intrinsic dimension.

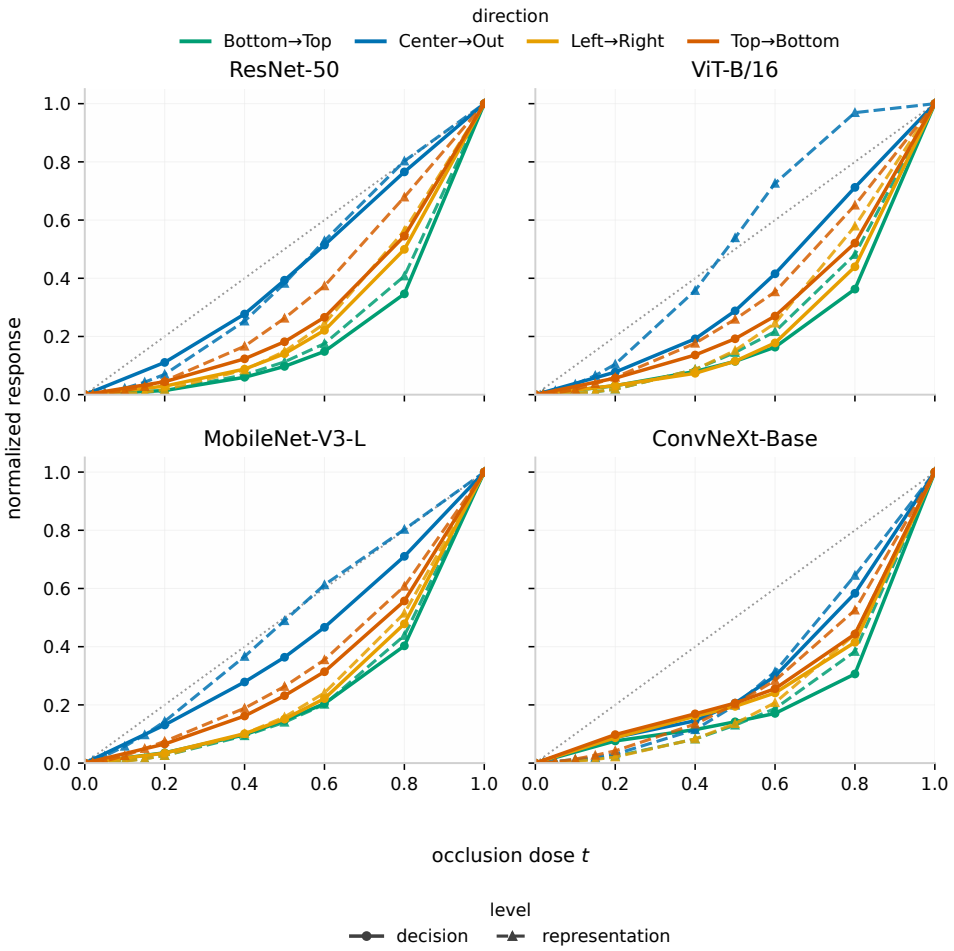


Figure D.3: **Normalized decision- and representation-level dose trajectories.** Solid lines with circles show decision-level effects, and dashed lines with triangles show representation-level MMD^2 drift. Each trajectory is normalized by its corresponding full-occlusion endpoint, so the comparison concerns trajectory shape rather than absolute magnitude. The gray diagonal denotes a linear trajectory with AUC $= 0.5$. Within each model, the direction-independent $t = 1$ endpoint is counted once and repeated across the four direction curves for display.

D.3 Detection Details and Audits

Setup. The Cityscapes training and validation splits contain 3,475 images from 21 city clusters, which we combine for the analysis. After excluding 18 images without an evaluable instance, we have $n = 3{,}457$ effectively paired samples. We consider seven COCO-compatible categories and two zero-shot detectors, Mask R-CNN R50-FPN and Mask DINO R50. For each image, we evaluate five directions at five dose levels, $t \in \{0.2, 0.4, 0.5, 0.6, 0.8\}$. The masks are defined relative to each instance and dilated deterministically according to its scale. The resulting covered fraction is close to the nominal dose, with a spread of approximately 0.08. We use soft recall equation 9 as the primary response. Uncertainty is assessed using a city-clustered bootstrap with $B = 5{,}000$ resamples, and multiple comparisons are handled using the Friedman/Wilcoxon protocol with Benjamini–Hochberg adjustment.

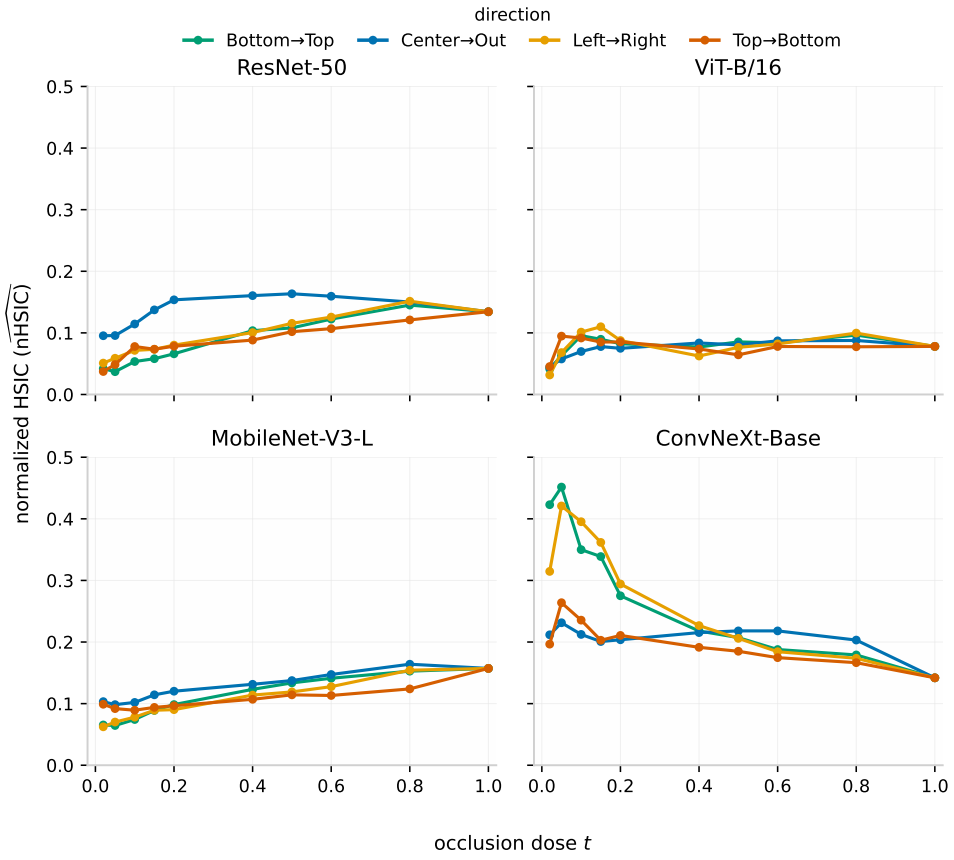


Figure D.4: **Dose-stratified coupling trajectories.** Normalized HSIC between representation changes and paired logit effects at fixed dose and direction. The permutation test rejects independence in all 148 distinct intervention settings after counting each model's direction-independent full-occlusion endpoint once ($B = 1{,}000$; $p = 1/1001$ before and after within-model BH adjustment). The shared $t = 1$ endpoint is repeated across direction curves only for display.

Audits. Three audits support the detection claims. The direction-label audit rebuilds the masks of all five directions from metadata and verifies the construction verbatim. We regress the mean drop within coverage-stratified bins on realized coverage and direction indicators. For Mask DINO, the direction indicators are jointly significant at all dose levels, with F-statistics ranging from 3.2 to 9.0 and $p \leq 0.036$. For Mask R-CNN, they are jointly significant through $t = 0.5$, with F-statistics ranging from 5.8 to 19.3 and $p \leq 0.002$, and are marginally significant at the two highest doses ($p = 0.072$ and 0.062). The estimated signs agree with the observed ordering at low and intermediate doses. The F1-bridge suite examines the same directional pattern using macro- and micro-F1 under three fixed confidence thresholds and thresholds selected on the clean baseline (Table D.5). At $t = 0.2$, Bottom→Top is the most damaging direction for both detectors across all four macro-F1 variants and under micro-F1 at threshold 0.3. The exact micro-F1 leader becomes threshold-dependent at stricter confidence cutoffs. Across the 80 detector–dose–metric comparisons, Bottom→Top produces a larger performance loss than Top→Bottom in 68 settings. Thus the broad low-dose directional pattern extends beyond the primary response, although its exact ordering depends on the aggregation rule and confidence threshold.

Decision level. Across the five directional doses $t \in \{0.2, 0.4, 0.5, 0.6, 0.8\}$, all 50 model–direction–dose ATEs are strictly negative, and the loss magnitude increases monotonically within every model–direction trajectory. Table D.7 reports the complete ATE grid. The direction-independent $t = 1$ endpoint is reported separately and is not included in this monotonicity statement. Table D.8 shows five category-level response patterns. Car exhibits near-homogeneous damage. Bi-

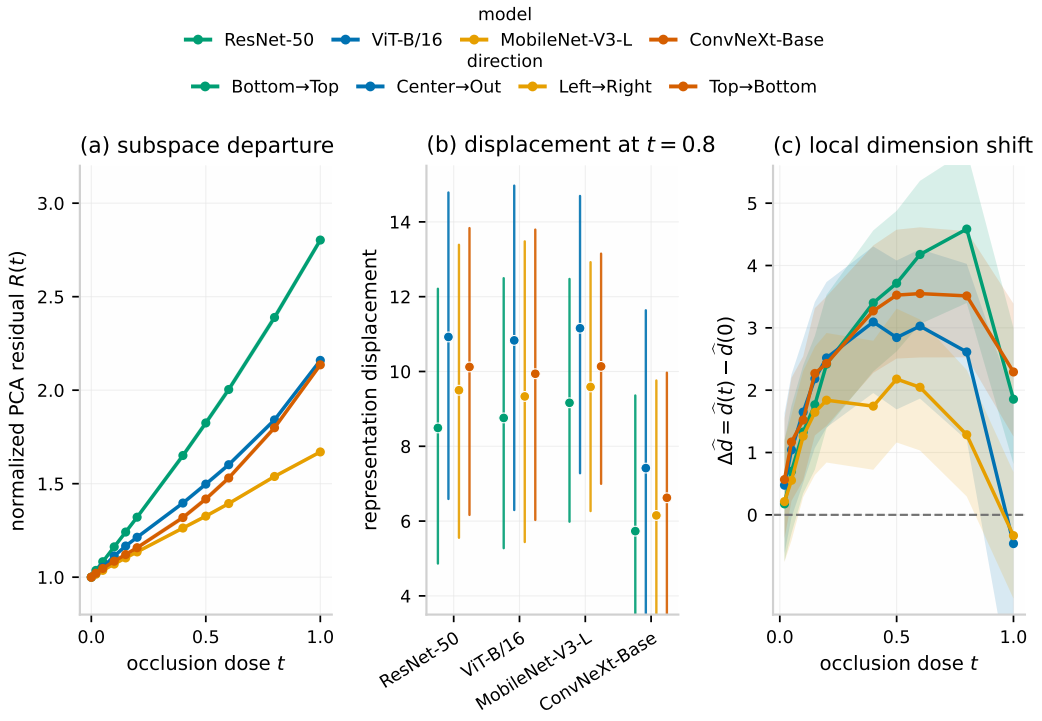


Figure D.5: **Geometric structure of representation drift.** (a) PCA projection residuals normalized by the clean reference for each model and averaged across directions. The bands are pointwise bootstrap intervals rescaled by the clean mean. (b) Mean per-image representation displacement at $t = 0.8$ with vertical segments showing the q_{10}–q_{90} range. (c) Change in Two-NN intrinsic dimension relative to the clean baseline. The trajectories show no systematic contraction, while some architectures exhibit moderate expansion. Bands in panel (c) are shown descriptively. Panels (a) and (c) use model colors, whereas panel (b) uses direction colors.

Table D.5: **F1-bridge sensitivity across detection metrics.** The two middle columns report the direction producing the largest F1 decrease at $t = 0.2$. The final column counts detector–dose settings in which Bottom→Top produces a larger decrease than Top→Bottom. Fixed values denote confidence thresholds, and t_0-opt denotes thresholds selected on the clean baseline.

Metric	Mask R-CNN	Mask DINO	B→T > T→B
Macro-F1, 0.3	B→T	B→T	10/10
Macro-F1, 0.5	B→T	B→T	9/10
Macro-F1, 0.7	B→T	B→T	6/10
Macro-F1, t_0-opt	B→T	B→T	10/10
Micro-F1, 0.3	B→T	B→T	10/10
Micro-F1, 0.5	T→B	L→R	8/10
Micro-F1, 0.7	T→B	L→R	6/10
Micro-F1, t_0-opt	T→B	B→T	9/10

cycle, Motorcycle, and Bus retain a resistant upper tail, while Person is intermediate between near-homogeneous damage and a resistant-subpopulation pattern. Train has a near-zero median ATE together with a positive upper tail, whereas Truck has zero estimated effect in every cell. These results show that category-level response distributions cannot be summarized by a single homogeneous-versus-resistant distinction. Trajectory shapes differ across detector–direction pairs. Table D.6 reports the corresponding AUC and t_{50} values. Mask R-CNN is brittle under Center→Outward and the two vertical directions but elastic under both horizontal directions, while Mask DINO is elastic under every direction. In classification, elasticity is consistently associated with the center config-

Table D.6: **Detection decision-level trajectory shape.** AUC and t_{50} of the dose-normalized soft-recall effect trajectories. AUC below 0.5 denotes a brittle trajectory, whereas AUC above 0.5 denotes an elastic trajectory.

Detector	Direction	AUC	t_{50}	Class
Mask R-CNN	Bottom→Top	0.466	0.552	brittle
	Center→Outward	0.451	0.573	brittle
	Left→Right	0.543	0.431	elastic
	Right→Left	0.530	0.442	elastic
	Top→Bottom	0.438	0.594	brittle
Mask DINO	Bottom→Top	0.587	0.363	elastic
	Center→Outward	0.580	0.410	elastic
	Left→Right	0.713	0.287	elastic
	Right→Left	0.678	0.305	elastic
	Top→Bottom	0.519	0.509	elastic

uration across architectures. In detection, it varies across model-direction pairs, with no common configuration across models.

Directional hierarchy across estimands. Under the primary estimand both detectors share a single hierarchy from $t = 0.4$ onward, Left→Right $>$ Right→Left $>$ Bottom→Top $\approx$ Center→Outward $>$ Top→Bottom. Among the 100 city-level Wilcoxon pairwise comparisons, 92 remain significant after multiplicity correction, including 44 of 50 comparisons for Mask R-CNN and 48 of 50 for Mask DINO. The eight exceptions concentrate at the lowest dose and among middle-ranking adjacent directions (Table D.9). The matched-coverage regression shows that the positional effects remain after accounting for realized coverage. The ordering is determined by the estimand rather than by image-specific characteristics.

The two horizontal directions show both an axis effect and an asymmetry between the two directions. Horizontal occlusion causes larger performance drops than vertical occlusion across all models and doses. Left-to-Right also causes slightly larger drops than Right-to-Left in every model-by-dose combination. This difference is small but consistent and is most apparent for Car and Motorcycle. The direction-label audit confirms the mask construction for all five directions. Dataset-level COCO-protocol mAP shows three patterns that are also observed with soft recall. Bottom-to-Top has the largest performance drop at the lowest dose, while the horizontal directions have the largest drops from the mid-dose levels onward. Center-to-Outward does not have the largest drop at any dose. The within-horizontal ordering is not identical across metrics. For Mask R-CNN, Right-to-Left is slightly more damaging than Left-to-Right for doses ranging from 0.4 to 0.8. Table D.11 reports the per-cell mAP results. The shared patterns show that the hierarchy is not solely an artifact of the soft estimand. Directional vulnerability in detection is thus a dose- and estimand-dependent hierarchy with a stable vertical asymmetry, and any single-metric claim of a strongest direction is incomplete. The ordering differs from that observed in the classification setting, where Center-to-Outward has the largest performance drop at every dose (Section 3.1). The relative effects of the directions therefore differ between classification and detection.

Representation and coupling. We examine three measures of representation drift, including per-unit displacement, PCA projection residuals, and intrinsic dimension. As in classification, the representation layer shows clear dose-related changes in nearly all model-by-direction pairs. Per-unit displacement and PCA projection residuals increase with dose, while intrinsic dimension remains within a relatively narrow range. The decision-level hierarchy has no counterpart in the geometry of the drift itself.

Figure D.6 shows that representation displacement increases with dose in both detectors. For Mask R-CNN, Bottom→Top produces the largest displacement at low doses, while Center→Outward leads at higher doses. For Mask DINO, Center→Outward produces the largest displacement from $t = 0.4$ onward. These directional orderings do not reproduce the soft-recall hierarchy in Figure 4.

Dose-stratified cluster-aware HSIC between per-unit representation displacement and decision change rejects independence in all 50 strata after within-detector BH adjustment. Table D.10 reports

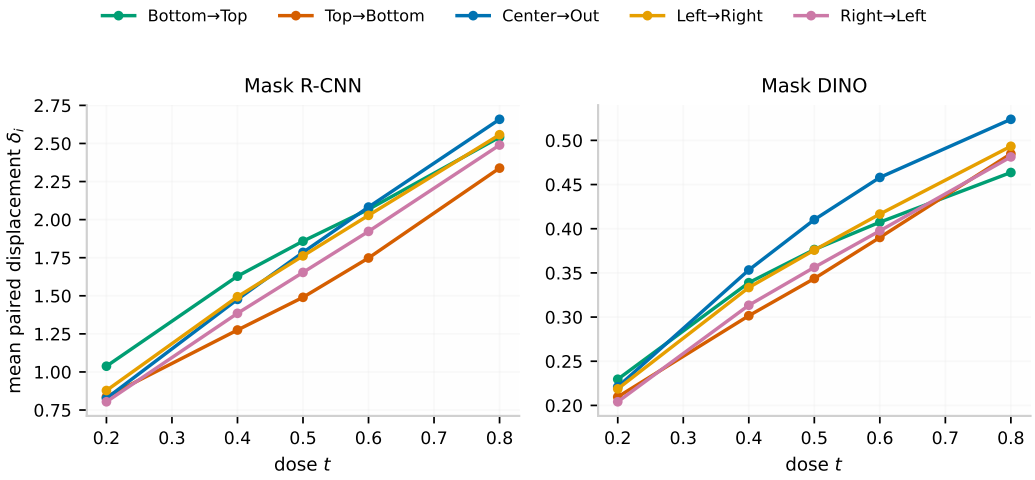


Figure D.6: **Detection representation drift.** Mean paired representation displacement δ_i by dose and direction for Mask R-CNN (left) and Mask DINO (right). The two panels use separate vertical scales, so displacement magnitudes are compared within detectors.

the cross-city coupling statistic in each stratum. All 50 strata remain significant after within-detector BH adjustment, while the corresponding descriptive nHSIC values remain small. The ordering of coupling magnitude also does not reproduce the decision-level hierarchy. Within-dose coupling is markedly weaker in detection than in classification. In a fixed-kernel leave-one-city-out analysis, 47 of 50 strata remain significant after every city deletion. The three exceptions are Center→Outward at $t = 0.2$ for Mask R-CNN and at $t = 0.4$ and $t = 0.6$ for Mask DINO.

Result tables. Tables D.7–D.11 collect the full detection results. Under the primary estimand at $t = 0.8$, Left→Right reaches -0.34 for Mask R-CNN and -0.29 for Mask DINO, compared with -0.27 and -0.23, respectively, for the weakest direction. The city-level permutation test rejects directional homogeneity at every dose ($p < 10^{-5}$). Among the 100 city-level Wilcoxon pairwise comparisons, 92 remain significant after BH correction, including 44 of 50 for Mask R-CNN and 48 of 50 for Mask DINO. Dose-stratified cluster-aware HSIC rejects independence in all 50 strata after within-detector BH adjustment.

D.4 Robustness Checks

Fill-rule sensitivity. We compared the Gaussian boundary fill with four alternatives: black, constant gray, Gaussian noise matched to the image-level channel statistics, and Telea inpainting. The comparison was conducted for ResNet-50 and ViT-B/16 at $t = 0.6$ and at the full-occlusion endpoint $t = 1$.

At $t = 0.6$, the rankings across the five fill rules had Kendall's coefficient of concordance $W = 0.936$ for both models. Center→Outward produced the largest ATE decrease under four of the five fills. Under Telea inpainting, Top→Bottom was slightly more damaging than Center→Outward. Bottom→Top remained the least damaging direction under every fill rule for both models.

The magnitude of the response was more sensitive to the fill. At $t = 1$, the ResNet-50 ATE ranged from -3.02 under constant gray to -4.74 under inpainting, while the ViT-B/16 ATE ranged from -3.83 to -6.32. These results support the stability of the main directional ordering in the two audited models, but they do not imply that effect magnitudes are invariant to the fill rule. Because only $t = 0.6$ and $t = 1$ were evaluated, this experiment does not establish invariance of the complete dose–response trajectory.

Bandwidth and basis sensitivity. The descriptive MMD bandwidth is estimated once per model by the median heuristic on baseline features and held fixed across all directions and doses, so de-

Table D.7: **Detection ATE grid** under the primary estimand (soft recall), for both detectors and five directions ($n = 3{,}457$ paired images). The table reports five directional doses and the shared direction-independent full-occlusion endpoint at $t = 1$.

Detector	Dir.	t=0.2	t=0.4	t=0.5	t=0.6	t=0.8	t=1.0*
Mask R-CNN	B→T	−0.08	−0.15	−0.18	−0.21	−0.28	−0.39
	C→O	−0.05	−0.11	−0.16	−0.21	−0.32	−0.39
	L→R	−0.08	−0.18	−0.23	−0.27	−0.34	−0.39
	R→L	−0.07	−0.17	−0.23	−0.26	−0.33	−0.39
	T→B	−0.07	−0.13	−0.16	−0.20	−0.27	−0.39
Mask DINO	B→T	−0.08	−0.15	−0.18	−0.20	−0.24	−0.28
	C→O	−0.06	−0.13	−0.18	−0.21	−0.26	−0.28
	L→R	−0.09	−0.20	−0.24	−0.26	−0.29	−0.28
	R→L	−0.08	−0.19	−0.22	−0.24	−0.28	−0.28
	T→B	−0.07	−0.12	−0.14	−0.16	−0.23	−0.28

*shared endpoint, direction-independent.

Table D.8: **Detection QIE summary by category.** Each category contains 50 cells from two detectors, five directions, and five doses. Median ATE is the median of the 50 cell-level ATEs. The three q_{90} columns count cells in which the upper decile of the image-level effect distribution is negative, zero, or positive.

Category	n_{img}	med. cell ATE	sign of q_{90}: cells			Regime
			< 0	$= 0$	> 0	
Car	3309	−0.205	46	1	3	near-homogeneous damage
Bicycle	1869	−0.198	0	41	9	resistant upper tail
Person	2728	−0.150	28	11	11	mixed
Motorcycle	604	−0.191	0	40	10	resistant upper tail
Bus	351	−0.177	0	28	22	resistant upper tail
Train	167	−0.001	0	0	50	positive upper tail
Truck	443	0.000	0	50	0	degenerate zero effect

Table D.9: **Detection hierarchy.** Top: orderings under the two evaluation protocols. Bottom: Friedman Q statistics computed from the 21 city-mean response profiles at each dose ($\mathrm{df} = 4$; within-city direction-label permutation $p = 1/200001$ throughout) and city-level Wilcoxon signed-rank comparisons surviving BH correction (92 of 100 overall; 44/50 for Mask R-CNN and 48/50 for Mask DINO). The eight nonsignificant comparisons all involve small between-direction differences; three occur at $t = 0.2$.

Setting	Ordering
Primary estimand, $t \geq 0.4$	L→R > R→L > B→T ≈ C→O > T→B
Primary soft recall, vertical comparison	B→T > T→B in 10/10 cells
F1-bridge, vertical comparison	B→T > T→B in 68/80 cells

Dose	Mask R-CNN Q	Mask DINO Q
$t = 0.2$	65.829	76.990
$t = 0.4$	77.219	80.267
$t = 0.5$	73.105	78.781
$t = 0.6$	68.533	81.714
$t = 0.8$	76.076	79.200

scriptive magnitudes are comparable within a model and are never compared across models. The within-image swap test instead uses a stratum-specific pooled-pair bandwidth that is invariant to every allowed swap. The HSIC bandwidths are pooled per model and likewise fixed. Detection inference uses the cross-city statistic $T_\times$ with city-level Rademacher multipliers; raw Gaussian multipliers give the same 50/50 decisions, and 47 of 50 strata remain significant under every fixed-kernel leave-one-city-out deletion. As a robustness check for classification, we compute distance correla-

Table D.10: **Detection coupling, cluster-aware within-dose paired protocol.** Cross-city statistic $10^3 T_\times$ between per-unit representation displacement $\delta_i = \|\Delta m_i\|_2$ and decision change τ_i within each direction–dose stratum (bandwidths pooled once per model, city-level Rademacher multiplier bootstrap, per-model BH). All 50 strata are significant; the largest adjusted values are 0.042 for Mask R-CNN and 0.020 for Mask DINO. Descriptive nHSIC magnitudes remain small (0.0017–0.0406).

Detector	Dir.	t=0.2	t=0.4	t=0.5	t=0.6	t=0.8
Mask R-CNN	B→T	0.825	0.855	0.372	0.303	0.771
	C→O	0.017	0.102	0.360	0.458	0.938
	L→R	0.360	0.380	0.373	0.365	1.237
	R→L	0.627	1.397	0.718	0.989	0.762
	T→B	0.095	0.211	0.331	0.675	0.377
Mask DINO	B→T	0.179	0.501	0.696	0.952	0.959
	C→O	0.142	0.239	0.416	0.285	0.610
	L→R	0.389	0.441	0.826	0.762	0.561
	R→L	0.416	0.657	0.504	1.132	0.432
	T→B	0.151	0.290	0.888	0.360	0.290

Table D.11: **Hard-metric dose–response.** Dataset-level COCO-protocol mAP (IoU 0.5:0.95, seven Cityscapes classes) per cell, computed with pycocotools on the stored per-image predictions; in parentheses, relative decay from the clean baseline (0.203 Mask R-CNN, 0.212 Mask DINO). The direction hierarchy of the soft estimand is reproduced: Bottom→Top leads at t=0.2, horizontal directions dominate at t=0.8, Center→Outward never leads.

Detector	Dir.	t=0.2	t=0.4	t=0.5	t=0.6	t=0.8
Mask R-CNN	B→T	0.137 (−33%)	0.090 (−56%)	0.078 (−62%)	0.070 (−66%)	0.048 (−76%)
	C→O	0.178 (−13%)	0.144 (−29%)	0.125 (−38%)	0.103 (−49%)	0.060 (−71%)
	L→R	0.146 (−28%)	0.079 (−61%)	0.060 (−70%)	0.056 (−73%)	0.038 (−81%)
	R→L	0.148 (−27%)	0.071 (−65%)	0.049 (−76%)	0.045 (−78%)	0.036 (−82%)
	T→B	0.163 (−20%)	0.120 (−41%)	0.112 (−45%)	0.102 (−50%)	0.068 (−67%)
Mask DINO	B→T	0.134 (−37%)	0.092 (−57%)	0.083 (−61%)	0.074 (−65%)	0.057 (−73%)
	C→O	0.177 (−17%)	0.134 (−37%)	0.111 (−48%)	0.091 (−57%)	0.066 (−69%)
	L→R	0.134 (−37%)	0.052 (−75%)	0.038 (−82%)	0.043 (−80%)	0.036 (−83%)
	R→L	0.140 (−34%)	0.060 (−72%)	0.047 (−78%)	0.051 (−76%)	0.044 (−79%)
	T→B	0.154 (−27%)	0.106 (−50%)	0.103 (−52%)	0.099 (−53%)	0.074 (−65%)

tion on the same strata. It is significant in all 148 distinct classification settings after BH adjustment, with $dCor$ ranging from 0.24 to 0.74. The Spearman correlation between distance correlation and nHSIC is 0.61. The shared PC-256 basis is fit on baseline features only, so no stratum defines its own geometry.

Use with other operators. The same paired estimation procedure can be applied to another deterministic operator. The resulting effects would describe that operator and need not equal those reported here.

E Limitations and Future Work

The estimands describe a specified model, intervention operator, and image population. Classification is conditional on successful SAM-3 masking, and the excluded images are harder at baseline (Section 3.1). Detection uses zero-shot models with low baseline mAP, making small response differences harder to resolve. The placebo interpretation adds two substantive assumptions: zero mean signal loss in the background arm and equality of mean artifact responses. The atypicality diagnostic describes differences between the observed arms but does not establish these assumptions. The contrast distribution remains directly observable, while its mean signal interpretation and sensitivity to the reference are governed by Section 2.3.

Two extensions follow naturally. The first is to make the latent information dose itself identifiable, rather than reading each direction as converting the nominal dose at its own unknown rate. The second is an instrumental-variable reading that exploits direction as exogenous variation at fixed dose, which would support mechanism claims beyond the descriptive profiles reported here.