PROBE: AUTOMATED SPURIOUS CUE DISCOVERY AND CROSS-ARCHITECTURE SHORTCUT SCORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Image classifiers often rely on shortcuts, yet systematically discovering and comparing such dependencies across classifiers remains challenging. We introduce PROBE, a fully automated two-stage auditing pipeline requiring no predefined attributes, group labels, or retraining. The first stage, DISCOVER, names the visual cues a classifier relies on for each class and tests each cue through a targeted counterfactual edit, measuring the resulting change in classifier confidence to separate object-intrinsic from spurious cues. The second stage, SCORE, turns the discovered cues into rendered probe images that show the object with its context, the context alone, the object alone, or the object under stacked confuser cues, and reports object-only accuracy, shortcut-stress accuracy, and confidence shift for each classifier. Across five ImageNet classifiers evaluated on a shared candidate set, 86% of shortcut cues that pass each classifier's own confidence threshold are flagged by at least two classifiers and 31% by all five. Probes that show the context alone rarely recover the label after filtering detected render leakage. The discovered catalog recovers 84% of human-annotated reference cues and achieves higher measured per-phrase precision against Salient ImageNet labels than three annotation-free baselines. Debiasing a classifier moves the audit's scores in the expected direction, while an untrained classifier yields no substantive findings, indicating that the scores respond to classifier behavior rather than being determined solely by the tools that construct the probes. The pipeline scales to thousand-class and open-vocabulary settings, where the extreme pairwise differences on the accuracy-based axes reach statistical significance.

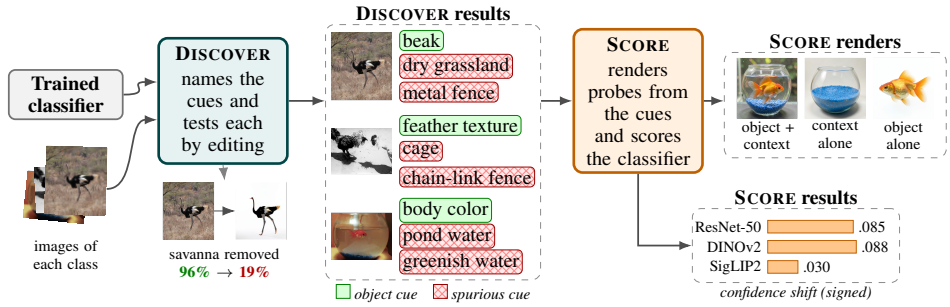


Figure 1: PROBE audits a classifier in two stages. DISCOVER proposes cues and tests each with a counterfactual edit to separate object-intrinsic from spurious cues[1]. SCORE turns the resulting cues into shared probe images and measures how classifier predictions change with and without context.

1 INTRODUCTION

Deep neural network models often exploit spurious correlations present in training data, such as contextual cues, background patterns, or co-occurring objects that are statistically associated with a class but not causally related to it (Geirhos et al., 2020). A classifier may learn to associate *ocean*

[1]We define a spurious cue as any cue that the classifier relies on disproportionately due to dataset bias, including external context and intrinsic cues whose specific state is spuriously correlated with the class.

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.

background with *sharks* or *glass bowls* with *goldfish*, achieving high accuracy on in-distribution benchmarks while failing under distribution shift (Xiao et al., 2021; Beery et al., 2018).

Many shortcut mitigation methods rely on predefined attributes or group annotations (Sagawa et al., 2020), and therefore address shortcuts that must be specified or represented in advance. Explanation methods can localize or characterize evidence associated with a prediction (Selvaraju et al., 2017; Wang et al., 2020), and they can also be audited for faithfulness (Adebayo et al., 2018) or used to diagnose spurious correlations (Sun et al., 2023), but these approaches do not automatically discover a named cue and test the audited classifier's dependence on that cue through the targeted counterfactual procedure used by PROBE. Counterfactual editors test how classifiers respond to visual changes but need a user-specified target class (Augustin et al., 2022; Jeanneret et al., 2022) or edit along a fixed set of caption perturbation types (Prabhu et al., 2023). Existing approaches do not jointly discover open-vocabulary cues, test each cue through a targeted counterfactual intervention on the audited classifier, and convert the resulting cue inventory into a shared probe set for cross-architecture comparison (§2).

We propose PROBE, a fully automated two-stage pipeline that first identifies visual cues influencing a classifier's predictions, then evaluates the classifier on synthetic probes derived from those cues. Neither stage requires domain expertise, predefined attribute lists, group labels, or model retraining. Unlike instance-level explanation approaches such as LIME and SHAP (Ribeiro et al., 2016; Lundberg & Lee, 2017), our pipeline operates at the class level and separates discovery from comparison. DISCOVER analyzes a single classifier and assigns each discovered visual cue to one or more of four categories, namely real (object-intrinsic), bias-spurious, bias-state, and inconclusive (Figure 1). SCORE then uses these cues to generate four probe variants per class, systematically varying the presence of the target object and its contextual cues. It evaluates each classifier using top-1 accuracy on each probe variant together with a confidence-shift score, defined as the difference in target-class confidence between bias-heavy and real-cue-only probes, intended to capture the confidence difference associated with contextual cues.

PROBE characterizes relative behavior on the audited probe set rather than overall model performance and reports the counterfactual evidence underlying each discovered cue. Across five ImageNet classifiers evaluated on a shared candidate set, 86% of shortcut cues that pass each classifier's own confidence threshold are flagged by at least two classifiers and 31% by all five (§4.1).

Our main contributions are as follows:

- **Counterfactual cue discovery.** DISCOVER uses a vision-language model (VLM) conditioned on the classifier's attention map to propose candidate cues, generates a diffusion-based counterfactual for each cue, and separates object-intrinsic cues from contextual ones, without requiring predefined attributes or group annotations.
- **Controlled probe generation.** SCORE aggregates the discovered object and spurious cues across the classifier pool and uses a fixed-seed text-to-image generator to create a shared set of controlled probe variants for evaluation.
- **Three-axis cross-architecture comparison.** Evaluating classifiers on the shared probe set yields three complementary diagnostic axes: object-only accuracy, shortcut-stress accuracy, and confidence shift. Because confidence scales differ across model families, confidence shift is compared across families by rank only (§4.3).
- **Validation and scope.** We validate DISCOVER's cue recovery against human-annotated and dataset-defined reference cues (§4.2) and evaluate whether the audit reflects classifier behavior and training procedure rather than artifacts of the generative pipeline (§4.3). We further test the stability of the confidence-shift ranking under an alternative concept source (§4.3) and demonstrate scalability to thousand-class and open-vocabulary settings (§4.1).

2 RELATED WORK

Shortcut learning and spurious correlations. Documented shortcuts (Geirhos et al., 2020) include texture bias (Geirhos et al., 2019), background reliance (Xiao et al., 2021), medical-imaging artifacts (DeGrave et al., 2021), and cues inherited from pretraining (Wang et al., 2024). Standard training can exploit predictive but non-robust features (Ilyas et al., 2019). Mitigation either uses attributes, groups, or environments (Sagawa et al., 2020; Kirichenko et al., 2023; Arjovsky et al., 2019), or avoids group labels without naming the shortcut (Liu et al., 2021a; Arefin et al., 2024). Prior works do not combine open-vocabulary cue discovery, per-cue counterfactual testing, and shared cross-architecture scoring.

Explainability and attribution methods. Saliency and class-activation-map (CAM) methods (Zeiler & Fergus, 2014; Simonyan et al., 2014; Smilkov et al., 2017; Zhou et al., 2016; Selvaraju et al., 2017; Chattopadhay et al., 2018; Wang et al., 2020) localize spatial evidence associated with a model's prediction but do not name the underlying concept. Concept-level methods name it but rely on predefined concept sets, as in TCAV (Kim et al., 2018) and Network Dissection (Bau et al., 2017), or return unnamed activation factors that need a post-hoc naming step, as in CRAFT (Fel et al., 2023). Concept-based approaches either train explicitly interpretable predictors, as in VLG-CBM (Srivastava et al., 2024), or extract and characterize concepts represented by pretrained VLMs (Zang et al., 2025). DEAL (Li et al., 2024) instead disentangles and localizes concept-level explanations of existing VLMs and studies their relationship to spurious correlations. These approaches do not construct per-cue counterfactual tests and a shared probe set for comparing arbitrary trained classifiers. TextCAM (Zhao et al., 2025) produces open-vocabulary textual descriptions of CAM evidence but does not perform a targeted counterfactual intervention for each proposed cue.

Counterfactual explanations. Counterfactual visual explanations ask what minimal semantic change would alter a model's prediction (Wachter et al., 2017; Goyal et al., 2019), first with GAN-based approaches (Lang et al., 2021; Sauer & Geiger, 2021; Singla et al., 2020) and now with diffusion-based editors (Jeanneret et al., 2022; Augustin et al., 2022; 2024; Kim et al., 2025; Guo et al., 2026), with counterfactual image generation itself now benchmarked (Melistas et al., 2024). Language-controlled editing (Brooks et al., 2023; Hertz et al., 2023; Jeanneret et al., 2024) adds semantic guidance, but each approach addresses only part of the problem. Gradient-guided methods such as DiME and DVCEs require user-specified edit targets, and language-guided methods such as LANCE (Prabhu et al., 2023) edit along a fixed set of caption perturbation types. GIFT (Zablocki et al., 2026) and UNBOX (Carnemolla et al., 2026) produce global or class-wise textual explanations without evaluating a classifier pool on a common controlled probe set. Agent-based systems (Shaham et al., 2024; Li et al., 2025) automate iterative hypothesis generation and testing on trained models, but do not turn the discovered cues into a shared controlled probe set for cross-architecture scoring.

Slice discovery and model debugging. Slice-discovery methods return underperforming groups from validation-set errors (Eyuboglu et al., 2022; Jain et al., 2023), with recent systems adding interpretable slice descriptions and predefined human-understandable dimensions (Chen et al., 2025; Gannamaneni et al., 2025). These methods are centered on identifying underperforming or error-associated subsets of data, whereas PROBE can intervene on a named cue and measure reliance even for examples the classifier currently predicts correctly.

3 METHOD

Problem Formulation. Let h be a trained image classifier over label space $\mathcal{Y} = \{1, \ldots, C\}$ and image space $\mathcal{X}$, and let $\mathcal{D}$ be a labeled image set over $\mathcal{Y}$. Let $p(c \mid x) \in [0, 1]$ denote the classifier's confidence for class $c \in \mathcal{Y}$ on image $x \in \mathcal{X}$. For each class c we seek an assignment of the named visual cues that h relies on when predicting c to four potentially overlapping categories. These categories are real (object-intrinsic cues), bias-spurious (spurious context cues), bias-state (state-dependent cues such as pose, lighting, or framing), and inconclusive (cues for which the audit reaches no decision). This per-classifier assignment $\mathcal{F}_c^{(h)}$ is the output of DISCOVER, and §3.1 describes how each cue is assigned to a category.

3.1 DISCOVER: PER-CLASSIFIER CUE DISCOVERY

DISCOVER answers one question for a given classifier and class: Which visual cues does the classifier actually rely on, and which are object-intrinsic rather than spurious? An attention map indicates where the classifier attends, but not whether the attended content affects its prediction. Similarly, a VLM can hypothesize candidate cues that may matter, but cannot establish which ones the classifier actually relies on. DISCOVER closes this gap by treating each hypothesized cue as a testable claim. It applies a counterfactual edit to each cue and measures the resulting change in classifier confidence. The audited classifier's confidence change provides the quantitative evidence for each decision. The VLM identifies the cue type and edit operation. Deterministic rules assign a category when applicable, and for cases not covered by a rule the VLM provides the category assignment.

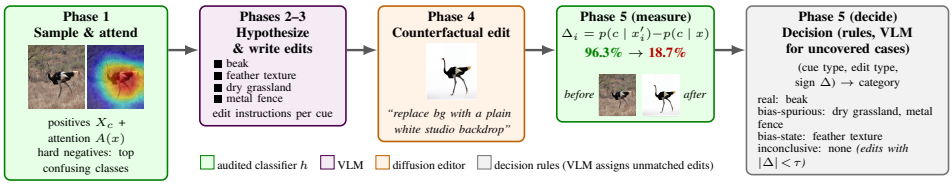


Figure 2: **DISCOVER for one classifier, illustrated on an ostrich image from a pilot run.** A VLM proposes named cues conditioned on the classifier's attention map, a diffusion editor renders one counterfactual per cue, and confidence changes gate the decision.

DISCOVER proceeds in five phases for a single (classifier h, class c) pair. Figure 2 visualizes the flow, and App. A gives the pseudocode.

Phase 1: Baseline sampling and attention. We sample positive images X_c labeled c in $\mathcal{D}$ and *hard negatives* X_c^- drawn from the most confusing classes, defined as the classes that most often appear among h's top-5 predictions on X_c. For every positive image x we compute an attention map $A(x)$ using Score-CAM (Wang et al., 2020) (or Grad-CAM (Selvaraju et al., 2017) / Grad-CAM++ (Chattopadhay et al., 2018)), highlighting the regions associated with h's prediction. The attention map serves only as a localization prior (App. G.16), and whether a cue is considered decision-relevant is determined by the counterfactual edit gate.

Phases 2–3: VLM cue hypothesis and edit instructions. The VLM φ is invoked *per image*. Each positive $x \in X_c$ is queried twice, once with its attention map $A(x)$ to identify target cues and once without it to identify environmental cues. Each hard negative $x^- \in X_c^-$ is queried once with $A(x^-)$ computed for class c, yielding image-level *candidate cues* $\tilde{\mathcal{F}}_c(x) = \varphi\big(x,\, A(x);\, \mathcal{P}_{\text{discover}}\big)$. The per-class candidate cue catalog is then assembled across images as $\tilde{\mathcal{F}}_c = \bigcup_{x \in X_c \cup X_c^-} \tilde{\mathcal{F}}_c(x)$, with duplicates merged by exact name. $\mathcal{P}_{\text{discover}}$ instructs φ to separate target-class cues, environmental cues, and cues shared with the hard negatives, with the latter serving as primary candidates for spurious correlations. For each hypothesized cue $f_i \in \tilde{\mathcal{F}}_c$ and each edit type applicable to its cue type (App. C.2), φ writes one or more concrete edit instructions $e_i = \varphi\big(x_c,\, f_i;\, \mathcal{P}_{\text{edit}}\big)$ conditioned on a representative image x_c of the class.

Phases 4–5: Counterfactual editing and category assignment. For each cue f_i with edit instruction e_i we produce a counterfactual image $x_i' = G(x,\, e_i)$ with an instruction-guided diffusion editor G. For every counterfactual image x_i' we record the confidence delta $\Delta_i = p(c \mid x_i') - p(c \mid x)$. Edits with $|\Delta_i| \geq \tau$ enter a decision step in which a rule set maps (cue type, edit type, sign of Δ) to one of $\{\text{essential},\ \text{spurious},\ \text{state-bias},\ \text{unknown}\}$ (full rule set in App. C.3). An edit that matches no rule is recorded as unknown under the strict-rules regime and receives a VLM category assignment in the as-run regime (Implementation note below). After every edit on every cue in $\tilde{\mathcal{F}}_c$ receives a decision, cues are assigned to the four categories of Eq. (1):

$$\mathcal{F}_c^{(h)} \;=\; \text{real}_c^{(h)} \;\cup\; \text{bias-spurious}_c^{(h)} \;\cup\; \text{bias-state}_c^{(h)} \;\cup\; \text{inconclusive}_c^{(h)}, \tag{1}$$

where a cue is assigned to a category if at least one of its edits receives the corresponding decision (essential to real, spurious to bias-spurious, state-bias to bias-state). A cue whose edits yield decisions from multiple categories is assigned to each corresponding category. The inconclusive category contains cues whose edits never change confidence by at least τ or receive only unknown decisions.

Implementation note. In our runs, edits not covered by a deterministic rule receive a VLM category assignment instead of being recorded as unknown; this as-run regime covers most classifier-edit pairs (78.6% on the ImageNet-100 audit of §4). The agreement figures of §4.1 count a vote only when the classifier's own $|\Delta_i|$ crosses τ. The VLM therefore assigns the category after gating; it does not determine whether a case passes the confidence threshold. Under the strict-rules-only regime, the probe-stage results and the dominance of bias-spurious are unchanged; only the absolute inventory sizes change (App. G.13).

3.2 SCORE: PROBE-BASED SHORTCUT SCORING

DISCOVER produces a per-class list of named cues, which identifies what a classifier relies on but not how strongly that reliance affects its behavior. SCORE answers that question. It uses the discovered real and spurious cues to render controlled probe images designed to vary the presence of the object

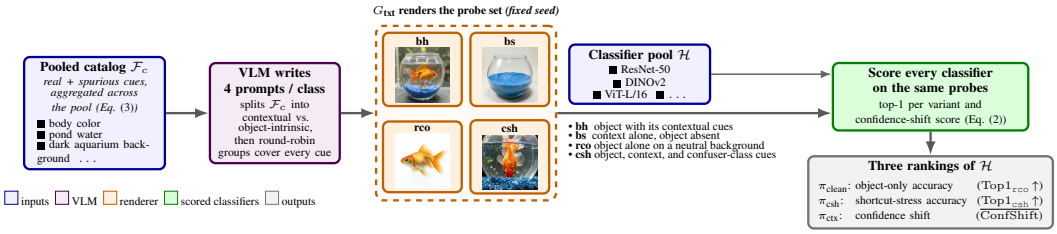


Figure 3: **SCORE on a pool.** Inputs are the class-level cue set and the classifier pool. A VLM writes four prompts per class, a fixed-seed renderer produces the four probe variants, every classifier is scored on the same probes, and three rankings follow.

and its context, and then evaluates the classifier on those images. Because each variant is designed to control what is visible, its accuracy measures a different form of reliance, turning the qualitative cue catalog into a quantitative evaluation.

SCORE takes the audited classifier h and its class-level cue set $\mathcal{F}_c = \text{real}_c^{(h)} \cup \text{bias-spurious}_c^{(h)}$, formed from the two categories of Eq. (1) that identify the cues on which h relies. Its output is four probe accuracies and one confidence-shift score. Rendering is performed once per class, so one probe set scores every classifier over $\mathcal{Y}$ (§3.3).

Probe synthesis. For every class c, the VLM φ reads $\mathcal{F}_c$ and writes natural-language prompts for four probe variants. The text-to-image generator G_{txt} renders each prompt with a fixed seed. The bias-heavy variant (`bh`) renders the target together with the contextual cues selected from $\mathcal{F}_c$. The bias-stripped variant (`bs`) renders the same contextual cues *without* the target, serving as a false-positive probe. The real-cue-only variant (`rco`) shows the target on a neutral background with only object-intrinsic cues visible, serving as the object-only reference. The confuser-shortcut variant (`csh`) renders the target under unusual pose or lighting together with its contextual cues and cues from the top confusing classes identified in Phase 1, providing a composite stress test for a shortcut-reliant classifier. Although cues in $\mathcal{F}_c$ are assigned to the real and bias-spurious categories, the VLM may regroup cues as contextual or object-intrinsic when composing each prompt, to keep each prompt semantically consistent with the requested probe variant. The contextual cues are assigned round-robin to N_c prompt groups so that every cue appears in at least one prompt per variant. The number of prompt groups N_c is a per-audit configuration choice (App. G.11).

Probe metrics. Let $\Pi_c^{\texttt{bh}}, \Pi_c^{\texttt{bs}}, \Pi_c^{\texttt{rco}}, \Pi_c^{\texttt{csh}}$ denote the four probe sets for class c. The classifier is run over every probe image. For each variant $v \in \{\texttt{bh}, \texttt{bs}, \texttt{rco}, \texttt{csh}\}$ we compute the class-averaged top-1 accuracy $\text{Top1}_v(h)$ (Eq. (4) in App. G.1) and the class-averaged *confidence-shift score*:

$$\overline{\text{ConfShift}}(h) = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \big(\bar{p}(c \mid \Pi_c^{\texttt{bh}}) - \bar{p}(c \mid \Pi_c^{\texttt{rco}})\big), \quad \text{where } \bar{p}(c \mid \Pi) = \frac{1}{|\Pi|} \sum_{x \in \Pi} p(c \mid x). \tag{2}$$

3.3 AUDITING A CLASSIFIER POOL

Both stages extend to a pool $\mathcal{H} = \{h_1, \ldots, h_M\}$ of classifiers sharing the label space $\mathcal{Y}$, where $p_m(c \mid x)$ denotes the class confidence of h_m. DISCOVER runs once and assigns categories separately for each classifier, and SCORE evaluates every classifier on one shared probe set. Figure 3 shows the pool flow.

DISCOVER on a pool. When the pool is available at discovery time, Phases 1–4 run once on a reference classifier h_1 and Phase 5 scores every h_m on the shared edits, recording $\Delta_{i,m} = p_m(c \mid x_i') - p_m(c \mid x)$ and one decision for each (cue, edit, classifier) triple. A single edit can therefore receive different decisions for different classifiers, and the resulting assignments $\mathcal{F}_c^{(m)}$ differ in category membership rather than in the candidate cue set. The τ gate is applied at the pool level (App. B.2). A classifier whose own $|\Delta_{i,m}| < \tau$ therefore still receives a decision when any pool member crosses τ, and such cases are VLM-assigned (App. G.13). The per-classifier assignments are aggregated into a strict-agreement tier (every classifier concurs) and an any-observed tier (at least one classifier flags it, App. B). The class-level cue set passed to SCORE is the any-observed tier

$$\mathcal{F}_c \;=\; \text{bias-spurious}_c^{\text{any}} \cup \text{real}_c^{\text{any}}, \tag{3}$$

with duplicates merged by exact name. The strict-agreement tier is reported as a diagnostic of the degree of overlap among classifier inventories (§4.1) and is not used to construct probes.

SCORE on a pool. Every classifier $h_m \in \mathcal{H}$ is evaluated on the same probe set, yielding $\mathrm{Top1}_v(h_m)$ and $\overline{\mathrm{ConfShift}}(h_m)$. Because each probe variant targets a different failure mode, we report three orderings over $\mathcal{H}$ instead of one scalar. π_{clean} sorts $\mathcal{H}$ by $\mathrm{Top1}_{\texttt{rco}}$, descending (object-only accuracy). π_{csh} sorts by $\mathrm{Top1}_{\texttt{csh}}$, descending (shortcut-stress accuracy under stacked confuser cues). π_{ctx} sorts by $\overline{\mathrm{ConfShift}}$, *ascending* (confidence shift). The sort is on the signed value, so a classifier whose confidence is lower on bias-heavy than on real-cue-only probes ranks ahead of one whose confidence is unchanged. The signed ordering is used for comparisons across concept sources and datasets (§4.3, concept-source substitution), whereas within one backbone the magnitude of the shift is the informative reading (§4.3, discrimination test). App. G.3 discusses the render penalty that contributes to negative shifts. Design rationale is summarized in App. G.15.

4 EXPERIMENTS

Experimental setup. We audit a 100-class subset (ambityga, 2021) of the ImageNet-1k validation split (Russakovsky et al., 2015) on five classifiers spanning the major pretraining regimes, namely RESNET-50 (He et al., 2016), DINOV2-VIT-B/14+LC (frozen features with a linear head) (Oquab et al., 2024), VIT-L/16 (Dosovitskiy et al., 2021), CLIP-VIT-B/32 (Radford et al., 2021), and SIGLIP2-BASE (Tschannen et al., 2025). QWEN2.5-VL-7B-INSTRUCT (Bai et al., 2025) is the VLM and FLUX.2-KLEIN-9B-KV (Labs, 2025) the diffusion editor throughout, with confidence-delta threshold $\tau = 0.15$. Further settings and the statistical protocol are in App. G.1.

4.1 WHAT THE AUDIT REVEALS ABOUT FIVE IMAGENET CLASSIFIERS

Discovered inventories. DISCOVER produces cue inventories of similar size across architectures. In the pool run all five classifiers assign categories to the same candidate set (§3.3), and per-classifier category sizes differ by roughly one cue of each other (as-run regime, App. Table 7), with bias-spurious dominating (≈ 9 cues per class) and real the smallest (≈ 1). The real category is small by design, since an object-intrinsic cue enters it through a single rule: removal must reduce confidence. Cross-classifier agreement depends on how sub-threshold cases are gated. Under the pool-level gate of §3.3, in which a classifier whose own $|\Delta|$ stays below τ still receives a VLM decision, 806 of 994 bias-spurious cues (81%) are flagged by all five classifiers. Requiring each classifier to cross τ on its own edits, 725 of 846 bias-spurious cues (86%) are flagged by at least two classifiers and 262 (31%) by all five. In addition, 71 of 100 classes contain at least one cue on which all five concur (App. G.13). Because the candidate set is shared by construction, these figures measure agreement on the effect of a shared edit, not independent rediscovery. Two prompt-echo strings are excluded from these counts (App. G.2, edit-test cases in App. H.1).

Probe accuracy and confidence shift. SCORE renders the four probe variants of §3.2 for every class, and Table 1 scores all five classifiers on them (an example probe set in Figure 4, further sets in App. H.2). For every classifier, top-1 accuracy drops from ≈ 0.6 on bias-heavy probes to ≈ 0.07 on bias-stripped probes, against a $1/1{,}000$ chance floor, since every classifier predicts over the full ImageNet-1k label space. The residual bias-stripped accuracy is an upper bound on context-only recovery, not a measurement of it. A VLM grader finds the target present in 41 of the 400 bias-stripped renders (10%) and absent from 102 of the 381 real-cue-only renders (27%). After filtering those renders, bias-stripped top-1 falls to 0.011–0.030 and confidence shift ranges from -0.019 to 0.064 while leaving the π_{ctx} ordering unchanged (App. G.12, App. Table 25). Most renders that still fire contain an object the grader missed. On this benchmark, context alone thus rarely triggers the label, and the unfiltered real-cue-only rate is a lower bound on object-only accuracy. The three rankings disagree at the point-estimate level. SIGLIP2-BASE ranks first on π_{clean} and has the lowest confidence shift. DINOV2-VIT-B/14+LC ranks first on π_{csh} but has the largest measured confidence shift. RESNET-50 sits at the bottom of π_{csh}. With 99 resampled classes, only the π_{ctx} extreme, the gap between the top and bottom classifier of that ranking, reaches $p = 0.017$ uncorrected under a paired-bootstrap test ($\Delta = -0.058$, $p = 0.051$ after Holm correction), so no extreme remains significant after correction at this scale.

Scaling audits. When DISCOVER is scaled to all 1,000 ImageNet-1k classes via the long-tailed ImageNet-LT split (Liu et al., 2019), 832 classes contain a discovered spurious cue and enter SCORE. There, the pairwise differences between the observed extrema of all three rankings reach

Table 1: Per-classifier top-1 accuracy on the four probe variants and average confidence-shift score (Eq. (2)). Columns are the four probe variants of §3.2, abbreviated as there. Each cell shows the point estimate over the 99 scored classes (one class contributed no any-observed cue to Eq. (3)) with its 95% bootstrap confidence interval (CI) as a subscript ($n_{\text{boot}} = 10{,}000$, paired class resampling).

Classifier	bh↑	bs↓	rco↑	csh↑	Conf. shift
ResNet-50	$0.626_{[.55,.70]}$	$0.068_{[.04,.10]}$	$0.512_{[.43,.60]}$	$0.462_{[.39,.54]}$	$0.085_{[.04,.14]}$
DINOv2-ViT-B/14+lc	$0.619_{[.54,.69]}$	$0.081_{[.05,.12]}$	$0.524_{[.44,.61]}$	$0.510_{[.43,.59]}$	$0.088_{[.05,.13]}$
ViT-L/16	$0.606_{[.53,.68]}$	$0.076_{[.04,.12]}$	$0.519_{[.43,.61]}$	$0.482_{[.40,.56]}$	$0.061_{[.02,.11]}$
CLIP-ViT-B/32	$0.588_{[.52,.66]}$	$0.068_{[.04,.10]}$	$0.492_{[.41,.58]}$	$0.472_{[.40,.54]}$	$0.055_{[.01,.10]}$
SigLIP2-base	$0.596_{[.52,.67]}$	$0.066_{[.03,.10]}$	$0.545_{[.46,.63]}$	$0.500_{[.42,.58]}$	$0.030_{[-.01,.08]}$

significance ($p < 10^{-3}$), and class frequency shows no detectable association with confidence shift (App. Table 22). The full pipeline also runs on a 1,032-category open-vocabulary LVIS v1 (Gupta et al., 2019) benchmark with a zero-shot pool (App. G.10). Both audits reproduce a large performance drop from bias-heavy to bias-stripped probes, and their bias-stripped rates (≈ 0.13 on ImageNet-LT) are read as leakage-inclusive upper bounds, since those probe sets were not fidelity-filtered. On LVIS, every vision-language classifier posts negative confidence shift.

4.2 DISCOVER RECOVERS REFERENCE CUES

DISCOVER's open-vocabulary catalog names cues that human annotators and dataset designers recognize, and does so more precisely than existing annotation-free discovery methods.

Salient ImageNet protocol. To check whether the catalog aligns with human-labeled saliency-visible cues, we compare against *Salient ImageNet* (SI) (Singla & Feizi, 2022), whose public 1,000-class release (Singla et al., 2022) lists 630 spurious-cue labels across 357 ImageNet-1k classes. We re-run DISCOVER on RESNET-50, the architecture SI's labels were derived from (App. G.4), over the same set of SI classes. Every DISCOVER phrase and SI label is embedded with `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019), and an SI label counts as *surfaced* if its maximum cosine similarity to any catalog entry for the same class exceeds a matcher threshold θ (distinct from the confidence-delta threshold τ). The sweep $\theta \in \{0.30, \ldots, 0.60\}$ was fixed before matching, and held-out F1 selects the grid's lower bound $\theta^\star = 0.30$ in every fold (App. G.4).

Cue-level recovery. At $\theta = 0.30$ the catalog *surfaces* 529 of 630 SI labels (84%, App. Table 10), counting a label as surfaced if it matches an entry in any of the real, bias-spurious, or bias-state categories. Two checks support this. First, a permutation null matches SI labels against the catalogs of *random other* classes and yields $16.1\% \pm 1.6\%$ coverage at the same threshold ($p < 10^{-4}$). Coverage is therefore class-specific signal, not threshold looseness. Second, an exact-phrase lexical baseline reaches 63% and a class-name-stripped token-overlap baseline 69%, so the embedder adds $+15$ to $+21$ percentage points (pp). Calibration and a miss taxonomy of the residual 16% are in App. G.4.

Reference cues beyond Salient ImageNet. Two further settings provide reference cues independent of our stack. Hard ImageNet (Moayeri et al., 2022) curates 15 ImageNet classes for their strong spurious cues and documents the cues. For all 10 classes with an explicitly named cue, DISCOVER's bias-spurious category contains that cue ($10/10$, and the remaining 5 name no cue to match against, App. Table 15). NICO++ (Zhang et al., 2023) photographs everyday categories in six labeled contexts (App. G.6). We audit the 60 categories using two ResNet-50 classifiers, one trained with ERM and one with GroupDRO. For both models, the bias-spurious category names a labeled context for 55 of 60 categories (91.7%, App. Table 16). Both settings evaluate cue *identification* only and do not involve SCORE.

Annotation-free baselines. We benchmark DISCOVER against three annotation-free discovery methods from different paradigms. B2T (Kim et al., 2024) mines bias keywords from captions of a classifier's misclassified images. SpLiCE (Bhalla et al., 2024) decomposes CLIP embeddings into sparse concepts from a 10k-term vocabulary. The full LANCE pipeline (Prabhu et al., 2023) runs on a 40-class subset, scored over its 36 SI-labeled classes. B2T and SpLiCE run on the Salient ImageNet setting above with the same classes, images, matcher, and threshold sweep (SpLiCE never sees RESNET-50, so its recall against RESNET-50-derived labels is a cross-model comparison). DiME (Jeanneret et al., 2022) and DVCEs (Augustin et al., 2022) carry no text at any stage and cannot be scored against SI's named labels, leaving LANCE as the runnable pipeline-level comparison.

Table 2: Salient ImageNet baseline comparison (357 classes, RESNET-50, $\theta = 0.30$). Label-/catalog-side denote the two matching-direction coverages (definitions in App. G.7). Precision is per-phrase macro precision and phrases are per class. †LANCE uses a 40-class subset (36 SI-labeled).

Method	Label-side↑	Catalog-side↑	Precision↑	Phrases↓	Edit test
PROBE (DISCOVER)	**44.9**% (283)	**34.4**% (217)	**0.200**	**4.9**	yes
SpLiCE (Bhalla et al., 2024)	43.2% (272)	30.3% (191)	0.052	≈14.7	no
B2T (Kim et al., 2024)	32.4% (204)	22.2% (140)	0.081	≈8.9	no
LANCE (Prabhu et al., 2023)	13.3%† (24)	—	0.098	≈9.9	no

Table 2 reports coverage in both matching directions, scored against each method's spurious phrases alone and not its full catalog, and our label-side entry accordingly sits below the 84% above. DISCOVER achieves the highest recall in both directions by 1.7–4.1pp over SpLiCE. The clearest differences are in per-phrase precision (0.200 vs. 0.052/0.081/0.098, i.e. 2–4×), parsimony (≈4.9 phrases/class vs. ≈14.7/8.9/9.9), and the interventional per-cue edit test, which none of the baselines performs (all three produce non-interventional cue lists). Because SI's label set is incomplete, these precisions are lower bounds for every method alike, and a matched five-phrase budget preserves the gap (App. G.7). LANCE trails on both axes (13.3% label-side recall on its subset at 0.098 precision), with its sensitive edits dominated by domain and style perturbations instead of object-level shortcuts.

4.3 SCORE RESPONDS TO CLASSIFIER BEHAVIOR

Null control. Re-scoring the archived ImageNet-100 edits and probes of §4.1 with an untrained (random-init) RESNET-50 yields probe accuracy at or below the chance floor, confidence shift 0.000 to three decimals, and *zero* findings in every substantive category, since no edit moves the near-uniform confidence by τ, so every cue is inconclusive and the VLM assignment path never fires (App. Table 28). The stack alone therefore yields no substantive findings under this null control.

Waterbirds sign agreement. Waterbirds (Sagawa et al., 2020) is a binary task with a known background shortcut (water vs. land), providing a controlled setting in which the expected shortcut is specified in advance. We run PROBE on a 14-classifier pool, comprising an ERM-trained RESNET-50 from the LADDER repository (Ghosh et al., 2025) together with four CLIP and nine SIGLIP2 variants (full roster and per-classifier results in App. G.8). *Every one of the* 14 *classifiers posts a negative confidence shift* (range $[-0.241, -0.054]$, mean -0.156), while the pooled any-observed catalog contains 83 bias-spurious cue entries across the two Waterbirds classes, dominated by the canonical background (App. G.8). The negative sign is not evidence against background reliance. Under these rendered probes the informative observation is the consistent response direction across classifiers, while the sign itself reflects the probe construction discussed in App. G.3 and App. G.9. The result is the 14/14 sign agreement, not any magnitude (§4.3).

Discrimination test. To isolate sensitivity to robustness interventions from architectural differences, we train two debiased variants of the same ERM RESNET-50 backbone audited above, last-layer retraining on group-balanced data (DFR, Kirichenko et al., 2023) and GroupDRO fine-tuning (Sagawa et al., 2020), with the recipes in App. G.8. All three arms are scored on the *identical, already-rendered* Waterbirds probe set. App. Table 18 pairs each arm's SCORE metrics with its worst-group accuracy (WGA) on the real Waterbirds test split ($n = 5{,}794$). Both interventions raise real-image WGA (0.751 to 0.908 for DFR and 0.858 for GroupDRO) and move both SCORE metrics in the expected direction. Every Waterbirds shift is negative. Within one backbone we therefore interpret the magnitude of the shift, not the signed pool ordering of §3.3, since a debiased classifier should move toward zero shift from either side. |Confidence shift| falls from 0.212 to 0.103 (DFR) and 0.015 (GroupDRO), and confuser-shortcut top-1 rises under both.

Concept-source substitution. Here we replace DISCOVER's candidate-cue source entirely. In a reduced instantiation of the Z-CBM retrieval pipeline (Yamaguchi et al., 2026), we retrieve 12 concepts per class from a Conceptual Captions phrase bank (Sharma et al., 2018) by CLIP similarity (Radford et al., 2021) and group them as intrinsic or contextual with QWEN2.5-VL (Bai et al., 2025) (App. G.10). The catalog is handed to SCORE unchanged, with the same 100 classes and pool, 1,427 rendered probe images, and DISCOVER not run at all (App. Table 20). The confidence-shift ordering agrees with Table 1 at Spearman $\rho = 0.9$, with a single inversion (RESNET-50 and VIT-L/16, whose intervals overlap almost entirely). Matching the ViT-L/16 checkpoint to Table 1 (App. G.11) moves



Figure 4: Representative SCORE probes for the Waterbirds waterbird class. Two renders are shown for each variant: bias-heavy, bias-stripped, real-cue-only, and confuser-shortcut. The same probe set is evaluated by all 14 classifiers. Full probe sets and per-classifier results are reported in App. H.2 and App. Table 19.

it from fourth to fifth and gives $\rho = 0.7$. The π_{ctx} extreme is significant under both checkpoints ($p < 10^{-3}$), SIGLIP2-BASE has the lowest confidence shift in both, and $\pi_{\text{clean}}/\pi_{\text{csh}}$ are non-significant in both, as in §4.1. The ordering that reaches uncorrected significance in §4.1 is therefore preserved under this alternative concept source, although the magnitudes and the signs of individual shifts change (lower absolute levels, a higher bias-stripped floor, App. G.10). Concepts written by the VLM from the class name alone, without images, barely overlap DISCOVER's catalog and weaken the shortcut signal, suggesting that per-classifier discovery is not restating what a class name implies (App. G.17).

Scope of interpretation. Salient ImageNet labels capture saliency-visible cues rather than interventional reliance, so they do not directly validate DISCOVER's counterfactual edit test, and the two yield different per-class assessments. Confidence shift likewise shows little association with real-image group robustness on Waterbirds or ImageNet-9, so it is an audit-specific response measure rather than a surrogate for robustness. Because confidence scales differ across model families, we interpret confidence shift magnitude, and the cue inventories gated by the fixed τ, only within an audit and within an architecture family. Across families its ordering is an audit-relative descriptive comparison rather than a calibration-invariant measure of shortcut reliance, and magnitude comparisons rest on the accuracy-based axes. The full analysis, with the per-class SI distribution, the robustness correlations, and the Waterbirds two-class decomposition, is in App. G.9.

5 DISCUSSION AND CONCLUSIONS

We present PROBE, a two-stage pipeline that discovers classifier-relevant visual cues and compares classifiers along three diagnostic axes, without predefined attributes, group labels, or retraining. The 100-class ImageNet audit shows substantial cross-classifier agreement on the effects of shared shortcut candidates. Of the spurious cues that pass a classifier's own confidence threshold, 86% are flagged by at least two classifiers and 31% by all five, indicating substantial agreement on which contextual cues affect classifier confidence, although category assignments differ across architectures. Context alone rarely recovers the label after filtering detected render leakage, and the three SCORE rankings disagree, with their extremes reaching significance when the audit scales to ImageNet-LT.

The audit responds to classifier behavior and training interventions (§4.3). The confidence-shift ordering is preserved under an alternative concept source ($\rho = 0.9$, or 0.7 with the checkpoint matched) but varies across datasets ($\rho = 0.4$ between ImageNet-100 and ImageNet-LT, App. G.10). The ordering thus depends on the audit setting as well as the classifiers, while new architectures can be added to the existing SCORE evaluation at inference-only cost (App. G.14).

Limitations. Absolute cue inventories are stack-dependent (App. F, App. G.17), and render fidelity makes context-only accuracies upper bounds. The ImageNet-LT and LVIS probes were not fidelity-filtered (App. G.12). Masked-autoencoder and adversarially trained regimes remain unaudited, and the ImageNet-LT and LVIS rankings cover only shortcut-bearing classes (App. F).

AI USE STATEMENT

A vision-language model is a central, non-incidental component of the method itself. It performs candidate cue discovery and generates the edit instructions used for counterfactual scoring. In our implementation the same model also assigns categories to counterfactual edits that match none of the decision rules (78.6% of classifier-edit pairs on the ImageNet-100 run), and both the as-run and the strict-rules regimes are reported (§3). This use is described in full in §3 and audited in the appendix, including decision-provenance and component-sensitivity analyses.

Beyond the method, we used large language model assistants for code generation when implementing the pipeline and the analysis scripts. AI-assisted code was checked by unit tests, replay checks against archived run records, and independent recomputation of previously reported numbers. The framework itself, including the two-stage design, the rule set with VLM assignment for rule-uncovered cases, and the probe variants, is the authors' original contribution and was not generated by a language model. Generative models are components of the method. A VLM writes the hypotheses and edit instructions and assigns categories to edits that match no rule, and a diffusion model renders every edit and probe. Outside those pipeline roles, no generative AI was used to produce, select, or alter an experimental result.

We also used large language model assistants on the manuscript itself, both as a reviewing aid that returned comments on drafts (typos, ambiguous phrasing, notation inconsistencies, mismatches between text and tables, incomplete arguments) and for editing passes that shortened or restructured sentences. They also identified and summarized candidate related work, which the authors retrieved and read before citing. The authors reviewed every suggested change, discarded those judged incorrect, and checked every number and claim against the run records. No scientific content, claim, or reported number was determined by this process. We take responsibility for the final content of this paper, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work evaluates existing, publicly available image classifiers on public benchmark datasets. It involves no human subjects, no crowdsourced annotation, and no collection or release of new personal data. The method is an auditing tool. Its intended use is to surface the spurious cues a classifier relies on so that they can be reported and mitigated. The same reliance information may identify conditions under which a deployed classifier is vulnerable to failure, which creates a potential misuse risk. We release no attack artifacts beyond the audit outputs already described, and App. F discusses this risk together with the method's other limitations. Because the discovery stage inherits whatever biases are present in the underlying vision-language model, its outputs should be read as hypotheses for audit rather than as ground-truth statements about a classifier.

REPRODUCIBILITY STATEMENT

Code and the archived run records will be released upon acceptance. The reported ImageNet-100, SI, ImageNet-LT, and LVIS runs used fixed generator seeds but predate the later configurable seeding flag, and they used a feature parser that kept one cue per VLM call (App. G.1), so re-runs from the current implementation reproduce the protocol rather than byte-identical inventories. The release includes the as-run parser and the configurations used for the reported experiments alongside the corrected implementation. The experimental setup needed to reproduce our results is documented in §4, with the full protocol (model checkpoints, datasets and splits, prompt templates, generation and editing settings, and the aggregation and gating rules) given in the appendix. The appendix also records the bootstrap confidence-interval procedure behind the reported intervals and the null-control and fidelity audits used to check that the measured effects are not artifacts of the editing pipeline.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

ambityga. ImageNet100. Kaggle dataset, https://www.kaggle.com/datasets/ambityga/imagenet100, 2021. Accessed 2026.

Md Rifat Arefin, Yan Zhang, Aristide Baratin, Francesco Locatello, Irina Rish, Dianbo Liu, and Kenji Kawaguchi. Unsupervised concept discovery mitigates spurious correlations. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems*, 35:364–377, 2022.

Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. Dig-in: Diffusion guidance for investigating networks - uncovering classifier differences, neuron visualisations, and visual counterfactual explanations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11093–11103, 2024. doi: 10.1109/CVPR52733.2024.01055.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3319–3327. IEEE, 2017.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, June 2023.

Simone Carnemolla, Chiara Russo, Simone Palazzo, Quentin Bouniot, Daniela Giordano, Zeynep Akata, Matteo Pennisi, and Concetto Spampinato. Unbox: Unveiling black-box visual models with natural-language, 2026. URL https://arxiv.org/abs/2603.08639.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Muxi Chen, Chenchen Zhao, and Qiang Xu. Hibug2: Efficient and interpretable error slice discovery for comprehensive model debugging. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l30moNjSY9.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

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*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Sabri Eyuboglu, Maya Varma, Khaled Kamal Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Re. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=FPCMqjI0jXN.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2711–2721, 2023.

Sujan Sai Gannamaneni, Rohil Prakash Rao, Michael Mock, Maram Akila, and Stefan Wrobel. Detecting systematic weaknesses in vision models along predefined human-understandable dimensions. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=yK9pvt4nBX.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Shantanu Ghosh, Rayan Syed, Chenyu Wang, Vaibhav Choudhary, Binxu Li, Clare B Poynton, Shyam Visweswaran, and Kayhan Batmanghelich. Ladder: language-driven slice discovery and error rectification in vision classifiers. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22935–22970, 2025.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International conference on machine learning*, pp. 2376–2384. PMLR, 2019.

Changlu Guo, Anders Nymark Christensen, Anders Bjorholm Dahl, and Morten Rieger Hannemose. Maskdime: Adaptive masked diffusion for precise and efficient visual counterfactual explanations. *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2026.

Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5351–5359, 2019. doi: 10.1109/CVPR.2019.00550.

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)*, June 2016.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_CDixzkzeyb.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=99RpBVpLiX.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian conference on computer vision*, pp. 858–876, 2022.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Text-to-image models for counterfactual explanations: a black-box approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Won Jun Kim, Hyungjin Chung, Jaemin Kim, Sangmin Lee, Byeongsu Sim, and Jong Chul Ye. Derivative-free diffusion manifold-constrained gradient for unified xai. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23795–23805. IEEE, 2025.

Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11082–11092. IEEE, 2024.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, and Inbar Mosseri. Explaining in style: Training a gan to explain a classifier in stylespace. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 673–682, 2021. doi: 10.1109/ICCV48922.2021.00073.

Christy Li, Josep Lopez Camuñas, Jake Thomas Touchet, Jacob Andreas, Agata Lapedriza, Antonio Torralba, and Tamar Rott Shaham. Automated detection of visual attribute reliance with a self-reflective agent. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=LyH2ISbOV8.

Tang Li, Mengmeng Ma, and Xi Peng. Deal: Disentangle and localize concept-level explanations for vlms. In *European Conference on Computer Vision*, pp. 383–401. Springer, 2024.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International conference on machine learning*, pp. 6781–6792. PMLR, 2021a.

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 *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Thomas Melistas, Nikos Spyrou, Nefeli Gkouti, Pedro Sanchez, Athanasios Vlontzos, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios A Tsaftaris. Benchmarking counterfactual image generation. *Advances in Neural Information Processing Systems*, 37:133207–133230, 2024.

Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. *Advances in Neural Information Processing Systems*, 35:10068–10077, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick

Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Viraj Uday Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. LANCE: Stress-testing visual models by generating language-guided counterfactual images. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=BbIxB4xnbq.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BXewfAYMmJw.

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*, pp. 618–626, 2017.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=mDw42ZanmE.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=XVPqLyNxSyh.

Sahil Singla, Mazda Moayeri, and Soheil Feizi. Core risk minimization using salient imagenet. *arXiv preprint arXiv:2203.15566*, 2022.

Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1xFWgrFPS.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Divyansh Srivastava, Ge Yan, and Tsui-Wei Weng. Vlg-cbm: Training concept bottleneck models with vision-language guidance. *Advances in Neural Information Processing Systems*, 37:79057–79094, 2024.

Susu Sun, Lisa M Koch, and Christian F Baumgartner. Right for the wrong reason: Can interpretable ml techniques detect spurious correlations? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 425–434. Springer, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 111–119. IEEE, 2020.

Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. A sober look at the robustness of clips to spurious features. *Advances in Neural Information Processing Systems*, 37:122484–122523, 2024.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gl3D-xY7wLq.

Shin'ya Yamaguchi, Kosuke Nishida, Daiki Chijiwa, and Yasutoshi Ida. Zero-shot concept bottleneck models, 2026. URL https://arxiv.org/abs/2502.09018.

Eloi Zablocki, Valentin Gerard, Amaia Cardiel, Eric Gaussier, Matthieu Cord, and Eduardo Valle. GIFT: A framework towards global interpretable faithful textual explanations of vision classifiers. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=OwhW5MpFmD. Featured Certification.

Yuan Zang, Tian Yun, Hao Tan, Trung Bui, and Chen Sun. Pre-trained vision-language models learn discoverable visual concepts. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Vq0wMFBjo2.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16036–16047. IEEE, 2023.

Qiming Zhao, Xingjian Li, Xiaoyu Cao, Xiaolong Wu, and Min Xu. Textcam: Explaining class activation map with text. *arXiv preprint arXiv:2510.01004*, 2025.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

A PSEUDOCODE

Algorithms DISCOVER and SCORE give the two stages of §3 as pseudocode. Algorithm DISCOVER follows the five phases of §3.1 for one classifier, and Algorithm SCORE gives the pool form of SCORE from §3.3.

Algorithm DISCOVER Per-Classifier Cue Discovery

Require: classifier h, class c, data $\mathcal{D}$, VLM φ, editor G, threshold τ
Ensure: four-category assignment $\mathcal{F}_c^{(h)}$

Phase 1. sample positives X_c and hard negatives X_c^-; compute attention $A(x)$
for each image $x \in X_c$ **do**
    Phases 2–3. φ names cues; writes edit instructions e_i per cue
end for
merge cues across images; deduplicate
for each cue f_i and image x **do**
    Phase 4. $x_i' \leftarrow G(x, e_i)$
    Phase 5a. $\Delta_i \leftarrow p(c \mid x_i') - p(c \mid x)$
    Phase 5b. **if** $|\Delta_i| \geq \tau$: rule set, VLM if no rule fires (§3.1)
end for
assign each cue to categories from its decisions
return $\mathcal{F}_c^{(h)}$

Algorithm SCORE Probe-Based Cross-Classifier Ranking

Require: cue set $\mathcal{F}_c$ (Eq. (3)) and confuser classes (Phase 1) per class, pool $\mathcal{H}$, VLM φ, renderer G_{txt}, seed
Ensure: rankings $(\pi_{\text{clean}}, \pi_{\text{csh}}, \pi_{\text{ctx}})$

for each class c and variant $v \in \{\texttt{bh}, \texttt{bs}, \texttt{rco}, \texttt{csh}\}$ **do**
    φ writes the variant-v prompts from $\mathcal{F}_c$
    G_{txt} renders probe set Π_c^v (fixed seed)
end for
for each classifier $h_m \in \mathcal{H}$ **do**
    compute $\mathrm{Top1}_v(h_m)$ (Eq. (4)) and $\overline{\mathrm{ConfShift}}(h_m)$ (Eq. (2))
end for
$\pi_{\text{clean}} \leftarrow$ sort $\mathcal{H}$ by $\mathrm{Top1}_{\texttt{rco}}$, descending
$\pi_{\text{csh}} \leftarrow$ sort $\mathcal{H}$ by $\mathrm{Top1}_{\texttt{csh}}$, descending
$\pi_{\text{ctx}} \leftarrow$ sort $\mathcal{H}$ by $\overline{\mathrm{ConfShift}}$, ascending
return $(\pi_{\text{clean}}, \pi_{\text{csh}}, \pi_{\text{ctx}})$

B AGGREGATION TIERS AND EDIT GATING

This appendix collects three details deferred from §3, namely (i) why the cross-classifier aggregation tiers of Eq. (3) are needed at all, (ii) the two-stage gate that screens each counterfactual edit before category assignment, and (iii) the role of the threshold τ as a per-edit significance gate.

B.1 WHY AGGREGATE ACROSS CLASSIFIERS

Aggregation is needed only when comparing classifiers on a shared probe set, not when auditing one. DISCOVER on a single h (Eq. (1)) returns a complete partition without any cross-model information, and SCORE can evaluate h directly on probes derived from that partition. The aggregation tiers of Eq. (3) are introduced for the multi-classifier setting because all classifiers in $\mathcal{H}$ must be evaluated on identical probes. Differences in score must reflect classifier robustness, not probe difficulty. The any-observed tier holds cues surfaced by at least one classifier, and is the operational aggregation rule. It is the pooled risk inventory from which the shared probe set is built, so that every classifier is evaluated on identical probes. *It is not a claim that every h_m relies on every pooled cue.* The strict-agreement tier holds cues flagged in the same category by every classifier and is reported as a diagnostic statistic about how much the pool's risk inventories overlap (large on both ImageNet, §4.1, and Waterbirds, §4.3). Eq. (3) uses the any-observed tier throughout, and the strict tier feeds no probes.

B.2 TWO-STAGE EDIT GATE (PHASES 4–5A)

Before per-classifier category assignment, every counterfactual edit x_i' passes a two-stage gate. (i) Phase 4. If G produces an image that is nearly identical to x (MSE similarity $1-\mathrm{MSE}/255^2 \geq 0.99$

on 256×256 downsamples, 0.98 in the ImageNet-100 run of §4.1), the edit is flagged `edit_failed` and excluded from the downstream decision step. (ii) Phase 5a. Otherwise, if no classifier in the pool (the single audited classifier when $|\mathcal{H}| = 1$) sees $|\Delta_{i,m}| \geq \tau$ (default $\tau = 0.15$), the edit is flagged `not_significant`. Only edits that pass both steps reach the per-classifier decision step of Phase 5b in Algorithm DISCOVER. Because this gate is pool-level, a classifier whose own $|\Delta_{i,m}| < \tau$ still receives a decision when another pool member crosses τ. Such cases match no rule and are VLM-assigned in the as-run regime (App. G.13).

B.3 THRESHOLD (τ) AS A PER-EDIT SIGNIFICANCE GATE

The threshold τ acts as a per-edit significance gate, not as a cue-level retention filter. Cues below the gate survive in the catalog but are categorized as inconclusive.

C VLM PROMPT TEMPLATES

This appendix lists every prompt the VLM sees during a run of the pipeline. All prompts are reproduced from the implementation, modulo line wrapping. The only edit is the variant name in App. C.5, where the code's `adversarial` is printed as confuser-shortcut, the name used throughout the paper. Placeholders in braces (e.g. `{class_name}`) are substituted at call time. Every prompt is followed by the same JSON-only instruction:

```
IMPORTANT: Respond with ONLY valid JSON. No markdown, no explanation,
no text before or after the JSON.
```

C.1 ALGORITHM DISCOVER, PHASE 2: CUE HYPOTHESIS

Phase 2 makes two calls per positive image (target, environmental) and one per hard-negative image (negative). Each call sends the image (plus the Score-CAM heatmap for the target and negative prompts) and the prompt below. The VLM returns a JSON list of named cues.

Target cues ($\mathcal{P}^{\text{target}}_{\text{discover}}$). Sent with the original image and the Score-CAM heatmap.

```
Target class: {class_name}

I'm showing you an image and the model's attention heatmap
(red = high focus, blue = low).

List every visual CONCEPT the model focuses on. Give short names
(1-3 words). Include body parts, textures, colors, shapes.

Example output:
{"features": [{"name": "dorsal fin", "attention": "high"},
              {"name": "body stripes", "attention": "medium"},
              {"name": "eye shape", "attention": "low"}]}
```

Environmental cues ($\mathcal{P}^{\text{env}}_{\text{discover}}$). Sent with the original image only (no heatmap), to elicit context that is not part of the object.

```
Target class: {class_name}

List all environmental and contextual elements in this image
that are NOT part of the {class_name} itself. Include: background,
setting, lighting, co-occurring objects, surfaces, other
animals/people.

Example output:
{"features": [{"name": "ocean background"},
              {"name": "fishing net"},
              {"name": "person holding fish"}]}
```

Negative cues ($\mathcal{P}^{\text{neg}}_{\text{discover}}$). Sent with a hard-negative image, its Score-CAM heatmap *computed for the target class*, and the model's confidence on it.

```
Target class: {class_name}
This image is actually a {true_label}, NOT a {class_name}.
The model gives it {confidence:.0%} confidence as {class_name}.

I'm showing you the image and the model's attention heatmap
for {class_name} (red = high focus).

What visual concepts cause the model to confuse this {true_label}
with {class_name}?

Example output:
{"features": [{"name": "similar body shape", "attention": "high"},
              {"name": "blue water", "attention": "medium"}]}
```

The example strings in these prompts leak into the catalog. "similar body shape" and "ocean background" appear as cue names in 73 and 6 of the 100 ImageNet-100 classes (79 of the 1,342 candidate pairs), and App. G.2 excludes both names from its counts.

C.2 Algorithm Discover, Phase 3: Edit Instruction Generation

For every hypothesized cue f_i the pipeline applies the edit-type templates that match its kind and wraps each in a concise instruction prompt sent with a representative class image. Target cues receive `feature_removal` and `state_change`, environmental cues receive `environment_removal` and `environment_change` on a positive image and `negative_environment_add` on a hard-negative image, and negative cues receive `negative_modify`. On ImageNet-100, `negative_environment_add` accounts for 8,862 of the 16,813 edits (53%). The wrapper is identical across types:

```
Target class: {class_name}

{type_prompt}

Rules: NEVER mention "{class_name}". Be SPECIFIC. 50-150 chars each.

Example: {"edits": ["Remove the dorsal fin, blend smoothly with body"]}
```

The six `{type_prompt}` templates below determine the kind of counterfactual edit:

```
feature_removal:
  Generate edit instructions to REMOVE "{feature}" from this image.
  Describe how to blend/fill.

state_change:
  Generate edit instructions to CHANGE THE STATE of "{feature}"
  (size, color, pose, open/closed).

environment_removal:
  Generate edit instructions to REMOVE "{feature}". Replace with
  neutral background.

environment_change:
  Generate edit instructions to REPLACE "{feature}" with something
  completely different.

negative_modify:
  Generate edit instructions to MODIFY "{feature}" to look more like
  what a {class_name} has.

negative_environment_add:
  Generate edit instructions to ADD "{feature}" to the image.
```

C.3 Algorithm Discover, Phase 5: Category Assignment

After the diffusion edit produces x_i' and every classifier in $\mathcal{H}$ has been re-scored, a rule set assigns a per-classifier outcome $\in \{essential, spurious, state-bias, unknown\}$ from the (cue_type, edit_type, sign-of-Δ) triple. Edits matched by no rule receive a VLM fallback assignment in the as-run regime (§3.1, App. C.4, App. G.13). The `target=negative` rule covers both `negative_modify` and `negative_environment_add` edits. The full rule set:

```
PHASE 5 VERDICT RULES (per classifier, |delta| >= tau):

  feature_type=target,        edit_type=feature_removal,
  delta < -tau                                    -> ESSENTIAL
      (body part removed; confidence drops -> the classifier
       was using it to identify the class)

  feature_type=target,        edit_type=state_change,
  delta < -tau                                    -> STATE_BIAS
      (pose/lighting changed on a target part; confidence drops
       -> the classifier requires that specific state)

  feature_type=environmental, edit_type in
      {environment_removal, environment_change},
  delta < -tau                                    -> SPURIOUS
      (background or co-occurring object altered; confidence
       drops -> the classifier was relying on context)

  target=negative, delta > +tau                   -> SPURIOUS
      (non-target image edited; confidence in the target class
       rises -> the classifier was fooled by a shortcut cue)

  any other case                                  -> UNKNOWN
      (excluded from category aggregation)
```

C.4 Algorithm Discover, Phase 5: VLM Fallback Assignment

An edit that matches no rule is sent to the VLM with the original and edited images, the per-classifier confidences, and the prompt below (the code's arrow glyph is printed as `->`). The prompt restates the rule set without the τ threshold, so a case assigned by the VLM can receive a substantive decision on a confidence change below τ (App. G.13). If the call fails or returns no JSON, the outcome for every classifier is recorded as unknown.

```
Class: {class_name}
Feature: {feature_name} (discovered as: {feature_type})
Edit type: {edit_type}
Edit: "{edit_instruction}"
Image: {target class (correctly classified) |
        non-target class (different species)}

Original (first) and edited (second) images shown.

Results:
  - {model_name}: {original:.0%} -> {edited:.0%} (delta: {delta:+.0%})
  ...

CLASSIFICATION RULES:
- If a TARGET body part (fin, eye, scale, beak, etc) is removed and
  confidence DROPS -> ESSENTIAL (model correctly uses it)
- If a TARGET body part state is changed and confidence DROPS
  -> STATE_BIAS (model requires specific pose)
- If an ENVIRONMENTAL element (background, water, person, net) is
  removed and confidence DROPS -> SPURIOUS (model relies on context)
- If an edit on a NON-TARGET image INCREASES confidence -> SPURIOUS
  (model is fooled by superficial features)
```

```
Classify each model: essential | spurious | state_bias

Example: {"per_model": {"{model_name}": {"verdict": "essential",
  "reasoning": "body part removal correctly drops confidence"}}}
```

C.5 Algorithm Score: Four-Variant Probe Prompt Synthesis

For each class c the VLM is given the pooled cue inventory $\mathcal{F}_c$ (split into bias cues, real cues, and the top confusing classes), and writes n text-to-image prompts in each of four variants. The bias cues are pre-grouped round-robin into N_c groups (App. G.11 lists N_c per run) so that prompt i uses group i and every cue reaches at least one prompt per variant.

```
Target class: {class_name}

Biased context split into feature groups -- for bias_heavy and
bias_stripped, each of the {n} prompts MUST use a DIFFERENT group
(prompt i uses group i):
  Group 0: {bias_features_group_0}
  Group 1: {bias_features_group_1}
  ...
Intrinsic/real features of the class: {real_features}
Classes the model confuses with {class_name}: {confusing_classes}

Write {n} text-to-image prompts in EACH of FOUR variants:
  bias_heavy: the {class_name} surrounded by its biased context cues.
  bias_stripped: ONLY the biased context, WITHOUT a {class_name}
    visible. Do NOT use the word '{class_name}' in these prompts.
  real_feature_only: the {class_name} on a plain background showing
    only its real features.
  confuser_shortcut: an image DESIGNED TO TRICK the classifier --
    a {class_name} shown with visual features shared with the
    confusing classes above, under an unusual pose or angle or
    lighting, on the biased context background. Goal: the classifier
    should struggle or misclassify this image.

Each prompt 50-220 chars, photographic style, ONE concrete scene per
prompt. No 'or', 'such as', 'for example'.

Output JSON:
{"bias_heavy": [...], "bias_stripped": [...],
 "real_feature_only": [...], "confuser_shortcut": [...]}
```

If the VLM call fails or the response cannot be parsed, the pipeline falls back to four deterministic per-class templates so the probe set is always populated. In the ImageNet-100 run the fallback fired for one class (`sulphur-crested cockatoo`, which had no probe-feeding cue), and that class was excluded from scoring:

```
bias_heavy:        "A photograph of a {class_name} in its typical
                    environment"
bias_stripped:     "A photograph of an empty environment"
real_feature_only: "A photograph of a {class_name} on a plain white
                    studio background"
confuser_shortcut: "A photograph of a {class_name} in unusual
                    lighting and pose"
```

D Threshold (τ) Ablation

Algorithm DISCOVER treats τ as a per-edit significance gate (default $\tau = 0.15$, §4). An edit whose confidence delta stays below τ is recorded `not_significant`, and a cue all of whose edits stay below τ lands in inconclusive. The choice of τ is currently a single point estimate. Table 3 reports a partial ablation on the 357-class SI run of §4.2. Holding the recorded counterfactual edits fixed

and re-applying the Phase 5 rule set at $\tau \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$, we recompute the per-class fraction of edits that land in each outcome, averaged over the 357 classes (16,739 non-failed edits with recorded deltas).

Table 3: Per-class outcome fraction at six values of τ, on the 357-class SI run (RESNET-50). Each row is a mean over the 357 classes of the edit-level fractions assigned each outcome by the Phase 5 rule set at the given τ. The default $\tau = 0.15$ retains 29% of edits. Lowering to $\tau = 0.05$ retains 47% but admits noisier decisions. Doubling to $\tau = 0.30$ retains only 18%. Bold marks the default threshold, not the best value.

τ	% essential	% spurious	% state-bias	% unknown	% retained
0.05	5.1	16.1	9.8	15.8	46.7
0.10	4.4	12.5	7.7	11.3	35.8
0.15	**4.1**	**10.0**	**6.4**	**8.4**	**29.0**
0.20	3.8	8.3	5.5	6.7	24.3
0.25	3.6	6.9	4.7	5.3	20.5
0.30	3.3	6.0	4.1	4.2	17.7

The outcome fractions vary smoothly with τ. Each outcome loses roughly proportional mass as τ rises, with no inversion in the relative ordering (spurious dominates at every τ, and essential is consistently the smallest non-unknown outcome). The choice of $\tau = 0.15$ as default trades a slight retention drop for a substantial reduction in edits with $|\Delta|$ in the $[0.05, 0.15)$ noise band. The partial ablation supports this default but does not constitute a full sensitivity test.

Multi-classifier completion. Two follow-ups complete the sensitivity picture without regenerating probes. (i) *SI agreement across τ* (Table 4). A rules-only replay over the stored per-edit deltas trades recall for precision smoothly (macro spurious recall $0.301 \to 0.145$, precision $0.180 \to 0.241$ from $\tau = 0.05$ to 0.50), leaving the §4.2 conclusion unchanged at every τ. (ii) *Probe-input stability on ImageNet-100* (Table 5). At every τ above the default, the probe-feeding cue set is a strict subset of the default set (subset fraction 100%), so SCORE's rankings cannot reorder through catalog composition. Raising τ above the default shrinks the probe inputs without substituting new ones (below the default, 4–9% of the pairs are new). The one axis still uncovered is re-rendering the probe stage itself at each τ (the FLUX-bound step).

Table 4: SI agreement across τ (SI-357, RESNET-50, rules-only replay over stored per-edit deltas, matcher threshold fixed at 0.30). "Recovered" counts SI labels matched at $\theta = 0.30$ over the SI labels of classes that retain at least one spurious cue at that τ, so the denominator falls with τ. The $\tau = 0.15$ row differs from Table 11 because this replay uses rules only. Recall trades for precision smoothly and the sign of the §4.2 conclusion holds at every τ. Bold marks the default threshold, not the best value.

τ	Macro spurious recall	Precision	Recovered
0.05	0.301	0.180	163/601
0.10	0.281	0.188	152/596
0.15	**0.247**	**0.195**	**135/591**
0.20	0.234	0.204	127/586
0.30	0.208	0.225	111/572
0.40	0.172	0.231	94/555
0.50	0.145	0.241	78/535

E SALIENT IMAGENET MISS-TAXONOMY LISTING

This appendix lists the 101 Salient ImageNet spurious-cue labels for which our framework's DISCOVER catalog contains no paraphrase at $\theta = 0.30$ (§4.2, Table 13). Each row reports the ImageNet class for which the SI label was written, the first sentence of the SI annotator's description (truncated to ≈85 characters), and the best cosine similarity to any cue in the same class's discovered catalog. The taxonomy categories are hand-coded by a single author. The full untruncated descriptions are available in the supplementary artifact `miss_taxonomy.json`.

Table 5: Probe-input stability across τ (ImageNet-100, pooled catalog, rules-only replay, hence fewer pairs than the as-run any-observed tier). "Probe-feeding pairs" counts (class, cue) pairs in the real and bias-spurious tiers. "Subset frac." is the share of the τ set already contained in the default $\tau = 0.15$ set. A value of 100% at every higher τ means the probe inputs shrink monotonically with no substitutions, so SCORE rankings cannot reorder through catalog composition. Bold marks the default threshold, not the best value.

τ	Probe-feeding pairs	Retention of $\tau = 0.15$ set	Subset frac.
0.05	1,116	100.0%	91.2%
0.10	1,056	100.0%	96.4%
0.15	**1,018**	**100.0%**	**100.0%**
0.20	960	94.3%	100.0%
0.30	829	81.4%	100.0%
0.40	667	65.5%	100.0%
0.50	528	51.9%	100.0%

Table 6: The 101 unmatched SI spurious-cue labels at $\theta = 0.30$, grouped by hand-coded category. "Best sim." is the maximum cosine to any cue in the same class's discovered catalog (any category). A value of 0.000 means the class's catalog held no entry to compare against.

Class	SI annotator description (truncated)	Best sim.
Vegetation / flowers / trees (29)		
sloth bear	focus is on the background of sloth bear	0.292
llama	Most of the images are mainly focused on background only.	0.288
house finch	Tree branch is not a part of bird	0.284
box turtle	focus is on the ground in the background of the turtle	0.284
dogsled	Focus is on the high trees in the background of the dogsled.	0.282
admiral	Focus is highlighting the flowers around the butterfly	0.277
dogsled	focus is on the trees which is a background of dogsled	0.266
lesser panda	most of the images focusing on some other object , not main object.	0.261
agama	Focus is on the face of the agama.	0.260
junco	The branches below the birds are highlighted in red.	0.258
junco	The main object is junco bird but focus is on other object like branch.	0.257
golfcart	focus is on sky, greenery and walls in the background of golfcart	0.256
ptarmigan	The focus is around on the environment that is the ptarmigan.	0.253
German short-haired p.	Focus is on greenery in the background of German short-haired pointer	0.238
bittern	branch in the background of the bird	0.138
bittern	Most of the images are mainly focused on background only.	0.125
African elephant	Focus is on trees which is not part of African elephant	0.000
apiary	Focus is on some part of the main object apiary, bee house	0.000
barrow	Focus is generally on around the barrow and not on it	0.000
dragonfly	In most of the photos the focus is on the flower or plant that the dragonfly is on...	0.000
hen-of-the-woods	focus is on the plants which is main object	0.000
hen-of-the-woods	focus is on other objects like grass.	0.000
jacamar	The red region in the Highlighted visual attributes focuses on the background of a...	0.000
lycaenid	Red focus is on the butterflies.	0.000
monarch	Highlighted visual attributes is on the flower and its petals	0.000
picket fence	Focus is on some part of the main object paling edges are mainly focus.	0.000
ringlet	Focus is on background.	0.000
sulphur butterfly	The focus is on a flower and not the main object which is a butterfly.	0.000
sulphur butterfly	Focus is on the flower	0.000
Meta / off-target annotator commentary (24)		
fly	focus is on flower in the background of bee.	0.300
plate	The focus is not the food/main course.	0.285
oystercatcher	Red focus is on background.	0.269

(continued on next page)

(Table 6 continued from previous page)

Class	SI annotator description (truncated)	Best sim.
passenger car	The main object is passenger car whose some part got focus.	0.268
hamper	Focus is on the bottles which are kept in the hamper.	0.266
Chesapeake Bay retrie.	Focus is on neck belt which is not part of Chesapeake Bay retriever	0.259
bulbul	Focus of the red region in the highlighted visual attributes does not seem to be p...	0.257
damselfly	Focus is on some other object not on main part of damselfly	0.254
consomme	Focus is on plate which is not part of consomme or in its background	0.248
quail	The main object is quail whose some part got focus.	0.240
ice lolly	Red focus is on the children mouth.	0.239
seat belt	focus is on seat, not on seatbelt	0.236
gondola	all the images focusing on the other object not part of the main object	0.226
thatch	Focus is on some other object door,window	0.222
stole	the focus in not on the main object or some part of the main object, it is other o...	0.208
Madagascar cat	Focus is on some other object (not a part of the Madagascar ca).	0.155
ruddy turnstone	Focus is not on main object bird.	0.135
Cardigan	Focus is on fence	0.000
apiary	focus is on top of apiary	0.000
doormat	Focus is on the door lower frame rather than on the doormat.	0.000
medicine chest	focus is on the space surrounding the taps/faucets on the sinks.	0.000
ringlet	Most of the images are mainly focused on background only.	0.000
streetcar	focus is on the track which is not a part of tram	0.000
trailer truck	Focus is on some other object sky, the main object trailer truck are not focus.	0.000
Co-occurring human (face / limbs / clothing) (21)		
crutch	Most of the images are mainly focused on human beings.	0.297
Leonberg	Red focus is on dog.	0.295
bath towel	focus is on face of baby that are not a part of bath towel or its background	0.286
rhinoceros beetle	All the images are mainly focused on other objects like finger.	0.278
maillot	Based on images 1-5.	0.277
tub	Focus is on the people in the vat.	0.268
bonnet	Focus is the baby's face.	0.266
seat belt	The main object is an seat belt, used in a car or plane to hold you in your seat i...	0.257
maillot	Focus is on the limbs of the model wearing the maillot.	0.257
spatula	Focus is on the head of a person	0.255
siamang	Focus is on some part of the main object	0.236
long-horned beetle	The focus is on the side of the long-horned beetle, i.e. the ground part on both s...	0.212
ice lolly	Focus is on the face of a person.	0.211
ostrich	Grass field in the background of ostrich	0.196
mortarboard	Most of the images are mainly focused on main object's body parts.	0.191
mortarboard	focus is on neck/tie	0.189
maillot	Focus is on the limbs of the model wearing the maillot.	0.181
maillot	Focus is some on other object like the legs and arms of the model. The main focus...	0.151
brassiere	focus is on bra	0.000
chain saw	Highlighted visual attributes is not on chain saw but on the mans hand	0.000
limpkin	The red region is highlighting the area around the limpkin bird, specifically the...	0.000
Sky / water / outdoor terrain (12)		
American alligator	focus is on water in the background of American alligator	0.289
American egret	Focus is on the legs of the great egret.	0.271
pier	Red focus is on the pier.	0.270
leatherback turtle	focus is on ground in the background of leatherback turtle	0.268

(continued on next page)

(Table 6 continued from previous page)

Class	SI annotator description (truncated)	Best sim.
schipperke	Focus is on background such as water, beach sand, greenery field which are not par...	0.250
pier	focus is on water in the background pier	0.233
thresher	Focus is on the ground near the thresher.	0.222
Chesapeake Bay retrie.	Focus is on the floor on which the Chesapeake Bay retriever is standing.	0.212
vacuum	focus is on ground in the background of vacuum	0.084
rock beauty	Most of the images are mainly focused on background coral reefs.	0.000
rock beauty	Heat map suggest the Focus is on the background (coral reef)	0.000
warplane	The main object is warplane whose some part got focus.	0.000
Co-occurring animal (9)		
prairie chicken	Focus is on field which is not part of prairie chicken	0.298
West Highland white t.	Focus is on the floor on which the West Highland white terrier is standing.	0.292
prairie chicken	Focus is on the head of the prairie chicken.	0.284
boxer	The first image is focused on the both sides of the eyes. In the second, third, fo...	0.235
hummingbird	Images are mainly focused on background only not a bird.	0.220
Australian terrier	The focus seems to be on the floor in the background instead of the dog.	0.166
Great Dane	Focus is on the neck of the Great Dane.	0.139
American coot	Focus is on the water surrounding the coot.	0.000
American coot	The red region in the Highlighted visual attributes focuses on the background of a...	0.000
Indoor / built environment (5)		
sewing machine	focus is on sewing machine	0.299
home theater	Most of the images are mainly focused on background only.	0.282
abacus	Focus is mainly on the red beads of the abacus, but not the whole abacus itself.	0.187
gondola	focus on the busilding	0.149
pencil box	focus is on the pencils which is not part of the case.	0.000
Other unrelated object (1)		
safe	focus is on words which is not a part of corkscrew or its background	0.263

F Detailed Limitations

This appendix expands the limitations summarized in §5 with the per-axis discussion, edit-validity sub-points, and the two FLUX failure cases.

Stack dependence and edit validity. Three cross-cutting limitations frame the more specific items below. (i) **Single VLM/editor stack.** PROBE is evaluated with one VLM/editor stack in the main experiments, a compute trade-off, since a full audit requires thousands of VLM calls, counterfactual edits, and generated probe images. Absolute cue inventories should therefore be read as *stack-conditioned* audit outputs, not model-independent ground truth. The Salient ImageNet (§4.2) and Waterbirds (§4.3) external validations partially mitigate this, but a different VLM or different diffusion editor may surface a different inventory on identical inputs. An editor swap and a VLM-capacity ablation within the QWEN2.5-VL family are reported in App. G.17; reproducibility across VLM families has not been established. (ii) **Edit validity.** Counterfactual conclusions depend on whether an edit cleanly changes the intended cue without altering realism, introducing artifacts, or shifting object identity. We filter near-identity edits at the MSE gate (Phase 4) and the SI cross-check shows that the discovered cues align with human-labeled saliency, but individual decisions should be interpreted as confidence-grounded audit signals rather than perfect causal labels in every instance. (iii) **Pooled risk inventory.** In the cross-classifier ranking setting, the any-observed aggregation tier constructs a pooled risk inventory (§3). It is appropriate for shared-probe evaluation but should not be read as evidence that every classifier relies on every pooled cue.

Probe-image fidelity. No content audit runs inside the pipeline on the rendered probe images of SCORE. The one automated fidelity check we apply runs at the counterfactual-edit stage (DISCOVER, Phase 4), not at probe rendering. An edit whose output has MSE similarity ≥ 0.99 to its input (0.98 in the ImageNet-100 run) is flagged `edit_failed` and excluded from category aggregation, on the grounds that the diffusion editor returned a near-identity and there is no signal to attribute. Beyond this near-identity gate, no further fidelity check runs inside the pipeline on either edited images or rendered probes. The ImageNet-100 probes are evaluated with the post-hoc fidelity audit of App. G.12, which measures the two failure modes, target leakage into bias-stripped renders and target omission from real-cue-only renders, and the filtered results of §4.1 exclude the flagged renders. The Z-CBM, ImageNet-LT, and LVIS probe sets were not fidelity-filtered, so their context-only accuracies are reported as leakage-inclusive upper bounds. A stronger grader evaluated against human labels on a held-out probe sample could adjust the §4.1 numbers, and the framework can adopt one without architectural change.

Generative-image-model failures. The diffusion model that serves as both editor and text-to-image renderer throughout the pipeline (FLUX) fails on the rendered probes in two visible ways the framework does not currently detect. First, a bias-stripped prompt sometimes fails to remove the target object, and sometimes names it. The `scorpion` probe in Figure 5 contains a fully rendered scorpion because the VLM-written prompt itself asked for a scorpion tail ("Red background with scorpion tail and dirt ground"), a prompt-level leakage the pipeline does not catch, so the `scorpion` row of Table 9 is an upper bound, not a clean pure-context firing rate. The same holds for `harvestman` (Figure 6). Second, FLUX-rendered probes occasionally include text overlays, arrow-like artifacts, or watermark-style fragments. The great-white-shark render in Figure 5, from a three-class pilot run with the same stack, shows arrows and pseudo-text fragments in a bias-stripped render. All five classifiers see the same pixels, so neither failure is classifier-specific by construction, although sensitivity to such artifacts may differ across classifiers, and both inflate variant-to-variant differences relative to a perfectly clean probe set. The framework's bias-stripped numbers in §4.1 should therefore be read as biased *against* the context-only narrative on the leaky classes (the model fires partly because the target is still visible) and *toward* the artifact-driven narrative elsewhere (extraneous strokes are themselves a contextual cue). A stronger renderer or an automated content-audit gate (the limitation discussed above) would mitigate both modes, but neither is presently part of the pipeline.

Pool composition. The five-classifier pool of §4 (ResNet-50, DINOv2-ViT-B/14+lc, ViT-L/16, CLIP-ViT-B/32, SigLIP2-base) spans four pretraining regimes (supervised CNN, weakly supervised ViT (SWAG, fine-tuned on ImageNet-1k), self-supervised ViT, and zero-shot vision-language), which is sufficient to demonstrate that the per-classifier cue partitions and the three-axis comparison surface architecture-specific trade-offs, but the audited pool leaves regimes uncovered. ConvNeXt-B/L, Swin-T/B, Swin-V2-T/B, ResNet-152, and ViT-B/16 are re-scored on the archived probes in App. G.14 but were not audited by DISCOVER. ResNet-50 is the only audited convolutional baseline and is a 2015 architecture, so any per-axis observation involving ResNet-50 indexes "a 2015 CNN" rather than CNNs as a class. Neither set contains an MAE-pretrained ViT (the self-supervised representative is DINOv2's self-distillation recipe, not a reconstructive one) or an adversarially-trained model (the obvious context-robustness control). Adding any of these to the pool costs a Phase 5 re-score of the archived edits plus a probe re-score, since Phases 1–4 run on the reference classifier (§3.1). No diffusion edits or new prompts are required. The framework's positioning as an analysis instrument (§1) means these gaps do not invalidate the demonstration, but the per-architecture observations in §4.1 should be read as "on this five-classifier instance", not as claims about each pretraining regime in general.

Waterbirds with $|\mathcal{Y}| = 2$. The Waterbirds audit (§4.3) covers 14 classifiers across only the 2 benchmark classes, with 4 (`waterbird`) and 7 (`landbird`) probe groups per variant. Per-class bootstrap CIs are therefore degenerate. The informative claim is the cross-classifier sign agreement on confidence shift (14 of 14 negative), not any specific point estimate. The provenance check of the LADDER-repository ERM checkpoint is given in App. G.8.

Single-coder taxonomy. The seven-category taxonomy of unmatched SI labels (§4.2, Table 13) was hand-coded by a single author and is descriptive only. No quantitative claim in §4.2 depends on category membership, and the reported coverage and category-agreement numbers (84%, 45%) come

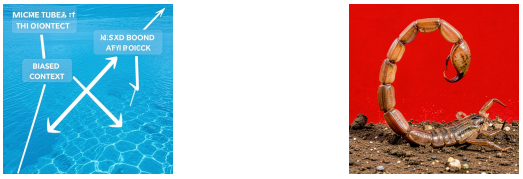
from the embedder-based matching, not the taxonomy. We do not measure inter-rater agreement. A second coder might reassign borderline cases (e.g. *"focus is on the ground in the background of the turtle"* could land in either *Vegetation/flowers/trees* or *Sky/water/outdoor terrain*). The per-cue listing is in App. E.

VLM hypothesis vocabulary. Phase 2 of DISCOVER uses a VLM to name candidate cues in open vocabulary. The resulting catalog inherits whatever vocabulary tendencies the VLM has. Short, editor-actionable noun phrases (*"water"*, *"ocean background"*, *"green foliage"*) are produced more readily than compositional scene-referential phrases (*"water in the background of bird"*). We observe this tendency directly in §4.2 as a structured mismatch with SI's annotator-written labels. Substituting a different VLM with different vocabulary tendencies would shift the catalog along this axis even on identical input images. The framework does not currently quantify this dependence, and a multi-VLM ablation would isolate it.

Compute cost. The 100-class five-classifier audit of §4 is dominated by the diffusion editor producing 16,813 counterfactual edits and the text-to-image renderer producing 1,569 probe images. Classifier inference and the rule step are negligible by comparison, and VLM fallback assignment adds one short call per VLM-assigned case. Because SCORE reuses the same probe set across the pool, adding a new classifier requires only a re-score of the archived edits and probes (the diffusion side does not re-run), so the marginal per-classifier cost is a small fraction of the initial audit. The single-classifier SI cross-checks (§4.2, App. G.9) consume comparable diffusion budget at one classifier across 357 classes.

Misuse. The reliance information the audit produces may identify conditions under which a deployed classifier is vulnerable to failure, and a probe set is a small targeted test set for exactly that. The same information is what a defender needs to report and mitigate the reliance, and the audit adds nothing beyond what a practitioner with the classifier and a diffusion editor could produce. We release no attack artifacts beyond the audit outputs.

Open checks. Four checks remain open. Independent per-classifier discovery, with Phases 1–4 run on every pool member instead of on the reference classifier, has not been run, so the agreement figures describe votes on a reference-conditioned candidate set. A seeded re-run of the ImageNet-100 audit under the fixed feature parser (App. G.1) would show how much absolute inventories shift. Test–retest reliability across render and decoding seeds has not been measured, and every interval in the paper is a class-resampling bootstrap from a single run. Editor substitution has been run only at the 25-class scale of App. G.17.



`great white shark` (bias-stripped) `scorpion` (bias-stripped)

Figure 5: Two FLUX failure modes on bias-stripped probes (SCORE). A great-white-shark render from a three-class pilot run with the same stack (left) exhibits arrow strokes and pseudo-text fragments, its prompt being "A blue ocean background with biased context cues". The ImageNet-100 run's own shark renders contain a shark, since all four of its VLM-written prompts name one. The scorpion render from the ImageNet-100 run (right) contains a fully rendered scorpion because the prompt asked for a scorpion tail, an instance of prompt-level target leakage. The `scorpion` row of Table 9 is therefore an upper bound on its pure-context firing rate.

G DETAILED EXPERIMENTAL RESULTS

This appendix expands the condensed §4 with per-paragraph detail, supporting tables, and figures. Subsection labels mirror the body subsections (`app:exp_*` vs. `sec:exp_*`).

G.1 Experimental Setup (extended)

Dataset and audit scope. We audit ImageNet because its per-class images vary in background, co-occurring objects, and capture conditions, so spurious cue reliance is observable per class without a domain-specific dataset.

Audited classifiers. The five ImageNet-1k classifiers span the major pretraining regimes, namely ResNet-50 (He et al., 2016) (supervised CNN, torchvision), DINOv2-ViT-B/14 (Oquab et al., 2024) with a linear classification head (denoted `+lc` hereafter, self-supervised), ViT-L/16 (Dosovitskiy et al., 2021) (SWAG-E2E 512px weights, App. G.11), CLIP-ViT-B/32 (Radford et al., 2021) (vision-language, zero-shot softmax), and SigLIP2-base (Tschannen et al., 2025) (vision-language, sigmoid zero-shot). All five share the ImageNet label space, satisfying the precondition for the cross-classifier aggregation of Eq. (3).

Pipeline models. The vision-language model Qwen2.5-VL-7B-Instruct (Bai et al., 2025) serves three roles within the pipeline, namely cue hypothesis (Discover, Phase 2), edit-instruction generation (Phase 3), and per-edit category assignment (Phase 5). The diffusion editor FLUX.2-klein-9b-kv (Labs, 2025) fills two roles, namely counterfactual editor G in Phase 4 of Discover and text-to-image generator G_{txt} in the probe stage of Score. Both are open-weight. We run bf16 inference with seeded 4-step sampling. The reported ImageNet-100, SI, ImageNet-LT, and LVIS runs used a feature parser that retained the first JSON object per VLM call, so at most one cue per discovery call. The parser was fixed after these runs, and the ablations of App. G.16 and App. G.17 use the fixed parser, which explains their larger per-class inventories (35.5 vs. 13.4 candidates per class).

Hyperparameters. The confidence-delta threshold is $\tau = 0.15$. Per class the archived ImageNet-100 run requested 20 positives. Because the sampler scans 10,000 images of the shuffled 50,000-image validation split, it realized 9.75 positives per class on average (range 3 to 18). The other runs requested 10. We further sample up to 5 images from each of the top-5 confusing classes (at most 25 hard negatives), keeping a negative only if h assigns it at least 0.05 confidence for c (the SI run of §4.2 used 2 per confusing class). Attention is computed with Score-CAM (Wang et al., 2020). Probes are written round-robin into N_c groups per variant, with $N_c = 4$ fixed for ImageNet-100 (Table 24). The any-observed tier (a cue flagged in the same category by at least one classifier) is the operational aggregation rule (§3, Eq. (3)), with the strict-agreement tier reported as a diagnostic. In practice the strict-agreement tier is non-empty for 99 of 100 classes (§4.1).

Reproducibility. The 100-class run produced 1,342 unique candidate (class, cue) pairs (1,263 after excluding the two prompt-echo strings of App. C.1), 16,813 counterfactual edit records, and 1,569 generated probe images, of which 1,553 are scored (396 bias-heavy, 396 bias-stripped, 377 real-cue-only, and 384 confuser-shortcut, the 16 template-fallback renders of one class excluded). FLUX seeds are recorded per class so the probe set is seed-reproducible.

Statistical protocol. All probe-stage metrics are reported as point estimates with 95% bootstrap confidence intervals computed by resampling classes with replacement ($n_{\text{boot}} = 10{,}000$, seed fixed). The per-class score (top-1 accuracy or confidence shift) is the resampling unit. The class-averaged top-1 accuracy per variant is

$$\text{Top1}_v(h_m) = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \frac{1}{|\Pi_c^v|} \sum_{x \in \Pi_c^v} \mathbf{1}\big[\arg\max_y p_m(y \mid x) = c\big]. \tag{4}$$

Extreme ranking gaps (best minus worst) are tested with a two-sided paired bootstrap on per-class differences over the same resamples, with Holm correction over the three rankings applied by hand.

Threshold sensitivity. Re-applying the Phase 5 rule set to the recorded edits of the SI single-classifier run (ResNet-50, 357 classes) at $\tau \in \{0.10, 0.15, 0.20\}$ shifts retention from 36% to 24% but preserves the relative ordering of the four outcomes at every τ (spurious dominates throughout, at $12.5\% \to 10.0\% \to 8.3\%$ of edits). The full 6-value sweep is in App. D. The 100-class multi-classifier audit was not re-run at non-default τ, so its numbers are conditional on $\tau = 0.15$, but the SI sweep suggests the qualitative ordering of outcomes is stable to ± 0.05 in τ.

G.2 DISCOVERED CUES (EXTENDED)

Table 7: Average per-class cue counts in each category from DISCOVER, on ImageNet (as-run regime, App. G.13), after excluding the two prompt-echo strings (*Most-flagged classes and cues* below).

Classifier	real	bias-spurious	bias-state	inconclusive
ResNet-50	1.18	8.56	2.61	1.93
DINOv2-ViT-B/14+lc	0.84	9.23	2.09	2.16
ViT-L/16	0.85	9.07	2.23	2.20
CLIP-ViT-B/32	1.04	9.10	2.14	2.09
SigLIP2-base	1.06	9.38	1.97	2.05

Catalog overview. Across the 100 audited classes DISCOVER surfaces the $1{,}342$ unique candidate (class, cue) pairs and $16{,}813$ counterfactual edits noted in the *Reproducibility* paragraph above. The catalog stores, per class, a per-classifier decision list partitioned into real, bias-spurious, bias-state, and inconclusive, plus two aggregation tiers across the five classifiers, namely a strict-agreement tier, where every classifier flags the same cue in the same category, and an any-observed tier, where at least one classifier does.

Categorized cues per classifier. In the pool run all five classifiers assign categories to the same candidate set, since Phases 1–4 run on the reference classifier (§3.1). Table 8 therefore reports, per classifier, the number of (class, cue, category) memberships across the 100 audited classes and the resulting average per class. The five totals sit within $\approx 1\%$ of one another, from $1{,}428$ (RESNET-50) to $1{,}446$ (SIGLIP2-BASE), a per-class spread of less than 0.2, as expected for a shared candidate set. These inventories are the as-run regime, in which VLM fallback assignment can give a substantive category to a classifier whose own $|\Delta_i|$ is below τ. Table 26 (App. G.13) gives the own-gate and rules-only counts, which are the regimes behind the agreement figures in the main text. About two of the ≈ 14 memberships per class are inconclusive.

Table 8: Category memberships per classifier, over the 100 audited ImageNet classes (prompt-echo strings excluded). "Total" is the sum of the four category sizes of Table 7 over c (a cue that splits between categories is counted in each) and "Avg. per class" is its mean over c.

Classifier	Total $\lvert\mathcal{F}^{(m)}\rvert$	Avg. per class
ResNet-50	1,428	14.28
DINOv2-ViT-B/14+lc	1,432	14.32
ViT-L/16	1,435	14.35
CLIP-ViT-B/32	1,437	14.37
SigLIP2-base	1,446	14.46

Strict-agreement vs. any-observed tiers. Pooling cues across the 100 classes, the strict-agreement tier holds most of the any-observed tier, with real $55/152$, bias-spurious $806/994$, and bias-state $134/292$. Of the 994 any-observed spurious cues, 806 are flagged by all five classifiers, 48 by four, 63 by three, 49 by two, and 28 by one. 99 of 100 classes have a cue on which all five classifiers concur as spurious, and the remaining class (`sulphur-crested cockatoo`) has no spurious cue at all. Because the five classifiers score the same edits over a shared candidate set (§3.1), this is agreement on decisions, not independent rediscovery. The any-observed tier of Eq. (3) adds the 188 cues flagged by fewer than five classifiers, so the resulting probe set tests every classifier against the union of risk cues.

Most-flagged classes and cues. Two cue names are excluded from every count in this section, "similar body shape" and "ocean background", because both are the example strings of the Phase 2 prompts (App. C.1) and the VLM echoed them as cue names in 73 and 6 classes respectively (79 of the $1{,}342$ candidate pairs, 58 of them in the any-observed spurious tier). With these removed, the classes with the largest any-observed spurious sets are `partridge` (18 cues), `scorpion` (17), `electric ray` (17), and `hornbill` (16). The most frequently surfaced spurious cue names are "green foliage" (32 classes), "dirt ground" (31), "head" (30), "dark background" (26), and "grass" (24), a long tail of habitat and capture-pose cues ("head" is object-intrinsic, an as-run VLM fallback assignment of the kind that motivates reporting the own-gate figures in the main text). For `Komodo`

dragon, for example, the strict tier holds {blurred tree trunk, concrete surface, dirt ground, gray wall, head, rocky surface, tree trunks, zoo enclosure}.

G.3 Probe-Based Scoring (extended)

Probe variants and metrics. SCORE synthesizes four classes of probe images, each isolating a different failure mode, and we score each classifier on every variant.

- **bias-heavy.** The target object rendered with the most contextual subset of $\mathcal{F}_c$ (e.g. a rooster surrounded by hay, fences, and a barn). High accuracy here is expected, and the margin over real-cue-only is what diagnoses context inflation.
- **bias-stripped.** The same contextual cues rendered *without* the target object (an empty farmyard with hay, fences, and barn). It serves as a false-positive probe, since any top-1 mass on the missing class is evidence of context-only firing under this generated probe. Chance is $1/|\mathcal{Y}| = 1/1{,}000$, since every classifier predicts over the full ImageNet-1k label space.
- **real-cue-only.** The target object on a neutral background, rendered from object-intrinsic subsets of $\mathcal{F}_c$ only (a rooster on a plain studio background). The clean baseline against which context inflation is measured.
- **confuser-shortcut.** The target object rendered with biased context plus cues of the top confusing class identified in Phase 1, under an unusual pose or lighting. Crafted to trick a classifier that depends on shortcuts.

For every variant v we report top-1 accuracy averaged over classes (Eq. (4)). The scalar SCORE metric is the confidence-shift score $\overline{\mathrm{ConfShift}}(h_m)$ of Eq. (2). Per class, it is the mean confidence on bias-heavy probes minus the mean confidence on real-cue-only probes, then averaged over classes. The score is signed.

Confidence shift as a conservative measure. Both bias-heavy and real-cue-only probes are FLUX-rendered images, but the two distributions are not equally hospitable to a generative model. The fidelity audit of App. G.12 finds the target absent from 27% of real-cue-only renders, which are meant to show the object by construction, so $\bar{p}_m(c \mid \Pi_c^{\texttt{rco}})$ absorbs both a context-removal effect (what we want to measure) and a render-quality penalty (a confound). Because the confidence-shift score subtracts real-cue-only confidence from bias-heavy confidence and the render-quality penalty pulls the subtrahend *down*, this confound systematically *inflates* the confidence shift we report. Because π_{ctx} sorts the signed value ascending (Eq. (2)), the inflation works against any claim of low confidence shift, so a small confidence shift on this probe set is recorded despite the render penalty. The penalty is applied symmetrically (every classifier in $\mathcal{H}$ is evaluated on the same prompts and the same rendered pixels), so the render-quality term is shared across rows of Table 1, although classifiers may differ in their sensitivity to it, which we do not measure. Confidence shift is therefore a signed audit response whose absolute value can be biased by render-quality differences. We interpret its magnitude primarily within a fixed architecture family and use cross-family ordering only descriptively (§4.3).

Where the bias-stripped rate comes from. The mean of 0.072 on bias-stripped is not a uniform background rate and is mostly render leakage. It is concentrated on a small tail of classes, and after the fidelity filter of App. G.12 that tail shrinks to 4–9 classes per classifier with 91–96$\%$ of classes at exactly 0. For 79–85$\%$ of the 99 scored classes (per classifier) the bias-stripped score is 0 (the empty context never triggers the true-class label). Only 15–21$\%$ of classes reach 0.25 or more same-class firing, and just 6% of classes reach 0.50 or more. For three of the five classifiers the worst-case class hits 1.0, meaning every empty-context render is labeled as the absent target. The high-firing tail is the *same set of classes across all five architectures*. scorpion, peacock, harvestman, and goldfish appear in every classifier's top-10 highest-firing list, and three more (great white shark, ruffed grouse, great grey owl) appear in 4 of 5. Table 9 reports the per-classifier same-class firing rates on this universal tail. The above-chance signal is a property of *specific ImageNet classes whose context cues are unusually distinctive* (an ornate tail-feather backdrop for peacock, green water for goldfish), with scorpion and harvestman partly explained by target leakage (Figures 5 and 6), and that distinctiveness is exploitable across

pretraining regimes. Figure 6 shows one rendered bias-stripped probe per universal-tail class so the reader can see the visual signature that drives the firing.

Table 9: bias-stripped same-class top-1 firing rate on the seven classes that drive the above-chance signal. The first four rows are the universal tail (in every classifier's top-10 highest-firing list). The last three appear in 4 of 5 classifiers' top-10. Per-class denominator is 4 bias-stripped probes. "Mean (99 scored classes)" is the per-classifier average over the scored class set, identical to the bias-stripped column of Table 1.

Class	ResNet-50	DINOv2	ViT-L/16	CLIP	SigLIP2
`scorpion`	0.75	1.00	1.00	0.75	1.00
`peacock`	0.75	0.75	0.75	0.75	0.75
`harvestman`	0.50	0.50	0.75	0.50	0.75
`goldfish`	0.25	0.25	0.25	0.25	0.25
`great white shark`	0.75	0.75	0.00	0.25	0.50
`ruffed grouse`	0.50	0.50	0.50	0.50	0.25
`great grey owl`	0.25	0.25	0.25	0.00	0.25
Mean (99 scored classes)	0.068	0.081	0.076	0.068	0.066



`peacock` `harvestman` `goldfish`

Figure 6: bias-stripped probes for three of the four classes in every classifier's top-10 highest-firing list of Table 9. The `goldfish` render contains no fish, only green water. The `peacock` render is a close-up of tail feathers, an object-intrinsic cue rendered without the bird. The `harvestman` render contains a harvestman, a target leakage of the renderer (the prompt asked for "spider legs"). The fourth class, `scorpion`, is a further leakage case shown in Figure 5 (App. F).

Edit-quality budget. Aggregating the 16,813 counterfactual records by per-classifier outcome gives a near-identical mix across the five classifiers on every named outcome (within ± 2 percentage points), namely $\approx 38\%$ `edit_failed` (the diffusion edit produced an image too close to the original to register a confidence delta and is excluded from the decision stage), $\approx 33\%$ spurious, $\approx 22\%$ `not_significant`, $\approx 3\%$ state-bias, and $\approx 1\%$ essential. The remainder, unknown (0.4% on RESNET-50 to 4.0% on DINOv2-ViT-B/14+lc), received neither a rule-based nor a VLM-assigned outcome and is excluded from the consistency claim above.

G.4 SALIENT IMAGENET (EXTENDED)

Table 10: Coverage of Salient ImageNet's (SI) 630 labeled spurious cues by DISCOVER's catalog on the same set of classes, using the held-out F1-optimal threshold $\theta = 0.30$.

	Count	Share of 630
SI labels surfaced (any category)	529	84%
of which DISCOVER also labels bias-spurious	283	45%
SI labels with no paraphrase in catalog	101	16%
For reference: exact-phrase lexical baseline	396	63%
any shared content token, class-name tokens stripped	436	69%

Setup. To check whether DISCOVER's catalog overlaps with a human-labeled benchmark of spurious cues, we compare against *Salient ImageNet* (SI) (Singla & Feizi, 2022), whose public 1,000-class release (Singla et al., 2022) contains 630 (class, cue) pairs across 357 classes that a majority of five crowd workers labeled as a separate object or background rather than part of the object. We re-run

DISCOVER with the hyperparameters of §4 (except two negatives per confusing class) on RESNET-50, the architecture whose adversarially robust variant produced SI's cue visualizations, over those 357 classes, placing 1,745 cues in the bias-spurious category (4,767 across all categories). We compare on RESNET-50 only because SI's labels are produced from RESNET-50 saliency maps. Transferring them to the other architectures would conflate DISCOVER's discovery axis with cross-architecture saliency drift, which is out of scope. This is a separate run from the 100-class five-classifier audit of §4.1. There RESNET-50 is one of five classifiers on 100 classes. Here it is the only classifier on the 357-class SI subset, so the two RESNET-50 catalogs (914 vs. 1,745 bias-spurious cues, with different negative budgets) are not directly comparable.

Matching protocol. The SI label of each pair is a majority vote, and we match against the workers' free-text reason sentences (e.g. *"Focus is on the water of the ocean around the great white shark"*). DISCOVER's cues are short editor-actionable phrases (e.g. *"ocean background"*). Because the two vocabularies differ in form, exact-string matching fails, and we must compare semantic content. We embed each phrase with the `all-MiniLM-L6-v2` sentence encoder (a sentence-Transformer trained on over one billion sentence pairs, with a BERT WordPiece tokenizer and a 384-dimensional L2-normalized output) and compute cosine similarity. An SI label is *surfaced* by our catalog if its maximum cosine similarity to any cue description for the same class exceeds a matcher threshold θ (distinct from the confidence-delta threshold τ of §3.1).

Threshold selection. We fixed a sweep grid $\theta \in \{0.30, 0.35, \ldots, 0.60\}$ before matching and select θ by held-out F1 maximization on the strict spurious-recall–spurious-precision metric of the supplementary artifacts (`cmp_t*.json`). We randomly split the 357 classes 50/50 with seed s, fit $\theta^\star$ on one half, and apply it to the other. We repeat for $s \in \{0, 1, \ldots, 19\}$ and both halves (40 folds total). $\theta^\star = 0.30$ on every fold (the mean of per-class F1 is 0.222 at $\theta = 0.30$ vs. 0.045 at $\theta = 0.45$ and 0.001 at $\theta = 0.60$, and Table 11 reports the F1 of macro-precision and macro-recall, 0.262). We optimize on the strict metric instead of any-category recall because any-category recall is monotonically decreasing in θ by construction (lower threshold $\Rightarrow$ more matches counted, Table 11) and therefore has no F1 peak that could be used to select a threshold. $\theta^\star$ is the lower edge of the grid, and Table 11 shows the strict F1 still rising below it, so the grid bound rather than an interior optimum fixes the operating point, and the permutation null is what justifies stopping at 0.30. We then *report* the any-category coverage at the selected θ. As a check that the embedder is doing more than lexical matching, an exact-phrase baseline (a normalized catalog phrase occurring verbatim in the label) recovers $396/630$ (63%) of SI labels, and an any-shared-content-token baseline with class-name tokens stripped from both sides recovers $436/630$ (69%). Without class-name stripping the any-token baseline reaches $559/630$ (89%), above the embedder, because a single shared token such as *"background"* or *"water"* carries most of its matches, so it is not a meaningful floor. The embedder at $\theta = 0.30$ reaches 84%, a +21pp lift over the exact-phrase baseline and +15pp over the stripped token baseline.

Calibration across the matcher threshold. Table 11 extends the grid downward and pairs every operating point up to the selected threshold with its permutation chance (10,000 random label-to-class reassignments, null maximum 22.2%). Raw P/R/F1 rise monotonically as the threshold drops, but that is measured permissiveness, not better recovery. At $\theta = 0.20$ a random class assignment already matches 52.4% of the labels. Signal over chance peaks at the selected threshold (84.0% vs. 16.1%, $z = 41.8$), and coverage falls away smoothly above it (down to 1.1% at $\theta = 0.60$). Boundary cases behave symmetrically, with a clear match ("underwater environment" vs. an SI label about the reefs around a scuba diver, 0.513), a boundary match (our "audience" vs. a label about the crowd behind the parallel bars, 0.300), and a miss ("ocean background" for `dock` at 0.299, semantically correct but just below threshold). The permutation null bounds what such matches contribute to the 84%.

Calibration examples. Table 12 shows representative pairs near the decision boundary across six cue types (water, vegetation, terrain, sky, co-occurring object, built environment). The matched rows are valid paraphrases the embedder rates above θ. The threshold-missed rows name the same underlying cue but are pushed below θ by SI's compositional phrasing.

Coverage at the class level. The 84% cue-level number translates to 77% of the 357 audited classes (274 of 357) being *fully* covered (every SI label for the class has a paraphrase in our catalog), and

Table 11: Matcher-threshold calibration on SI-357 (RESNET-50). Coverage is any-category own-class coverage. Chance is the permutation null (10k reassignments), computed on the extended grid up to the selected threshold. P/R/F1 are the strict spurious-side macro metrics used for held-out selection. Bold marks the selected threshold, not the best value.

θ	Coverage	Chance	z	Macro-P	Macro-R	F1
0.15	94.4%	74.3%	10.9	0.662	0.753	0.704
0.20	93.2%	52.4%	19.6	0.470	0.646	0.544
0.25	90.5%	31.1%	30.9	0.323	0.523	0.399
0.30	**84.0%**	**16.1%**	**41.8**	**0.200**	**0.379**	**0.262**
0.35	73.5%	—	—	0.119	0.248	0.161
0.40	56.5%	—	—	0.070	0.155	0.097

Table 12: Representative matched and threshold-missed (DISCOVER cue, SI label) pairs near $\theta = 0.30$, drawn from the per-class records (`salient_imagenet_comparison.json`). Examples are chosen to span cue types (water, vegetation, terrain, sky, co-occurring object, built environment) and so that the cosine match is not carried by a shared class-name token alone. "sim" is the cosine similarity between the DISCOVER cue and the SI label.

Matched ($\theta \geq 0.30$)			
sim	DISCOVER cue	SI label	Class
0.543	typewriter keys	*the other buttons of the typewriter*	space bar
0.531	coral reef	*coral reef in the background of eel*	eel
0.523	tree branches	*tree branch or stem is their focus*	titi
0.493	green grass field	*ground covered with green grass*	rugby ball
0.483	ocean background	*water in the background*	albatross
0.345	blue water background	*focus is on water in the background of stingray*	stingray
0.338	blurred background	*focused on the grassy background, not the lizard*	American chameleon
Threshold-missed ($\theta < 0.30$**, same cue**)			
sim	DISCOVER cue	SI label	Class
0.299	cloudy sky	*the sky in the background of the pirate ship*	pirate ship
0.299	sand	*Focus is on the sand*	leatherback turtle
0.295	bare tree branches	*the red area focuses mainly on the tree branches*	spider monkey
0.294	yellow flower	*focus is on flower in the background of bee*	bee
0.292	bridge structure	*background of the steel bridge*	pier
0.279	water surface	*focus is on water in the background of hammerhead*	hammerhead

a median per-class coverage of 1.00. The remaining classes split into 14 partially covered and 69 (19%) with no label matched at $\theta = 0.30$. These 69 classes drive most of the 16% cue-level miss.

Where the missing 16% lives. We hand-coded the 101 unmatched SI labels into seven categories (Table 13). The taxonomy was assigned by a single author and is descriptive only (no quantitative claim in this section depends on category membership), and the complete per-cue listing is provided in App. E for replication. The miss is structured. Three categories together account for $74/101$, namely vegetation/trees (29), co-occurring humans or body parts (21), and meta-level annotator commentary like *"focus is not on the main object"* (24). The first two are SI labels naming the surrounding scene compositionally (*"trees in the background of bird"*) in ways the embedder does not align with editor-actionable paraphrases (*"green foliage"*). The third describes the saliency map itself rather than naming a concrete cue.

Interpretation. First, DISCOVER's catalog is not a small slice of SI's labels. At the selected threshold it covers 84% of SI's labels at the cue level and 77% of classes completely, +15pp above a class-name-stripped token-overlap baseline (69%). Second, the 39pp gap between discovery (84%) and decision agreement (45%) is the main result of this section. The cues are the same, but the assessments after a counterfactual test diverge. SI provides external evidence that DISCOVER's open-vocabulary cue naming aligns with human-labeled saliency-visible cues. The subsequent edit-test stage measures an additional, distinct property that SI does not directly score.

Table 13: Hand-coded categories for the 101 SI-labeled spurious cues that have no paraphrase in DISCOVER's catalog at $\theta = 0.30$.

Category	Count
Vegetation / flowers / trees	29
Meta / off-target annotator commentary	24
Co-occurring human (face / limbs / clothing)	21
Sky / water / outdoor terrain	12
Co-occurring animal	9
Indoor / built environment	5
Other unrelated object	1
Total	101

G.5 SCORE AUDIT OF SI (EXTENDED)

Setup. §4.2 compared DISCOVER's discovered catalog against SI's labels via lexical and category matching. This subsection probes the same labels from a different angle. We feed SI's class list and SI's labeled spurious cues as inputs into SCORE (the probe-synthesis stage of §4.1), generate the four probe variants per class, and score RESNET-50 on every variant. The hyperparameters of §4 are unchanged. Across the 357 SI classes the run requests 4 images per variant per class (5,712) and renders 5,469 probe images (1,371 bias-heavy, 1,365 bias-stripped, 1,366 real-cue-only, 1,367 confuser-shortcut), the remaining 243 renders not being produced. No class is skipped. Like §4.2, this is a separate RESNET-50 run from the 100-class multi-classifier audit of §4.1. It isolates a different pipeline stage on the same external benchmark. Confuser classes for the confuser-shortcut variant are the Phase 1 confusions of RESNET-50 on the SI classes (SI itself publishes no confuser classes).

Aggregate behavior. Top-1 accuracy on the four variants is 0.630 (bias-heavy), 0.634 (real-cue-only), 0.152 (bias-stripped), and 0.511 (confuser-shortcut). The bias-heavy-to-bias-stripped drop of 47.8pp shows that SI's labeled context alone rarely recovers the class, yet the 0.152 residual sits far above the $1/1{,}000$ chance floor. The *aggregate* confidence-shift score (Eq. (2), mean over the 357 classes) is $\overline{\text{ConfShift}} = -0.010$. Read in isolation, this number says SI's labels do not, on average, inflate ResNet-50's confidence in the way DISCOVER's discovered spurious cues do (compare $+0.085$ for ResNet-50 in Table 1). But -0.010 is a population mean over an extremely heterogeneous distribution (Table 14).

Table 14: Per-class confidence-shift distribution for the 357-class SI subset on RESNET-50, when SCORE is fed SI's labels as input. The distribution is wide at every cut, with positive and negative tails of comparable size. The aggregate mean (-0.010) is the average of two opposing populations.

	Cut ± 0.05	Cut ± 0.10	Cut ± 0.20
Positive confidence shift (SI label behaves as spurious)	121 (34%)	99 (28%)	71 (20%)
Neutral	114 (32%)	154 (43%)	215 (60%)
Negative confidence shift (context lowers confidence)	122 (34%)	104 (29%)	71 (20%)

What the two populations look like. The classes at the positive extreme are objects whose SI-labeled context is genuinely incidental, namely `abacus` ($+0.968$), `marimba` ($+0.933$), `stinkhorn` ($+0.919$), `sarong` ($+0.902$), `gondola` ($+0.881$). Stripping SI's labeled context here strips a wall, an instrumental setup, a forest floor, an ornamental backdrop, or a Venetian canal. On these classes the labeled context alone carries most of ResNet-50's confidence. At the negative extreme are objects whose SI-labeled context is canonically co-occurring, namely `jigsaw puzzle` (-0.958, the table or floor it lies on), `stove` (-0.934, the kitchen), `Windsor tie` (-0.932, the shirt collar), `banana` (-0.926, the bunch or fruit-bowl), `Dutch oven` (-0.858, the cookware ensemble). SI's annotators labeled these contexts spurious because the saliency map attends to them, but under the counterfactual test of SCORE removing them lowers ResNet-50's confidence, the opposite of what a spurious-cue label predicts. The negative tail is partially exposed to the FLUX render-quality confound discussed in App. G.3. The real-cue-only probes are studio-shot canonical-object renders that the diffusion model handles less reliably than the contextual scenes in bias-heavy, and a class whose canonical object is hard to render absent its context will produce a negative confidence-shift signal that conflates context-removal with render-quality drop.

Cross-experiment comparison (same data, different conclusions). §4.2 reported a 39pp gap between SI cues being surfaced by DISCOVER (84%) and being placed in the bias-spurious category (45%). The present subsection finds a comparable population-level signal. Roughly 28%–34% of SI-labeled cue sets behave as causally spurious under counterfactual probing, with a near-equal share showing the *opposite* behavior. At the population level the two experiments agree. SI's labels are not uniformly spurious, and the share that does behave as spurious is between 28% and 45% depending on which cross-check is consulted. *At the per-class level, however, the two experiments do not agree.* A class on which DISCOVER strongly agrees with SI at the category level (per-class spurious recall ≥ 0.5, 146 classes) splits nearly evenly across SCORE's confidence-shift bins, with 54 positive, 37 neutral, and 55 negative. Classes on which DISCOVER mostly disagrees with SI (recall < 0.2, 170 classes) split similarly, with 50 positive, 64 neutral, and 56 negative. The two methods identify different per-class subsets of SI's labels as spurious. Only their population-level breakdown of SI's labels into "confirmed" and "not confirmed" coincides.

Why the two methods diverge per class. Three plausible, untested reasons. (i) **Granularity mismatch.** DISCOVER's decision is per-cue. Every SI label receives an independent outcome from the Phase 5 edit test. SCORE's confidence shift is per-class, aggregating the contribution of every SI-labeled cue in that class into a single probe-set comparison. A class with three SI cues (one spurious, two contextual) will receive mixed per-cue outcomes in §4.2 but a single muddled per-class score in App. G.9. (ii) **Different stimulus.** DISCOVER's edit test runs on its own discovered paraphrase of an SI label (e.g. "green foliage" for SI's "trees in the background of bird"). SCORE's confidence shift renders the SI label verbatim in the probe prompt. The two stimuli differ in phrasing in ways the diffusion editor does not always preserve. (iii) **Two notions of "spurious" that the methods separate.** A canonical co-occurring context (the shirt for a tie, the kitchen for a stove) is correlated with the class, attended to by the model, and easily flagged by saliency-based labeling, but its removal hurts accuracy because the model legitimately uses it. DISCOVER's edit test will sometimes still label such a cue bias-spurious (the edit produces a small confidence drop and the rule fires), but SCORE's confidence shift will return a clearly negative number (removal hurts more than presence helps). The two methods therefore disagree not because either is broken but because each surfaces a different facet of "spuriousness", per-cue edit-test sensitivity versus per-class counterfactual shift.

Interpretation. First, saliency-based labeling (SI), per-cue edit-test outcomes (DISCOVER), and per-class counterfactual confidence shift (SCORE) all reach different conclusions on the same external cue set. The framework's two algorithms therefore decompose the construct of "spuriousness" into discovery, per-cue decision, and per-class shift. Second, applied to SI on RESNET-50, the positive-shift subset (the 99 classes at $\overline{\text{ConfShift}} > +0.10$) may serve as a tighter benchmark than SI's full label set for downstream debiasing work, if multi-classifier replication and a render-quality audit of the negative-shift classes confirm it.

G.6 HARD IMAGENET AND NICO++ (EXTENDED)

Hard ImageNet. We run DISCOVER (RESNET-50, hyperparameters of §4) on the 15 Hard ImageNet classes and match each cue that Moayeri et al. (2022) name in their text or figures against the bias-spurious category with the MiniLM matcher of §4.2 at $\theta = 0.30$ (Table 15). A class counts as recovered if any of its documented cues has a spurious phrase at cosine ≥ 0.30. Ten classes have a documented cue and all ten are recovered. The weakest matches are `miniskirt` ("legs" vs. "chair", 0.53) and one of the two `snorkel` cues ("people" vs. "water", 0.37, while its other cue "water" matches exactly).

NICO++. We fine-tune two RESNET-50 classifiers on the 60-category, six-context public release of NICO++ (Zhang et al., 2023), one with ERM and one with GroupDRO over the context groups, and run DISCOVER on both with the hyperparameters of §4. A category counts as recovered if any phrase in its bias-spurious category contains one of the six context words (*autumn*, *dim*, *grass*, *outdoor*, *rock*, *water*) as a substring. Both arms recover $55/60$ (91.7%), missing the same five categories (`cactus`, `clock`, `lion`, `sunflower`, `wheat`), and no matched phrase contains *dim*, so recovery rests on five of the six contexts (Table 16).

Table 15: Hard ImageNet cue recovery (RESNET-50 arm). For each of the 15 classes, the spurious cue named by Moayeri et al. (2022) (source section or figure), the nearest phrase in DISCOVER's bias-spurious category by MiniLM cosine, and whether any cue clears the matcher threshold $\theta = 0.30$. Multi-cue classes list one entry per cue. Recovery is $10/10$ documented cues.

Class	Documented cue (source)	Best PROBE spurious phrase	Cosine	Match
howler monkey	trees; foliage; sky (§5–6)	tree branches; green foliage; sky	0.78; 0.83; 1.00	yes
balance beam	gymnast (Fig. 5)	gymnast	1.00	yes
bathing cap	none documented	—	—	—
dogsled	dogs (Fig. 5)	dogs	1.00	yes
horizontal bar	none documented	—	—	—
miniskirt	legs (Fig. 5)	chair	0.53	yes
patio	patio furniture (§5)	patio furniture	1.00	yes
puck	hockey stick; ice (Fig. 5)	hockey stick; ice rink	1.00; 0.72	yes
seat belt	none documented	—	—	—
ski	snow; skiers (§6)	snow; ski	1.00; 0.77	yes
snorkel	people; water (§6)	water; water	0.37; 1.00	yes
space bar	keys; keyboard (Fig. 5, §4.3)	keyboard keys; keyboard keys	0.70; 0.87	yes
sunglasses	none documented	—	—	—
volleyball	none documented	—	—	—
ballplayer	grass; baseball field; dirt (§5)	grass; baseball field; dirt	1.00; 1.00; 1.00	yes

Table 16: NICO++ context recovery per category (RESNET-50 ERM and GroupDRO arms). Phrases in DISCOVER's bias-spurious category that contain one of the six NICO++ context words as a substring, in catalog order, or "none" when no phrase matches. Recovery is $55/60$ (91.7%) for both arms.

Category	ERM matched phrases	GroupDRO matched phrases
airplane	water; grass field	water; grass field
bear	rocky terrain; green grass; waterfall; grass field	grass texture; rocky terrain; green grass; waterfall; grass field
bicycle	rocky terrain; grass; green grass; water; reflection on water; grass and wildflowers	grass texture; rocky terrain; grass; green grass; water; reflection on water; grass and wildflowers
bird	water; water surface; outdoor setting; autumn lighting; green grass	water; water surface; outdoor setting; autumn lighting; green grass
bus	water; grass; blue water	water surface; water; grass; blue water
butterfly	background grass; green grass; natural outdoor setting; dry grass; green grass field	green grass; natural outdoor setting; dry grass; green grass field
cactus	none	none
car	grass; water surface; water body; grass field; water tower; boats on water; rocky ground; rocky terrain; blue water	grass; water surface; water body; grass field; water tower; boats on water; rocky ground; rocky terrain; blue water
cat	outdoor setting	outdoor setting
chair	grass; outdoor setting	grass; outdoor setting
clock	none	none
corn	natural outdoor setting; swimming pool water	natural outdoor setting; swimming pool water
cow	green grass; green grass field; green grassy field; dry grassy field; stream water; grassland	grass texture; green grass; green grass field; green grassy field; dry grassy field; stream water; grassland
crab	rock formation; rocky surface; water; water reflection; blue water	rock formation; rocky surface; water; water reflection; blue water
crocodile	water surface; grass surface; outdoor setting; algae-covered water surface	grass surface; outdoor setting; algae-covered water surface
dog	outdoor setting; reflection on water; rocky terrain; man standing on rocks; rocky surface; rocks on shore	outdoor setting; reflection on water; rocky terrain; man standing on rocks; rocky surface; rocks on shore
dolphin	water splash; water surface; blue water; water	water surface; blue water; water
elephant	dry grassland; small elephant in water	water; dry grassland; small elephant in water

(continued on next page)

(Table 16 continued from previous page)

Category	ERM matched phrases	GroupDRO matched phrases
fishing rod	water body; trees with autumn foliage; water surface; calm water surface; water; rocks; rocks on ground; man standing in water; fish in water; person standing on rock	water body; trees with autumn foliage; water surface; calm water surface; water; rocks; rocks on ground; man standing in water; fish in water; person standing on rock
flower	rock; rocks; water; rocky surface; gray rocks	rock; rocks; water; rocky surface; gray rocks
football	swimming pool water; poolside rocks; autumn setting	swimming pool water; poolside rocks; autumn setting
fox	reflections on water; green water surface	reflections on water; green water surface
frog	green grass; natural outdoor setting	green grass; natural outdoor setting
giraffe	rocky cliff	rocky cliff
goose	water surface; reflective water; reflection on water; water body; icy water	reflective water; reflection on water; water body; icy water
gun	grass field; sand or grassy ground; rocky surface; grass	grass field; sand or grassy ground; rocky surface; grass
hat	water; grass field; green grass; grass ground	grass field; green grass; grass ground
helicopter	rocky canyon walls; water	rocky canyon walls; water
horse	grass field; clear water; water body; tall grass field; green grassy field	clear water; water body; tall grass field; green grassy field
hot air balloon	grass field; water reflection	grass field; water reflection
kangaroo	grass surface	grass surface
lifeboat	water	water
lion	none	none
lizard	rocky terrain	rocky terrain
mailbox	water surface	water surface
monkey	rock surface; grass; rocky surface; muddy water	rock surface; grass; rocky surface; muddy water
motorcycle	blue water	blue water
ostrich	green grass; outdoor setting	green grass; outdoor setting
owl	rocks; outdoor setting	rocks; outdoor setting
pineapple	water ripples; water; water texture; swimming pool water; grass; water surface; pool water; water droplets on bottle	water texture; swimming pool water; grass; water surface; pool water; water droplets on bottle
pumpkin	grass; autumn setting; pool water	grass; autumn setting; pool water
rabbit	grassland in the distance; autumn foliage	grassland in the distance; autumn foliage
racket	grass surface; green grass; outdoor setting; overgrown grass	grass surface; green grass; outdoor setting; overgrown grass
sailboat	water surface; water; rock; calm water surface; reflection on water; rocky outcrop; blue water; water body	water surface; rock; calm water surface; reflection on water; rocky outcrop; blue water; water body
scooter	water; grass	water; grass
seal	rocky surface; water texture; water; water surface; clear blue water	water texture; water; water surface; clear blue water
sheep	grass field; sheep drinking water; rocky hillside	grass field; sheep drinking water; rocky hillside
ship	water; water surface; rocky shore; clear blue water; rocks; green grass; grass; rocky cliff; autumn trees; blue water	water; clear blue water; rocks; green grass; grass; rocky cliff; autumn trees; blue water
shrimp	water; water surface; rocks; rocky surface; underwater environment	water; water surface; rocks; rocky surface; underwater environment
spider	water surface; dried grass; outdoor setting	water surface; dried grass; outdoor setting
squirrel	grass field; blurred grassy background; rocky surface; moss-covered rock	grass field; blurred grassy background; rocky surface; moss-covered rock
sunflower	none	none
tent	green grass; grass surface	green grass; grass surface
tiger	rocks; autumn foliage	rocks; autumn foliage
tortoise	log in water; branch in water; reflection on water; dry grass	log in water; branch in water; reflection on water; dry grass
train	grass field	grass field

(continued on next page)

(Table 16 continued from previous page)

Category	ERM matched phrases	GroupDRO matched phrases
truck	dry grass; grass; grass surface; water puddles on grass; blue water	dry grass; grass; grass surface; water puddles on grass; blue water
umbrella	beachgoer in water; autumn foliage; rocky cliff; autumn setting	beachgoer in water; autumn foliage; rocky cliff; autumn setting
wheat	none	none
wolf	water surface; grass; rocky surface; water body; autumn trees	grass; rocky surface; water body; autumn trees

G.7 Baseline Comparison at Matched Phrase Budget

Metric definitions for Table 2. All methods are scored with the matcher and threshold of §4.2, against each method's spurious phrases alone and not its full catalog. Label-side coverage is the share of SI's 630 spurious labels whose best-matching spurious phrase of the method reaches θ. Catalog-side coverage is the share of those 630 labels that are the best match of at least one spurious phrase of the method, with each phrase counting toward one label only. Precision is per-phrase macro precision, and phrases is the mean number of spurious phrases per class. LANCE is the full upstream pipeline run on a 40-class subset and scored over its 36 SI-labeled classes, where it recovers 24 of 180 labels (13.3%). No catalog-side entry is reported for LANCE, since its subset does not cover the 630-label set.

Table 2 caps no method's phrase list, so the per-phrase precision gap could in principle reflect budget rather than phrase quality. To test this, we re-score the archived matcher records with each baseline capped to its first five phrases per class, matching PROBE's average budget of 4.9 phrases per class. The cap keeps each method's stored output order (SpLiCE's sparse-decomposition weight order and B2T's keyword ranking), so it retains each method's own top-ranked phrases. Table 17 reports the outcome. At the matched budget SpLiCE's precision rises to 0.089, still under half of PROBE's 0.200, while its catalog-side recall drops from 30.3% to 20.8%. B2T's precision is essentially unchanged and its recall drops from 22.2% to 14.0%. PROBE therefore leads both axes at equal phrase budget, and the precision gap of Table 2 is not an artifact of parsimony.

Table 17: Matched-budget re-scoring of the Table 2 comparison from the archived matcher records (same 357 classes, matcher, and $\theta = 0.30$). Capped rows keep each baseline's first five phrases per class in its own ranking order.

Method	Phrases/class	Precision	Catalog-side recall
PROBE (DISCOVER)	4.9	**0.200**	**34.4**% (217)
SpLiCE (full)	≈14.7	0.052	30.3% (191)
SpLiCE (top-5)	4.9	0.089	20.8% (131)
B2T (full)	≈8.9	0.081	22.2% (140)
B2T (top-5)	4.4	0.079	14.0% (88)

G.8 Waterbirds (extended)

Setup. Waterbirds provides a complementary validation setting. Unlike ImageNet, where class-specific shortcut cues are diverse, Waterbirds has a single, uniform background bias structure (water-background vs. land-background) that every classifier in a typical pool exploits. To probe whether the agreement and disagreement between DISCOVER and SCORE observed on Salient ImageNet (App. G.9) is intrinsic to the framework or contingent on the dataset, we run a parallel audit on Waterbirds (Sagawa et al., 2020), the canonical background-shortcut benchmark. The pool comprises 14 classifiers, namely an ERM-trained RESNET-50 obtained from the LADDER repository (Ghosh et al., 2025), four CLIP variants (ViT-B/32, ViT-B/16, ViT-L/14, ViT-L/14-336), and nine SIGLIP2 variants (BASE, BASE-256, BASE-384, BASE-512, LARGE, LARGE-384, LARGE-512, GIANT-256, GIANT-384). We sample 300 positives per class and disable confuser-class negative sampling because the benchmark is binary (the only confuser is the other class, which DISCOVER sources directly from baseline confusions). All other hyperparameters of §4 are unchanged. The run produces 218 raw VLM-discovered concepts, 94 unique cues after dedup, 650 counterfactual edits, and 40 rendered

Table 18: Discrimination test. The audited ERM RESNET-50 vs. two debiased variants of the same backbone that we train ourselves (App. G.8), on the identical Waterbirds probe set. Real-image WGA is measured on the Waterbirds test split.

Arm (same backbone)	Real WGA	Conf. shift	confuser-shortcut
ERM (audited pool member)	0.751	−0.212	0.786
DFR (last-layer retrain)	0.908	−0.103	0.857
GroupDRO (fine-tune)	0.858	−0.015	0.929

probe images (16 for `waterbird`, 4 per variant, and 24 for `landbird`, 7 per variant except 3 real-cue-only renders).

Catalog overview. DISCOVER's pool-aggregated catalog shows the same strict-tier density as the 5-classifier ImageNet-100 audit of §4.1. The bias-spurious category contains 47/37 (any/strict) cues for `waterbird` and 36/29 for `landbird`, a total of 83/66, against 994/806 on the 100-class ImageNet pool (prompt-echo strings excluded in both). With 14 classifiers concurring and a binary task whose well-known shortcut is exploited by every architecture in the pool, strict cross-classifier agreement is the rule. The most-flagged spurious cues include `bamboo forest`, `forest background`, `green foliage`, and `green trees`. This is the literature-defined Waterbirds shortcut, found with the ImageNet hyperparameters unchanged. Several bird-anatomy cues (`beak`, `head`, `feather texture`, `head feathers`) also land in bias-spurious, through VLM fallback assignment or because the VLM typed them as environmental, since the rule set maps target-cue removal to essential.

The LADDER-repository baseline. The RESNET-50 obtained from the LADDER repository (Ghosh et al., 2025) is the repository's ERM base checkpoint (Hugging Face `shawn24/Ladder`, `Waterbirds_ERM_hparams0_seed0`, with a measured real-image worst-group accuracy of 0.751). The DFR arm retrains the last layer on a group-balanced validation subset, and the GroupDRO arm fine-tunes the full network with GroupDRO over the four groups for 60 epochs, the reported epoch being the best by test worst-group accuracy, so its WGA is optimistic. It posts confidence shift −0.212, in the same direction and order of magnitude as the zero-shot CLIP/SigLIP2 variants. The discrimination test of §4.3 scores DFR and GroupDRO variants of this same backbone on the identical probe set and finds |confidence shift| shrinking to 0.103 and 0.015 respectively. The audit responds to the training method, not merely the architecture. The small probe budget (2 classes, 3–7 rendered probes per variant) inflates noise on any single classifier's point estimate, so the reliable evidence remains the cross-classifier sign agreement (14 of 14 negative), not any individual value.

Full 14-classifier point estimates. Table 19 reports the per-classifier top-1 accuracy on the four probe variants (per-class averages, Eq. (4)) and the per-class-averaged confidence shift for all 14 classifiers.

Table 19: Per-classifier top-1 accuracy on the four probe variants and the per-class-averaged confidence-shift score on Waterbirds. With $|\mathcal{Y}| = 2$ classes, per-class bootstrap CIs are degenerate. We report point estimates only. Top-1 values are per-class averages (Eq. (4)), consistent with Table 18. The informative observation is the cross-classifier sign agreement on confidence shift (14/14 negative), not any specific value. Columns `bh`, `bs`, `rco`, and `csh` abbreviate bias-heavy, bias-stripped, real-cue-only, and confuser-shortcut.

Classifier	`bh`	`bs`	`rco`	`csh`	Conf. shift
ERM ResNet-50 (LADDER repo.)	0.786	0.661	1.000	0.786	−0.212
CLIP-ViT-B/32	0.857	0.339	1.000	0.857	−0.099
CLIP-ViT-B/16	0.482	0.482	0.750	0.804	−0.128
CLIP-ViT-L/14	0.679	0.286	0.750	0.750	−0.099
CLIP-ViT-L/14-336	0.625	0.339	0.875	0.625	−0.155
SigLIP2-base	0.786	0.411	0.875	0.857	−0.054
SigLIP2-base-256	0.929	0.411	0.875	0.857	−0.114
SigLIP2-base-384	0.857	0.411	0.875	0.857	−0.099
SigLIP2-base-512	0.857	0.482	0.875	0.857	−0.093
SigLIP2-large	0.679	0.411	1.000	0.804	−0.241
SigLIP2-large-384	0.679	0.411	1.000	0.804	−0.233
SigLIP2-large-512	0.732	0.464	1.000	0.804	−0.223
SigLIP2-giant-256	0.607	0.411	0.875	0.679	−0.198
SigLIP2-giant-384	0.661	0.339	0.875	0.804	−0.231

G.9 What PROBE Does Not Measure (extended)

This section gives the full version of the scope paragraph that closes §4.3.

Saliency visibility versus interventional reliance. SI's labels record saliency *visibility* and contain no interventions, so they do not directly validate the Phase 5 edit test, which asks whether editing the same cue changes classifier confidence, a different, interventional question. Accordingly, category-level agreement in §4.2 is lower than cue-level recovery, with 283 of the 630 SI labels (45%, or 53% of the surfaced ones) also landing in the bias-spurious category and the remaining 246 forming a disagreement set. Feeding SI's labels directly into SCORE makes this divergence explicit. The aggregate confidence shift is -0.010, but the per-class distribution is wide, with comparable positive and negative tails (28%/29% at the ± 0.10 cut, Table 14). Per-class spurious recall and per-class confidence shift correlate at Pearson $r = 0.008$ ($n = 357$). These results indicate that saliency labels, edit tests, and counterfactual shifts capture different aspects of "spuriousness" at the per-class level. Depending on the intervention-based criterion, 28–45% of SI labels exhibit spurious behavior at the aggregate level. Details are in App. G.5 and App. H.2.

Cross-family limits. Confidence shift shows little association with real-image group robustness, either across the 14-classifier Waterbirds pool (Spearman $\rho = -0.05$ against WGA) or across the five-classifier ImageNet-100 pool against the ImageNet-9 background gap (Xiao et al., 2021) ($\rho = 0.10$), while confuser-shortcut top-1 is directionally consistent with ImageNet-9 mixed-random accuracy ($\rho = 0.60$, $n = 5$, underpowered). Confidence shift is therefore an audit-specific response measure rather than a surrogate for real-image group robustness. Confidence scales differ across model families (on Waterbirds, sigmoid SigLIP2-large mean confidences occupy 0.10–0.36 where softmax CLIP approaches 1.0) and the Waterbirds probe budget is small (App. F). We therefore interpret confidence shift magnitude, as well as cue inventories gated by the fixed τ, only within an audit and within an architecture family. Across families, confidence shift is compared only through its ordering, and magnitude comparisons rest on the accuracy-based axes. This ordering is an audit-relative descriptive comparison and should not be read as a calibration-invariant measure of relative shortcut reliance. On Waterbirds, with $|\mathcal{Y}| = 2$, the per-class average is dominated by `landbird` (the ERM ResNet-50 posts -0.426 on `landbird` and $+0.002$ on `waterbird`), which is why §4.3 interprets only the sign.

G.10 EXTENDED RESULTS: Z-CBM, IMAGENET-LT, LVIS

Z-CBM concept acquisition (§4.3). The concept bank follows Z-CBM's construction (Yamaguchi et al., 2026) at reduced scale. Noun phrases are chunked from 150k Conceptual Captions with the determiner–adjective–noun grammar `NP: DT? JJ* NN+`, cleaned and length-filtered to 80,966 unique phrases (Z-CBM's own bank draws on Flickr30K, CC3M, CC12M, and YFCC15M, about 5M concepts). For each of the 100 audited classes we embed 10 ImageNet-1k validation images and the full bank with CLIP-ViT-B/32, retrieve the top-12 concepts by mean image–text cosine similarity (our choice, where Z-CBM retrieves per image with a larger budget), and drop concepts whose similarity to the class name exceeds 0.9 (our class-leakage filter). QWEN2.5-VL then labels each concept as an intrinsic part of the object or a contextual element (single-word forced choice). Intrinsic concepts populate the real category and contextual ones the bias-spurious category of the SCORE input catalog. Probe rendering requests a fixed 4 images per variant per class, as in the ImageNet-100 run, and yielded 1,427 rendered probes. The run records of the original Z-CBM instantiation were lost, and Table 20 reports a re-instantiation (rebuilt bank, re-rendered probes, seeded decoding) scored with the five-classifier pool. Two differences from the DISCOVER-driven run stand out. Absolute levels drop (bias-heavy 0.36–0.44 vs. 0.59–0.63), consistent with generic caption phrases being weaker contextual anchors than classifier-tailored discoveries. The bias-stripped variant fires at 0.14–0.17 vs. 0.07 because retrieved contextual phrases often embed the object itself (e.g. "large plastic shark"). The bootstrap protocol of §4 applies unchanged.

Table 20: SCORE on ImageNet-100 with concepts acquired by a reduced Z-CBM-style retrieval (Yamaguchi et al., 2026) instead of DISCOVER (class-averaged point estimate over 100 classes and 95% bootstrap CI, $n_{\text{boot}} = 10{,}000$, paired class resampling). The confidence-shift ordering agrees with Table 1 at $\rho = 0.9$ (single inversion, RESNET-50/VIT-L/16). Columns `bh`, `bs`, `rco`, and `csh` abbreviate bias-heavy, bias-stripped, real-cue-only, and confuser-shortcut.

Classifier	`bh`	`bs`	`rco`	`csh`	Conf. shift
ResNet-50	$0.385_{[.31,.46]}$	$0.155_{[.10,.21]}$	$0.412_{[.34,.49]}$	$0.318_{[.25,.40]}$	$-0.016_{[-.08,.05]}$
DINOv2-ViT-B/14+lc	$0.438_{[.36,.52]}$	$0.158_{[.10,.22]}$	$0.426_{[.34,.51]}$	$0.315_{[.24,.39]}$	$+0.009_{[-.05,.07]}$
ViT-L/16	$0.395_{[.32,.47]}$	$0.148_{[.10,.20]}$	$0.425_{[.34,.51]}$	$0.323_{[.25,.40]}$	$-0.003_{[-.06,.05]}$
CLIP-ViT-B/32	$0.362_{[.29,.43]}$	$0.165_{[.11,.23]}$	$0.393_{[.31,.47]}$	$0.295_{[.22,.37]}$	$-0.039_{[-.09,.01]}$
SigLIP2-base	$0.385_{[.31,.46]}$	$0.138_{[.09,.19]}$	$0.394_{[.31,.48]}$	$0.288_{[.22,.37]}$	$-0.118_{[-.17,-.06]}$

ImageNet-LT (§4.1). We adopt the OLTR split files of Liu et al. (2019) and tier definitions head > 100, medium 20–100, tail < 20 training images (385/477/136 classes reachable in our label mapping, 2 unmapped). The full pipeline runs with the §4 protocol and the N_c formula of Table 24. SCORE covers the 832 classes with at least one discovered spurious cue. Table 21 reports the main results. Every classifier collapses from 0.62–0.70 on bias-heavy to 0.12–0.15 on bias-stripped. Aggregate confidence shift turns mildly negative for all five classifiers (-0.06 to -0.03), matching the SI-357 aggregate and the real-cue-only render-penalty confound (App. G.3). All three ranking extremes reach significance ($\pi_{\text{clean}}\ \Delta = +0.064$, $\pi_{\text{csh}}\ \Delta = +0.055$, $\pi_{\text{ctx}}\ \Delta = -0.033$, all $p < 10^{-3}$), the statistical power the 100-class audit lacked. Between ImageNet-100 and ImageNet-LT the π_{ctx} rank correlation is $\rho = 0.4$, with only DINOv2's position as the classifier with the highest confidence shift persisting. Per-tier mean confidence shift is in Table 22.

Table 21: Full-pipeline SCORE results on ImageNet-LT (all 1,000 classes, of which the 832 with a discovered spurious cue are scored, with image-pooled point estimate and class-resampled 95% bootstrap CI, see *Estimator audit*). The bias-heavy $\gg$ bias-stripped collapse of Table 1 replicates at $10\times$ scale. Columns `bh`, `bs`, `rco`, and `csh` abbreviate bias-heavy, bias-stripped, real-cue-only, and confuser-shortcut.

Classifier	`bh`	`bs`	`rco`	`csh`	Conf. shift
ResNet-50	$0.631_{[.60,.66]}$	$0.140_{[.12,.16]}$	$0.673_{[.64,.70]}$	$0.486_{[.46,.52]}$	$-0.063_{[-.09,-.04]}$
DINOv2-ViT-B/14+lc	$0.702_{[.67,.73]}$	$0.149_{[.13,.17]}$	$0.707_{[.68,.74]}$	$0.491_{[.46,.52]}$	$-0.030_{[-.05,-.01]}$
ViT-L/16	$0.654_{[.63,.68]}$	$0.148_{[.13,.17]}$	$0.682_{[.65,.71]}$	$0.467_{[.44,.50]}$	$-0.047_{[-.07,-.02]}$
CLIP-ViT-B/32	$0.618_{[.59,.65]}$	$0.117_{[.10,.14]}$	$0.643_{[.61,.67]}$	$0.435_{[.40,.47]}$	$-0.049_{[-.07,-.03]}$
SigLIP2-base	$0.678_{[.65,.71]}$	$0.135_{[.12,.16]}$	$0.695_{[.66,.73]}$	$0.487_{[.46,.52]}$	$-0.059_{[-.08,-.04]}$

Table 22: Mean confidence shift by ImageNet-LT frequency tier (full 1,000-class run, §4.1). All entries sit in a mildly negative band, and no tier consistently shows the highest confidence shift across the five classifiers.

Classifier	Head	Medium	Tail
ResNet-50	$-0.067_{[-.10,-.03]}$	$-0.058_{[-.09,-.02]}$	$-0.070_{[-.12,-.03]}$
DINOv2-ViT-B/14+lc	$-0.048_{[-.08,-.01]}$	$-0.021_{[-.05,.01]}$	$-0.015_{[-.07,.03]}$
ViT-L/16	$-0.053_{[-.09,-.02]}$	$-0.040_{[-.07,-.01]}$	$-0.055_{[-.10,-.02]}$
CLIP-ViT-B/32	$-0.042_{[-.08,-.01]}$	$-0.058_{[-.09,-.03]}$	$-0.030_{[-.08,.02]}$
SigLIP2-base	$-0.027_{[-.06,.01]}$	$-0.089_{[-.12,-.06]}$	$-0.035_{[-.09,.02]}$

LVIS benchmark construction (§4.1). From the LVIS v1 validation annotations we crop every positively-annotated instance with 10% box padding, capping at 40 crops per category and discarding degenerate boxes (<8 px min side). This covers $1{,}032$ of the $1{,}203$ categories. Most of the missing categories have no positively annotated validation instance under LVIS's federated annotation, and the remainder have only degenerate boxes. Federated annotations are respected by cropping only positively-annotated instances. Zero-shot classifiers score LVIS labels through per-category text prompts ("a photo of a {label}"). Category names are used verbatim (underscores and parentheticals stripped). The audited pool is CLIP-ViT-B/32, CLIP-ViT-L/14-336, SigLIP2-base, and SigLIP2-giant-384, run with the §4 protocol (10 positives per class, $\tau = 0.15$). The zero-shot classifiers provide no attention map, so Phase 2 sees the image alone in this audit. SCORE covers the 265 categories for which DISCOVER surfaces a spurious cue. Table 23 reports the main results. The bias-heavy $\gg$ bias-stripped collapse holds (0.42–0.55 vs. 0.12–0.14 against a $1/1{,}032$ floor), all four classifiers post negative confidence shift (-0.068 to -0.038), and the $\pi_{\text{clean}}/\pi_{\text{csh}}$ extremes are significant ($\Delta = +0.078/+0.082$, $p < 10^{-3}$) while π_{ctx} is borderline ($\Delta = -0.030$, $p = 0.053$).

Table 23: Open-vocabulary SCORE results on LVIS instance crops ($1{,}032$ categories and four zero-shot classifiers, with the 265 categories that have a discovered spurious cue scored, image-pooled point estimate and class-resampled 95% bootstrap CI). All classifiers post negative confidence shift, extending the Waterbirds $14/14$ sign agreement to open vocabulary. Columns `bh`, `bs`, `rco`, and `csh` abbreviate bias-heavy, bias-stripped, real-cue-only, and confuser-shortcut.

Classifier	`bh`	`bs`	`rco`	`csh`	Conf. shift
CLIP-ViT-B/32	$0.424_{[.37,.47]}$	$0.119_{[.09,.15]}$	$0.507_{[.45,.56]}$	$0.347_{[.30,.39]}$	$-0.068_{[-.10,-.04]}$
CLIP-ViT-L/14-336	$0.500_{[.45,.55]}$	$0.143_{[.11,.18]}$	$0.542_{[.49,.60]}$	$0.388_{[.34,.44]}$	$-0.060_{[-.09,-.03]}$
SigLIP2-base	$0.545_{[.49,.60]}$	$0.144_{[.11,.18]}$	$0.572_{[.52,.63]}$	$0.425_{[.38,.48]}$	$-0.056_{[-.08,-.03]}$
SigLIP2-giant-384	$0.549_{[.50,.60]}$	$0.134_{[.10,.17]}$	$0.585_{[.53,.64]}$	$0.428_{[.38,.48]}$	$-0.038_{[-.07,-.01]}$

Estimator audit. Table 1 is class-averaged exactly as Eq. (4) specifies (recomputation from the archived per-class records matches all 25 cells to $\Delta = 0.000$). The ImageNet-LT and LVIS tables were generated image-pooled. Recomputing them class-averaged from the archived per-class records shifts point estimates by at most 0.017, leaves every reported value inside its class-averaged 95% CI, and changes no extreme identity or significance conclusion (ImageNet-LT extremes all at the 10^{-4} bootstrap floor, and LVIS π_{ctx} $p = 0.053$ both ways).

G.11 COMPLETE HYPERPARAMETER RECORD (REPRODUCIBILITY)

Table 24 records every hyperparameter for the four ImageNet-scale SCORE audits so each run can be reproduced. All runs share the same pipeline stack, VLM QWEN2.5-VL-7B-INSTRUCT and diffusion editor/renderer FLUX.2-KLEIN-9B-KV (bf16, seeded 4-step sampling), and the same statistical protocol (95% bootstrap CIs, $n_{\text{boot}} = 10{,}000$, paired per-class resampling, fixed seed). Probe rendering is seeded (seed base 42, image size 768×768). The diffusion model, the classifiers, and the VLM all run in inference mode only (no training or fine-tuning anywhere in the audit pipeline, while the DFR and GroupDRO arms of §4.3 are trained separately).

Pool footnotes. [a] RESNET-50, DINOV2-VIT-B/14+LC, VIT-L/16, CLIP-VIT-B/32, SIGLIP2-BASE. [b] CLIP-VIT-B/32, CLIP-VIT-L/14-336, SIGLIP2-BASE, SIGLIP2-GIANT-384 (open label space forces a zero-shot-only pool). [c] Our reduced Z-CBM-style retrieval replaces DISCOVER's concept acquisition entirely. Its knobs are a concept bank of $80{,}966$ noun phrases from 150k

Table 24: Hyperparameter record for the four ImageNet-scale SCORE audits (the SI-357 and Waterbirds probe sets are described in App. G.5 and App. G.8). †Primary ImageNet-100 run of §4. Shared across all runs are VLM QWEN2.5-VL-7B + editor FLUX.2-KLEIN-9B-KV (bf16, 4-step), $\tau = 0.15$, Score-CAM attention where the classifier provides one, 10 positives and up to 25 hard negatives per class, min. negative confidence 0.05, top-5 predictions recorded, probe image size 768^2, probe seed base 42, probe cue source `any`, and bootstrap $n_{\text{boot}} = 10{,}000$.

Setting	ImageNet-100†	Z-CBM (§4.3)	ImageNet-LT (§4.1)	LVIS (§4.1)
Dataset / split	ImageNet-1k val	ImageNet-1k val	ImageNet-1k (OLTR LT)	LVIS v1 val crops
Categories audited	100	100	1,000	1,032
Scored by SCORE	99	100	832	265
Classifier pool	5^a	5^a	5^a	4^b
Concept source	DISCOVER	Z-CBM^c	DISCOVER	DISCOVER
Checkpoint batch size	—	—	30	30
Prompt groups per variant N_c	4 (fixed)	4	$\max(1, \lceil\sqrt{\lvert bias - spurious_c\rvert}\rceil)$	4
Scored probe images	1,553	1,427	6,156	4,091

Conceptual Captions, retrieval backbone CLIP-VIT-B/32, 10 retrieval images per class, top-12 concepts retrieved, class-name similarity cutoff 0.9, intrinsic/contextual grouping by QWEN2.5-VL, and the 2 nearest classes as confusers (App. G.10). *Dataset specifics:* the ImageNet-LT tiers (head >100 / medium 20–100 / tail <20 training images) use the OLTR split. The LVIS benchmark uses ≤ 40 crops/category with 10% box padding and an 8-px minimum side (App. G.10).

ViT-L/16 checkpoint disclosure. The Table 7/1 audit used the SWAG-E2E 512px checkpoint. The runs of §4.3–§4.1 used the IMAGENET1K-V1 224px checkpoint under the same "ViT-L/16" label. On identical re-rendered Z-CBM probes the V1 checkpoint scores confidence shift -0.003 while SWAG-512 scores $+0.025$. On real ImageNet-9 images the two checkpoints differ $2\times$ in background gap (0.029 vs. 0.059). Rankings within each audit are unaffected (each audit scores one checkpoint throughout). Cross-audit comparisons involving "ViT-L/16" should note the checkpoint difference, and §4.3 reports the checkpoint-matched agreement.

G.12 PROBE-FIDELITY AUDIT

We assess probe fidelity on the ImageNet-100 run by grading every bias-stripped and real-cue-only render with QWEN2.5-VL for target-object presence, and we retain the graded records. The target should be absent from bias-stripped probes and present in real-cue-only probes. The grader flags the target as present in 41 of 400 bias-stripped renders (10%, renderer leakage) and absent from 102 of 381 real-cue-only renders (27%, the studio-render penalty of App. G.3). Re-scoring the five classifiers without those renders gives Table 25. The grader is conservative. Of the 46 bias-stripped renders that still receive a correct label after filtering, visual inspection finds the object in most (a harvestman on moss, a scorpion, a shark, a snake held in a hand), so the filtered accuracies remain upper bounds. The grader's miss rate was not measured against human labels, so all rates are uncorrected for grader error. The Z-CBM, ImageNet-LT, and LVIS probe sets were not graded, and their bias-stripped accuracies are reported as leakage-inclusive upper bounds.

Table 25: Table 1 re-scored after removing the 41 bias-stripped renders the grader flags as leaked and the 102 real-cue-only renders it flags as target-absent. Point estimates over the 78 classes that keep at least one render of every variant, with 95% paired-bootstrap CIs. The π_{ctx} extreme (SIGLIP2-BASE vs. DINOV2-VIT-B/14+LC) is $\Delta = -0.083$, $p = 0.001$, and the π_{clean} extreme is $\Delta = +0.101$, $p = 0.014$.

Classifier	bh↑	bs↓	rco↑	csh↑	Conf. shift
RESNET-50	$0.702_{[.62,.78]}$	$0.011_{[.00,.02]}$	$0.611_{[.52,.70]}$	$0.516_{[.44,.60]}$	$0.061_{[.01,.11]}$
DINOV2-VIT-B/14+LC	$0.708_{[.62,.79]}$	$0.030_{[.01,.06]}$	$0.637_{[.54,.73]}$	$0.587_{[.50,.67]}$	$0.064_{[.02,.11]}$
VIT-L/16	$0.683_{[.60,.77]}$	$0.027_{[.00,.06]}$	$0.644_{[.54,.74]}$	$0.532_{[.44,.62]}$	$0.018_{[-.02,.06]}$
CLIP-VIT-B/32	$0.660_{[.58,.74]}$	$0.028_{[.00,.06]}$	$0.605_{[.51,.70]}$	$0.510_{[.43,.59]}$	$0.017_{[-.03,.07]}$
SIGLIP2-BASE	$0.676_{[.59,.76]}$	$0.016_{[.00,.04]}$	$0.706_{[.61,.80]}$	$0.577_{[.49,.66]}$	$-0.019_{[-.07,.03]}$

G.13 Decision-Provenance Sensitivity Analysis

Replaying the deterministic Phase 5 rule over the 16,813 archived ImageNet-100 edit records quantifies the assignment split. Of all edits, 37.9% fail the near-identity gate, 22.0% are below-τ for every pool member, and of the remaining 33,685 (significant-edit $\times$ classifier) pairs, 21.4% are covered by the rule and 78.6% are VLM-assigned (70.1%/29.9% on the single-model SI-357 run). Recorded decisions match the rule for every classifier-edit pair to which it applies. Of the VLM-assigned classifier-edit pairs, 79.6% belong to a classifier whose own $|\Delta|$ was below τ (the gate is pool-level, App. B.2), which is the main mechanism behind the smaller strict-rule inventories. The same mechanism inflates cross-classifier agreement. Table 26 recounts the strict-agreement tier of §4.1 under three gates on the same archived records. With the pool-level gate, 806 of 994 bias-spurious cues are flagged by all five classifiers. When each classifier must cross τ on its own edit and the recorded decision is kept, all-five agreement falls to 262 of 846 (31%) while at-least-two agreement stays at 725 (86%). Under strict rules only it falls further to 122 of 713 (17%). The abstract and §4.1 report the own-gate figures, so every headline agreement count rests on the classifier's own confidence change, and the VLM's role in those counts is limited to naming the category of a classifier-edit pair the gate has already admitted.

Table 26: Cross-classifier agreement on bias-spurious cues under three gating regimes on ImageNet-100, prompt-echo strings excluded. Columns give the number of cues flagged by exactly k of the five classifiers.

Regime	any	k=1	k=2	k=3	k=4	k=5	classes with k=5
Pool gate, rule or VLM decision (as run)	994	28	49	63	48	806	99
Own gate, recorded decision	846	121	153	155	155	262	71
Own gate, rules only	713	146	163	142	140	122	52

Table 27 reports Table 7 under both regimes. Robust to the regime choice, bias-spurious dominates every classifier's inventory (spur $>$ state $>$ real throughout), and all probe-stage results (§4.1) are regime-independent. Regime-dependent, in contrast, are the absolute inventory sizes (strict-rule bias-spurious drops to 3.0–5.3 per class) and the "within roughly one cue of each other" cross-architecture uniformity of §4.1, which holds only in the as-run regime and should be read accordingly.

Table 27: Category sensitivity. Average per-class cue counts under the as-run regime (rule + VLM fallback assignment) and under strict-rules-only (VLM-assigned classifier-edit pairs excluded), computed on the catalog before the prompt-echo exclusion of App. G.2, so the as-run columns exceed Table 7 by the echo entries. spur and state abbreviate the bias-spurious and bias-state categories.

	As-run			Strict rules only		
Classifier	real	spur	state	real	spur	state
ResNet-50	1.18	9.14	2.61	0.84	4.49	1.55
DINOv2-ViT-B/14+lc	0.83	9.79	2.09	0.67	3.74	1.12
ViT-L/16	0.87	9.65	2.28	0.57	2.97	1.01
CLIP-ViT-B/32	1.04	9.68	2.17	0.94	5.33	1.68
SigLIP2-base	1.06	9.96	1.98	0.89	5.14	1.45

G.14 Null Control and Architecture Extension

Table 28 re-scores the existing ImageNet-100 probe set and archived counterfactual edits with three additional classifiers at inference-only cost, namely ConvNeXt-B (Liu et al., 2022) (modern CNN), Swin-B (Liu et al., 2021b) (window-attention ViT), and an untrained random-initialization ResNet-50 as a null control. The null control is clean at both stages, with probe accuracy at or below the $1/1{,}000$ chance floor, confidence shift 0.000 to three decimals, and zero strict-rule assignments in every substantive category (all cues inconclusive). The control cannot exercise the VLM fallback assignment, since no edit moves a near-uniform confidence by τ. The two new architectures land within rounding of the audited pool's bands (ConvNeXt-B `rco` 0.550 vs. the pool's 0.492–0.545, Swin-B bias-heavy−bias-stripped 0.517 vs. 0.520–0.558), extending §4.1's findings to two previously uncovered regimes. A further re-scoring of the identical probes and archived edits adds six more architectures (ResNet-152, ConvNeXt-L, ViT-B/16, Swin-T, and Swin-V2-T/B). Their bias-heavy−bias-stripped gaps span $+0.520$–$+0.563$ against the pool's

$+0.520$–$+0.558$, so the thirteen-model band is $+0.518$ to $+0.563$, and strict-rule spurious dominance holds for all (4.5–$5.5\times$ spurious:real).

Table 28: Probe metrics and strict-rule categories for the extension pool on the existing artifacts (class-averaged over the 99 scored classes). The random-init null control produces no signal at either stage. Columns `bh`, `bs`, `rco`, and `csh` report top-1 accuracy on bias-heavy, bias-stripped, real-cue-only, and confuser-shortcut, shift is the confidence-shift score of Eq. (2), and real, spur, and state are the average per-class strict-rule cue counts in the real, bias-spurious, and bias-state categories.

	Probe metrics					Strict-rule categories		
	`bh`	`bs`	`rco`	`csh`	shift	real	spur	state
ConvNeXt-B	0.621	0.073	0.550	0.503	$+0.049$	0.67	3.03	1.17
Swin-B	0.593	0.076	0.530	0.487	$+0.061$	0.64	3.47	1.26
ResNet-152	0.604	0.078	0.507	0.472	$+0.070$	0.77	4.26	1.45
ConvNeXt-L	0.634	0.071	0.542	0.510	$+0.072$	0.61	2.98	1.04
ViT-B/16	0.614	0.063	0.522	0.480	$+0.065$	0.73	4.02	1.26
Swin-T	0.596	0.076	0.532	0.477	$+0.061$	0.79	3.54	1.47
Swin-V2-T	0.619	0.071	0.538	0.500	$+0.066$	0.72	3.66	1.47
Swin-V2-B	0.616	0.071	0.526	0.495	$+0.081$	0.66	3.06	1.27
ResNet-50 (random init)	0.000	0.003	0.000	0.003	$+0.000$	0.00	0.00	0.00

G.15 Design Rationale

Each pipeline decision rules out a specific failure of its alternative. Table 29 pairs every decision with that failure and the measurement that supports it.

Table 29: Design rationale. What each pipeline decision rules out, and the measured evidence.

Decision	Alternative it rules out	Measured evidence
VLM hypothesis generation	Predefined concept banks cannot name classifier-specific shortcuts	Class-name-only concepts: Jaccard 0.031 overlap, shortcut signal erased (App. G.17)
Attention-map conditioning	Coverage loss without a localization prior; the map is not the explanation	Jaccard 0.78–0.80 across attribution methods vs. 0.54–0.64 degraded (App. G.16)
Diffusion counterfactual edits	Correlational co-occurrence cannot separate a cue the model *uses* from one merely present	Null control: zero findings for an untrained classifier (App. G.14)
Rule fast path with VLM fallback assignment	A purely learned judge would prevent replay of the rule-covered classifier-edit pairs	Both regimes reported, 21.4% of classifier-edit pairs rule-based and 78.6% VLM-assigned (App. G.13)
Four probe variants	A single scalar conflates reliance with accuracy	Three-axis disagreement at the point-estimate level, non-significant after correction at 100 classes and significant at ImageNet-LT scale (§4.1, §4.1)
Cross-architecture pool	A single classifier's inventory cannot show which cues are shared across architectures	$806/994$ spurious cues flagged by all five classifiers, 188 by fewer (§4.1)

G.16 Heatmap-Source Ablation

DISCOVER conditions the VLM on an attention map (Phase 1). To measure how much the discovered hypotheses depend on that map, we re-run the discovery stage only (RESNET-50, 40 classes, 6 positives per class, fixed seed, fixed parser) under six conditions and compare each arm's per-class concept sets against the Score-CAM reference by Jaccard similarity (Table 30). Alternative attribution methods barely move the hypotheses (Grad-CAM 0.781, Grad-CAM++ 0.796). Degrading the map moves the hypotheses further (random 0.642, blurred 0.578, none 0.536) without shrinking the inventory (33.6–38.1 concepts per class). The map acts as a localization prior, while the edit gate maintains decision precision.

Table 30: Heatmap-source ablation. Per-class Jaccard of discovered concept sets vs. the Score-CAM reference arm (discovery only, RESNET-50, 40 classes, 6 positives per class, fixed seed, fixed parser). The reference arm averages 35.5 concepts per class.

Condition	Jaccard vs. Score-CAM	Concepts/class
Grad-CAM++	0.796	36.0
Grad-CAM	0.781	35.5
Random map	0.642	35.4
Blurred map	0.578	33.6
No map	0.536	38.1

G.17 COMPONENT SWAPS AND CATALOG ROBUSTNESS

Editor swap. Replacing FLUX.2 with the much weaker INSTRUCTPIX2PIX (Brooks et al., 2023) (based on Stable Diffusion 1.5, Rombach et al., 2022) changes which cues get verified (25 classes, RESNET-50 arm, Jaccard 0.128 between the two spurious sets), but on the 135 cues verified by both editors the decision agreement is 93.3% (126/135). A substantially weaker open editor reaches the same conclusions on the cues it can render.

VLM capacity. QWEN2.5-VL at 3B/7B/32B produces 24.9/35.5/56.0 concepts per class. Capacity moves hypothesis breadth monotonically, and the catalogs are not string-stable across sizes (Jaccard 0.097 and 0.108 against the 7B catalog). Decision correctness was not measured for the 3B and 32B arms, since only discovery was run. The verified catalog is the layer validated externally in §4.2.

Class-name-only concepts. Concepts written by the language model from the class name alone overlap DISCOVER's catalog at Jaccard 0.031 (40 classes) and weaken the shortcut signal in SCORE (confidence shifts of -0.022 to $+0.041$ against -0.013 to $+0.097$ with DISCOVER's catalog), showing that per-classifier discovery is not restating what a class name implies (§4.3).

Pool composition. Leave-one-classifier-out on the pooled catalog retains $\geq 94.8\%$ of every probe-feeding category (real 94.8%, bias-spurious 98.7%, bias-state 96.0%) under any single removal. Only the inconclusive category, which never feeds probes, is unstable.

Hypothesis phrasing. Within-class categorized hypothesis pairs land in the same category in 88.1–93.3% of cases at cosine similarity 0.5–0.7, 90.0–96.1% at 0.7–0.9, and 100% for near-paraphrases (≥ 0.9, 26/26 pairs across the five classifiers). Phrasing bias therefore changes naming granularity more than decision direction (strict-rules regime). If editing a subtly off phrasing also perturbs the true cue, the confidence delta fires and the decision is valid for the edit performed. If the edit misses the cue, the hypothesis lands inconclusive, unless the edit happens to perturb another cue that fires a rule.

H SELECTED PER-EXAMPLE VISUALIZATIONS

This appendix shows representative qualitative examples drawn from the runs underlying §4.1 (DISCOVER) and §4.1, App. G.5, and App. G.8 (SCORE). App. H.1 shows four DISCOVER edit-test cases, namely two state-change edits whose confidence drops (A1-03, A1-12), one essential cue removal (A1-07), and one state-change amplification that raises confidence (A1-17). App. H.2 shows five SCORE probe-set cases drawn from the Salient ImageNet, Waterbirds, and ImageNet runs.

H.1 DISCOVER: COUNTERFACTUAL-EDIT CASE STUDIES

Each figure shows the original image x and the edited image x'_i produced by the diffusion editor under the listed instruction. Confidence values are taken from the model named in each caption.



x (original) x_i' (edited)

Figure 7: **[A1-03] Waterbirds, target class `landbird`**, LADDER RESNET. Cue: *bird wings* (target). Edit type: `state_change`. Instruction: *"Adjust wing pose to a more relaxed, slightly spread position."* Outcome: state-bias. Confidence: $1.000 \rightarrow 0.000$ ($\Delta = -1.000$).



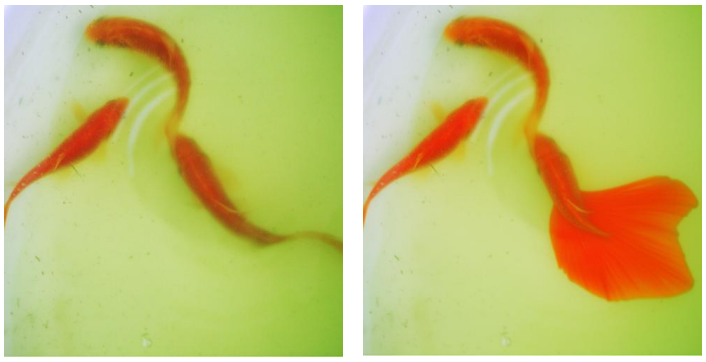
x (original) x_i' (edited)

Figure 8: **[A1-07] Waterbirds, target class `waterbird`**, LADDER RESNET. Cue: *wingspan* (target). Edit type: `feature_removal`. Instruction: *"Remove the wingspan, blend smoothly with feathers."* Outcome: essential (a sanity check, since removing the bird's wingspan collapses the prediction). Confidence: $0.998 \rightarrow 0.026$ ($\Delta = -0.973$).



x (original) x_i' (edited)

Figure 9: **[A1-12] ImageNet, target class `Gila monster, Heloderma suspectum`**, RESNET-50. Cue: *head* (target). Edit type: `state_change`. Instruction: *"Head color change to darker brown with lighter spots."* Outcome: state-bias. Confidence: $1.000 \rightarrow 0.001$ ($\Delta = -0.999$).



x (original) x_i' (edited)

Figure 10: **[A1-17] ImageNet, target class `goldfish, Carassius auratus`**, RESNET-50. Cue: *tail fin* (target). Edit type: `state_change`. Instruction: *"Increase tail fin size by 20%, make it more vibrant orange."* Outcome: state-bias, VLM-assigned, since no rule fires on a confidence gain for a target cue. The outcome here arises from a confidence *gain* after amplifying a state-specific cue rather than removing it. Confidence: $0.026 \rightarrow 0.966$ ($\Delta = +0.940$).

H.2 SCORE: PROBE-SET CASE STUDIES

Each figure shows a per-class SCORE probe set rendered by the four variant prompts (bias-heavy shows the target with its contextual cues, bias-stripped the same context without the target, real-cue-only the target on a neutral background with only object-intrinsic cues, and confuser-shortcut the target with biased context plus confuser-class cues under unusual pose or lighting). Available probe slots vary. A2-03 has only the first two probes per variant rendered. The other four examples have all four. Captions report per-class top-1 accuracy on each variant, Δaccuracy as the bias-heavy minus bias-stripped top-1 gap, and the confidence shift as Eq. (2) evaluated on that class. For the multi-model cases A2-07 and A2-13 they summarize cross-architecture agreement.



Figure 11: **[A2-03] Salient-ImageNet, target class `sarong`** on RESNET-50. Top-1: bias-heavy 1.00, bias-stripped 0.00, real-cue-only 0.00, confuser-shortcut 0.50. Mean confidence on bias-heavy 0.999, real-cue-only 0.097, and bias-stripped 0.001, and confidence shift +0.902. The real-cue-only renders show garments on hangers rather than a worn sarong, so part of the shift reflects render failure. Only the first two probes per variant were rendered for this class (the other two renders were not produced).

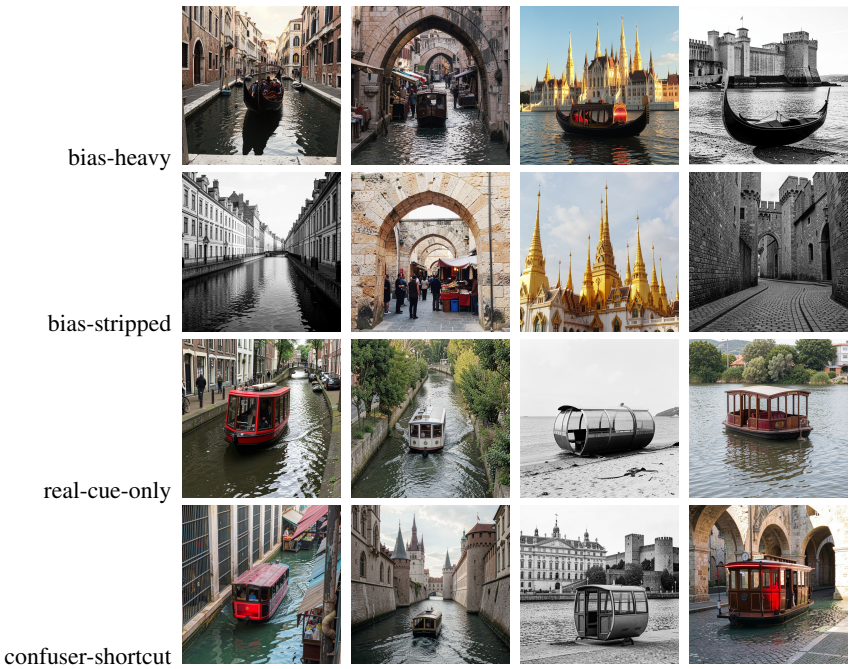


Figure 12: **[A2-04] Salient-ImageNet, target class `gondola`** on RESNET-50. Top-1: bias-heavy 1.00, bias-stripped 0.00, real-cue-only 0.00, confuser-shortcut 0.50. Mean confidence on bias-heavy 0.893, real-cue-only 0.012, and bias-stripped 0.052, and confidence shift +0.881. The real-cue-only renders show boats without a gondola's form, so part of the shift reflects render failure.

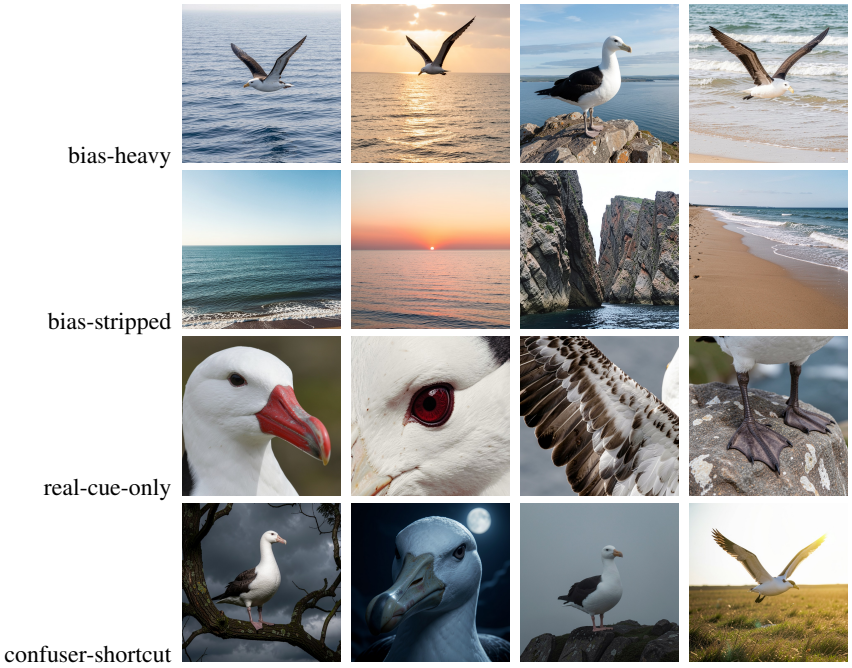


Figure 13: **[A2-05] Salient-ImageNet, target class `albatross, mollymawk`** on RESNET-50. Top-1: bias-heavy 1.00, bias-stripped 0.00, real-cue-only 0.00, confuser-shortcut 1.00. Mean confidence on bias-heavy 0.927, real-cue-only 0.127, and bias-stripped 0.001, and confidence shift +0.800.

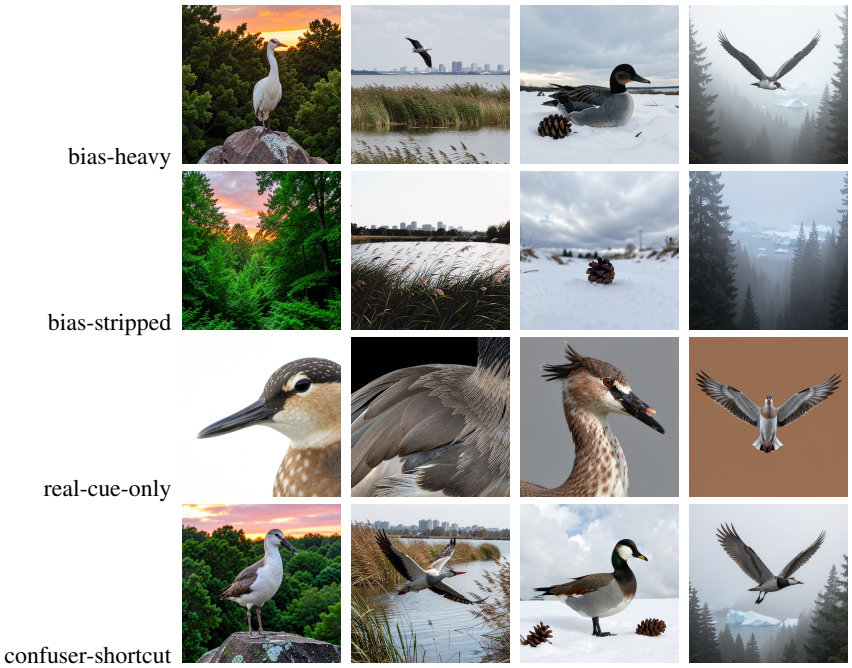


Figure 14: **[A2-07] Waterbirds, target class `waterbird`** (multi-model, 14 classifiers). Mean Δaccuracy across the pool $= +0.43$ (min $+0.00$, max $\mathrm{acc}(bias - stripped) = 0.75$). Per-classifier numbers for these probes are reported in Table 19.

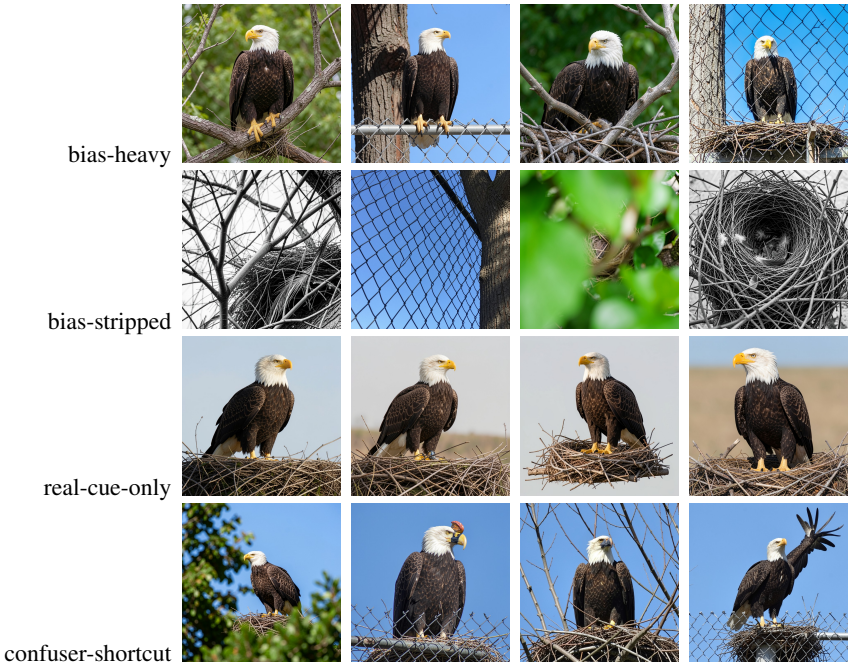


Figure 15: **[A2-13] ImageNet, target class `bald eagle`, `American eagle`, `Haliaeetus leucocephalus`** (multi-model, 5 classifiers, namely ResNet-50, DINOv2-ViT-B/14+LC, ViT-L/16, CLIP-ViT-B/32, SigLIP2-BASE). Every classifier produces top-1 1.00 on bias-heavy, real-cue-only, and confuser-shortcut, but 0.00 on bias-stripped, a Δaccuracy of $+1.00$ across all five (max $\mathrm{acc}(bias - stripped) = 0.00$). With the target absent, no classifier fires on the context alone, the expected behaviour when the decision ignores context on this class.

I Licenses and Terms of Use for Existing Assets

Table 31 lists the external datasets, benchmarks, pretrained models, and software tools used in this work. We cite the corresponding original papers in the main paper or appendix and link the repositories, dataset pages, or model cards below. We use these assets only for academic research and evaluation, comply with attribution requirements, and do not redistribute third-party datasets, checkpoints, or model weights unless permitted by their licenses or terms of use.

Table 31: External assets used in this work, together with their sources and license or terms-of-use information.

Asset	Type	License / terms	Source and usage notes
ImageNet-1k validation split	Dataset / benchmark	ImageNet terms of access; non-commercial research and educational use	Used for evaluation only. ImageNet states that it does not own the copyright of the underlying images and provides access to researchers and educators under specific terms. Source: `https://www.image-net.org/`.
ImageNet-100	Dataset subset	Derived from ImageNet; ImageNet terms of access apply	Used as an ImageNet-derived evaluation subset (a Kaggle dataset of 100 uploader-chosen classes, created 2021). We do not redistribute the images and rely on the original ImageNet access mechanism. Source: `https://www.image-net.org/`.
Salient ImageNet	Dataset / benchmark	Custom non-commercial research license of the official release (University of Maryland, research institutions only, no redistribution); ImageNet-derived terms also apply where ImageNet images are used	Used for evaluating saliency and spurious-cue behavior. We cite the Salient ImageNet paper and the paper documenting the 1,000-class release, and link the official repository. Source: `https://github.com/singlasahil14/salient_imagenet`.
Waterbirds	Dataset / benchmark	Non-commercial research and educational use; derived from CUB and Places	Used for group-shift and spurious-correlation evaluation. We cite the GroupDRO / Waterbirds paper, link its repository, and cite the upstream datasets where appropriate. Source: `https://github.com/kohpangwei/group_DRO`.
ResNet-50	Pretrained classifier	Depends on checkpoint source; for torchvision, BSD-style PyTorch / torchvision software license, with ImageNet-trained weights subject to ImageNet-related considerations	Used as an audited ImageNet classifier. We cite the ResNet paper and link the checkpoint source used in our experiments. Source: `https://pytorch.org/vision/stable/models.html`.

Continued on next page

Asset	Type	License / terms	Source and usage notes
DINOv2-ViT-B/14 with linear head	Pretrained classifier / representation model	Apache License 2.0	Used as an audited ImageNet classifier. We cite the DINOv2 paper and link the repository and model card. Source: `https://github.com/facebookresearch/dinov2`.
ViT-L/16	Pretrained classifier	BSD 3-Clause License for torchvision. The SWAG-E2E weights used in the main audit are released under CC BY-NC 4.0 (facebookresearch/SWAG); the IMAGENET1K-V1 weights follow torchvision's model-weight terms	Used as an audited ImageNet classifier under two checkpoints (App. G.11). We use the ViT-L/16 implementation and pretrained weights from torchvision, cite the Vision Transformer paper, link torchvision, and do not redistribute the checkpoints. Source: `https://pytorch.org/vision/stable/models.html`.
CLIP-ViT-B/32	Pretrained vision-language classifier	MIT License for OpenAI CLIP code/model release	Used as an audited classifier in both ImageNet and Waterbirds pools. We cite the CLIP paper and link the OpenAI CLIP repository. Source: `https://github.com/openai/CLIP`.
CLIP-ViT-B/16	Pretrained vision-language classifier	MIT License for OpenAI CLIP code/model release	Used as an audited Waterbirds classifier. We cite the CLIP paper and link the checkpoint source. Source: `https://github.com/openai/CLIP`.
CLIP-ViT-L/14	Pretrained vision-language classifier	MIT License for OpenAI CLIP code/model release	Used as an audited Waterbirds classifier. We cite the CLIP paper and link the checkpoint source. Source: `https://github.com/openai/CLIP`.
CLIP-ViT-L/14-336	Pretrained vision-language classifier	MIT License for OpenAI CLIP code/model release	Used as an audited Waterbirds classifier. We cite the CLIP paper and link the checkpoint source. Source: `https://github.com/openai/CLIP`.
SigLIP2-base	Pretrained vision-language classifier	Apache License 2.0	Used as an audited classifier. We cite the SigLIP2 paper and link the Google Hugging Face model card. Source: `https://huggingface.co/google/siglip2-base-patch16-224`.

Continued on next page

Asset	Type	License / terms	Source and usage notes
SigLIP2-base-256	Pretrained vision-language classifier	Apache License 2.0	Used as an audited Waterbirds classifier. We cite the SigLIP2 paper and link the Google SigLIP2 collection that hosts the checkpoint. Source: `https://huggingface.co/collections/google/siglip2`.
SigLIP2-base-384	Pretrained vision-language classifier	Apache License 2.0	Used as an audited Waterbirds classifier. We cite the SigLIP2 paper and link the Google SigLIP2 collection that hosts the checkpoint. Source: `https://huggingface.co/collections/google/siglip2`.
SigLIP2-base-512	Pretrained vision-language classifier	Apache License 2.0	Used as an audited Waterbirds classifier. We cite the SigLIP2 paper and link the Google model card. Source: `https://huggingface.co/google/siglip2-base-patch16-512`.
SigLIP2-large-256	Pretrained vision-language classifier	Apache License 2.0	Used as an audited Waterbirds classifier (denoted SigLIP2-large in the tables). We cite the SigLIP2 paper and link the Google model card. Source: `https://huggingface.co/google/siglip2-large-patch16-256`.
SigLIP2-large-384	Pretrained vision-language classifier	Apache License 2.0	Used as an audited Waterbirds classifier. We cite the SigLIP2 paper and link the Google model card. Source: `https://huggingface.co/google/siglip2-large-patch16-384`.
SigLIP2-large-512	Pretrained vision-language classifier	Apache License 2.0	Used as an audited Waterbirds classifier. We cite the SigLIP2 paper and link the Google SigLIP2 collection that hosts the checkpoint. Source: `https://huggingface.co/collections/google/siglip2`.
SigLIP2-giant-256	Pretrained vision-language classifier	Apache License 2.0	Used as an audited Waterbirds classifier. We cite the SigLIP2 paper and link the Google SigLIP2 collection that hosts the checkpoint. Source: `https://huggingface.co/collections/google/siglip2`.
SigLIP2-giant-384	Pretrained vision-language classifier	Apache License 2.0	Used as an audited Waterbirds classifier. We cite the SigLIP2 paper and link the Google model card. Source: `https://huggingface.co/google/siglip2-giant-opt-patch16-384`.

Continued on next page

Asset	Type	License / terms	Source and usage notes
LADDER ResNet	Pretrained classifier / released model	MIT License (Hugging Face model card)	Used as an audited Waterbirds classifier. The checkpoint is the ERM base model `Waterbirds_ERM_hparams0_seed0` hosted at `https://huggingface.co/shawn24/Ladder`. We cite the LADDER paper and link the official repository. Source: `https://github.com/batmanlab/Ladder`.
Qwen2.5-VL-7B-Instruct	Vision-language model	Apache License 2.0	Used as a VLM component in the pipeline. We cite the Qwen2.5-VL technical report and link the model card. Source: `https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct`.
FLUX.2-klein-9b-kv	Diffusion editor / generator	FLUX Non-Commercial License	Used only for non-commercial academic research and evaluation. We do not redistribute the model weights. Source: `https://huggingface.co/black-forest-labs/FLUX.2-klein-9b-kv`.
all-MiniLM-L6-v2	Text embedding model	Apache License 2.0	Used to compute text embeddings. We cite the Sentence-BERT paper and link the SentenceTransformers model card. Source: `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`.
Score-CAM	Saliency / explanation method	MIT License for common official implementations	Used to compute the Phase 1 attention maps. We cite the Score-CAM paper and link the implementation used. Source: `https://github.com/haofanwang/Score-CAM`.
Grad-CAM and Grad-CAM++	Saliency / explanation methods	Method papers. The reference repositories carry no license file (the Grad-CAM README names BSD). We implement both methods in our own code	Used as alternative attribution sources in the heatmap-source ablation (App. G.16). We cite the Grad-CAM and Grad-CAM++ papers and link the reference repositories. Sources: `https://github.com/ramprs/grad-cam`, `https://github.com/adityac94/Grad_CAM_plus_plus`.

Continued on next page

Asset	Type	License / terms	Source and usage notes
InstructPix2Pix	Diffusion editor	MIT License (model card). Built on Stable Diffusion 1.5 weights, which carry the CreativeML Open RAIL-M license	Used only as the weaker counterfactual editor in the editor-swap ablation (App. G.17). We cite the InstructPix2Pix and Stable Diffusion papers and link the model card. We do not redistribute the weights. Source: `https://huggingface.co/timbrooks/instruct-pix2pix`.
Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-32B-Instruct	Vision-language models	Qwen Research License for the 3B model, Apache License 2.0 for the 32B model	Used only in the VLM-capacity ablation (App. G.17). We cite the Qwen2.5-VL technical report and link the model cards. Sources: `https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct`, `https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct`.
ConvNeXt-B/L, Swin-T/B, Swin-V2-T/B, ResNet-152, ViT-B/16	Pretrained classifiers (torchvision)	BSD 3-Clause License for torchvision, with ImageNet-trained weights subject to ImageNet-related considerations	Used only for inference-time re-scoring of the archived probes and edits in the pool extension (App. G.14). We link torchvision and do not redistribute the checkpoints. Source: `https://pytorch.org/vision/stable/models.html`.
ImageNet-LT (OLTR split)	Dataset split / benchmark	Split files from the OLTR repository (see repository). Images under ImageNet terms of access	Used for the long-tail SCORE audit (App. G.10). We cite the OLTR paper and link its repository. Source: `https://github.com/zhmiao/OpenLongTailRecognition-OLTR`.
LVIS v1	Dataset / benchmark	Annotations under CC BY 4.0. Underlying COCO images are subject to their Flickr terms of use	Used to build the open-vocabulary instance-crop benchmark (App. G.10). We cite the LVIS paper and link the dataset page. Source: `https://www.lvisdataset.org/`.
Conceptual Captions	Dataset (image alt-text captions)	Conceptual Captions terms of use (see dataset page)	Used only as the noun-phrase source for the Z-CBM concept bank (App. G.10). We cite the Conceptual Captions paper and link the dataset page. Source: `https://ai.google.com/research/ConceptualCaptions/`.
Hard ImageNet	Dataset / benchmark	No license file in the repository. The dataset's datasheet states no limitations on use. Images under ImageNet terms of access	Used as a ground-truth spurious-cue benchmark (§4.2, App. Table 15). We cite the Hard ImageNet paper and link the repository. Source: `https://github.com/mmoayeri/HardImageNet`.

Continued on next page

Asset	Type	License / terms	Source and usage notes
NICO++	Dataset / benchmark	Non-commercial research use per the authors' README (no formal license file)	Used as a ground-truth context-cue benchmark in its 60-category, six-context public release (§4.2, App. Table 16). We train the ERM and GroupDRO ResNet-50 classifiers ourselves. We cite the NICO++ paper and link the repository. Source: `https://github.com/xxgege/NICO-plus`.
ImageNet-9 (Backgrounds Challenge)	Dataset / benchmark	No license file in the repository. ImageNet-derived images under ImageNet terms of access	Used to compute real-image background gaps (§4.3, App. G.11). We cite the Backgrounds Challenge paper and link the repository. Source: `https://github.com/MadryLab/backgrounds_challenge`.
B2T, SpLiCE, and LANCE	Baseline discovery methods (code)	B2T repository without a license file; SpLiCE and LANCE under Apache License 2.0	Used as annotation-free discovery baselines (§4.2, App. G.7). We cite the three papers and link their repositories. Sources: `https://github.com/alinlab/b2t`, `https://github.com/AI4LIFE-GROUP/SpLiCE`, `https://github.com/virajprabhu/lance`.