

Detectors Say 'Different', Not 'What': Score, Deviation and Name Are Three Claims With Three Referees, Only the Name Needs a Referee Outside the Detector, and Published Results Show Novelty Detected Far More Reliably Than It Is Named

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. An anomaly or novelty detector reports that an input is unlike the data it was built from. Users want more: they want to know what the unusual thing is. Explainable anomaly detection is often presented as supplying that second answer, but three different claims travel under the one word "explanation". A score claim says which parts of the input raised the detector's score; a deviation claim says in what respect the input departs from the reference data; a name claim says what the input is, in a vocabulary a person can act on. This paper argues that the three are checked by different referees: the detector can check the first, the reference data the second, and only something outside both, a label, a benchmark description or a person, can check the third. Assembling published results, it finds that explaining the score is valuable precisely because the score is often driven by something other than the novelty, as generative likelihoods driven by background pixels and an industrial detector partly driven by high-frequency background show; that detectors built on vision-language models do compute their scores from names, yet a correct detection can rest on wrong names; and that where naming has been measured against ground truth it has mostly lagged detection badly: in one open-vocabulary video study, novel categories were detected at 88.2 AUC but named correctly 37.1 percent of the time with the right names in the vocabulary, although a later model on the same task named novel categories on a second dataset almost as well as trained ones. No located study measures, on the same instances, both whether a name is correct and whether it names what the detector responded to. The paper states that missing measurement and the studies that would supply it.

Keywords: anomaly detection; novelty detection; out-of-distribution detection; explainable anomaly detection; open-vocabulary recognition; vision-language models; evaluation methodology; explanation faithfulness

1. Introduction

A long-standing complaint about anomaly detection in deployed systems is not that it misses things. It is that it cannot say what it found. Sommer and Paxson (2010), explaining why machine-learning anomaly detection had seen so little operational use in network intrusion detection, list among its difficulties "a semantic gap between results and their operational interpretation", and describe the needed final step as answering, from the operator's point of view, "What does it mean?" Their diagnosis is precise: an anomaly detector is not built to find attacks but to report what it has not seen before, benign or not, and "by definition, a machine learning algorithm does not make any mistakes within its model of normality; yet for the operator it is the results' interpretation that matters".

Fifteen years on, the question has a literature. Surveys of explainable anomaly detection organise dozens of methods by what they explain and how (Li, Zhu and van Leeuwen, 2023; Wang et al., 2023), and a unifying review of deep and shallow methods uses explanation techniques to audit the detectors it compares (Ruff et al., 2021). The open-world learning programme names the missing step outright: after detection comes novelty characterisation, then accommodation (Jafarzadeh et al., 2020; Chadwick et al., 2023), and a related formal

framework for theories of novelty defines a family of novelty types (Boult et al., 2021). Industrial inspection has datasets with pixel-precise defect masks and more than seventy defect types (Bergmann et al., 2019), and a newer dataset that adds logical anomalies, such as permissible objects in invalid locations, alongside local structural defects (Bergmann et al., 2022). Vision-language models now sit inside detectors and are asked to describe what they detect (Gu et al., 2024; Zhang, H. et al., 2024; Du et al., 2024).

The premise this paper examines is the one the topic states bluntly: a detector emits a purely negative signal, that the input is far from what it knows, while an explanation of novelty needs a positive characterisation, which requires a vocabulary the detector does not contain. If that is right, attribution methods explain the score and not the novelty, and a language model that supplies a name offers no guarantee that the name matches the deviation that triggered the score.

The published evidence supports part of that premise, refutes part, and sharpens the rest.

It supports the claim that explaining the score is not explaining the novelty, but not in the way the premise implies. The gap between score and substance is real and has been measured, and it was found by explaining the score. Likelihood-based detectors assign higher

likelihood to an entirely different dataset than to their own (Nalisnick et al., 2019), and a per-pixel decomposition of the likelihood shows it dominated by background pixels (Ren et al., 2019). Heatmaps of industrial detectors show one keyed to wood grain rather than the hole in the wood (Kauffmann, Ruff et al., 2020), and another drawing a non-negligible share of its score from high-frequency background that disappears under a better resizing algorithm, taking the detector's F1 from 0.92 to 0.80 (Kauffmann, Dippel et al., 2025). Score explanations are the instrument that exposes these failures. They are not a defective attempt at naming.

It refutes the claim that the signal is purely negative, for one large family of detectors. Zero-shot detectors built on vision-language models compute their scores against text: in-distribution class names (Ming, Cai et al., 2022), large banks of negative labels (Jiang, Liu et al., 2024), outlier class names envisioned by a language model (Cao et al., 2024), or candidate names for the unknown generated per input (Esmaeilpour et al., 2022). For these, a name is part of the computation. But the best-documented case also shows why that is not enough. In the example Esmaeilpour et al. use to illustrate their method, an image of a boat is correctly flagged as unknown while "boat" is absent from the candidate names the model generated for it; the score, which sums probability over all candidate unknown names, came out right with the names wrong.

And it sharpens the claim about correspondence into a measurement that has not been made. Where naming has been scored against ground truth, it has mostly trailed detection by a wide margin, even when the correct name is supplied. In open-vocabulary video anomaly detection, one model detected novel anomaly categories at 88.2 frame-level AUC on UCF-Crime and 76.0 AP on XD-Violence, better on XD-Violence than it detected the categories it was trained on, yet named novel categories correctly 37.1 and 30.9 percent of the time, against 49.0 and 89.3 percent for trained categories, with every novel name in the test vocabulary (Wu et al., 2024). A later model on the same task closed most of that naming gap on XD-Violence, naming novel categories correctly 87.4 percent of the time against 92.4 percent for trained ones, but not on UCF-Crime, where the figures were 41.6 and 56.9 percent (Li, F. et al., 2025). On an industrial benchmark of multiple-choice questions, the best multimodal model scored 94.98 percent on identifying the object and 65.80 percent on classifying the defect (Jiang, Li et al., 2025). Withholding the vocabulary costs naming heavily within a single study: on ordinary benchmark images, a vocabulary-free classifier recovered 17.7 percent word overlap with the true class name, averaged over ten datasets, against 69.4 percent for CLIP given the ground-truth names (Conti et al., 2023). None of these studies asks whether the name, when right, names the thing the detector responded to.

The contribution of this paper is a decomposition and what it predicts. Section 3 separates the score claim, the deviation claim and the name claim and argues that each has its own referee. Sections 4 to 6 assemble what each kind of explanation has been shown to deliver. Section 7 examines detectors whose scores are computed from names. Section 8 sets out what would tie a name to the deviation and which published checks supply part of it. Section 9 collects the numbers; Section 10 states the decision structure as an algorithm; Section 11 states the studies that would settle the open part. Sections 12 to 14 give the strongest case against the argument, what is not known, and the limitations.

The paper's strongest claims are these. Explaining a score, however faithfully, is a statement about the detector and its reference data, and cannot certify a name. A name can be certified only against something outside the detector. The published measurements of naming, taken together, show a large gap in most settings between detecting that something is new and saying what it is; in the two studies that put the right names in the vocabulary, the gap persisted on one dataset in both, and on the other it closed to about five points in the later study. On the question the topic asks directly, whether a name matches the deviation that triggered the score, the literature located here contains no measurement at all.

2. Scope, terms, and what this paper is not

Terms. Anomaly detection, novelty detection, open-set recognition, out-of-distribution detection and outlier detection differ in their assumptions about training data and test-time labels, and a unified framework treats them as special cases of one problem (Yang et al., 2021). This paper uses "detector" for any of them and "flagged input" for an input whose score crosses a threshold. It follows the classical definition in taking an anomaly to be a pattern that does not conform to expected behaviour (Chandola et al., 2009), and takes novelty to be judged, likewise, against a model of normal data. "Reference" means whatever the detector's notion of normal is built from: training data, a density model, a memory bank of normal features, or a set of in-distribution class names.

The three claims. A score claim is a statement of the form "these features, regions or time steps of x account for $s(x)$ ". A deviation claim is a statement of the form " x differs from the reference in this respect". A name claim is a statement of the form " x is, or contains, a T ", with T drawn from a vocabulary. Section 3 defines these more carefully. The paper calls the joint claim that a name T describes the respect in which x deviates, and that this respect is what raised $s(x)$, the correspondence claim.

What this paper is not. It is not a survey of explainable anomaly detection; two exist (Li, Zhu and van Leeuwen, 2023; Wang et al., 2023), and this paper cites methods

only where they bear on the three-claim argument. It is not about detection accuracy, and it takes no side on which detector is best. It is not about set-level description of distribution shift as a task in its own right, although Section 8 draws on that literature because it contains the most developed verification machinery for names. It does not address whether a flagged input should be acted on; a related line of work argues that the useful question for a runtime monitor is whether the model's prediction is wrong, not whether the input is new (Guerin et al., 2023), and that question is orthogonal to naming. It does not address hypothesis generation from anomalies in science.

Evidence standard. Measured values are reported as the source reports them, with the table or section named where a number was taken from the body. Where two sources use different metrics, the paper says so and does not compare them on one scale. The paper's own claims are of three kinds: definitional (Section 3), inferential from published results (Sections 4 to 8), and proposals (Sections 10 and 11). The verbs are meant to tell them apart.

3. Three claims, three referees

3.1 Why a score is contrastive by construction

Nearly every detector in the literature cited here computes a score by comparing an input with a reference. A density model scores the negative log-likelihood of x under a model fitted to normal data. A distance model scores the distance from x to its nearest normal neighbours, or to the nearest entries of a memory bank of normal patch features (Roth et al., 2022). A one-class model scores the distance of x 's embedding from a centre where normal data were concentrated (Liznerski et al., 2021). A reconstruction model scores how badly x is reconstructed by a network trained to reconstruct normal data. A vision-language detector scores how weakly x matches any in-distribution class name (Ming, Cai et al., 2022) or how strongly it matches a non-normal text (Jeong et al., 2023; Zhou et al., 2024).

In each case the score is a function of two arguments, the input and the reference, and it is large when they disagree. This is what the topic means by a negative signal. The score does not say that x has property P ; it says that x lacks the properties the reference has, to a degree. Chandola et al. (2009) organise their survey around the same structure: for each family of techniques they identify the key assumptions it uses to tell normal from anomalous behaviour, and those assumptions are stated relative to what normal behaviour looks like.

3.2 Score, deviation, name

Three things can be said about a flagged input x .

The score claim answers: what in x produced $s(x)$? Its referee is the detector itself. A score claim is correct if intervening on the named parts of x changes $s(x)$ as the

claim predicts, and if the claim depends on the detector and not only on the input (Adebayo et al., 2018). Deletion tests, model-randomisation tests and counterfactuals built against the detector all check score claims.

The deviation claim answers: in what respect does x differ from the reference? Its referee is the reference data. A deviation claim is correct if, in the named respect, x is separable from the reference data; outlier-explanation methods that search for the feature subspace in which an outlier stands apart from inliers use this criterion (Micenkova et al., 2013; Macha and Akoglu, 2018). A deviation claim can be true of x and the reference while being false of the detector: x may differ from the reference in a respect the detector ignores.

The name claim answers: what is x ? Its referee is neither the detector nor the reference data. A name claim is correct if x is in fact a T , which is a fact about the world and is certified by a label, a ground-truth description, or a competent person. The detector's reference contains, by construction, no instances of the novel T , and a detector that has never been told the word T cannot check whether it applies.

The correspondence claim, that name T describes the respect in which x deviates and that this respect is what raised $s(x)$, is a conjunction of all three. To check it one needs the detector (for the score part), the reference (for the deviation part), and an external referee (for the name part), and one needs them on the same instances.

3.3 Three consequences

First, the three claims can come apart in every combination. A score claim can be correct while the deviation it identifies is not the one a person cares about: Kauffmann, Ruff et al. (2020) show heatmaps that correctly attribute a kernel density detector's score to wood grain, while the defect is a small perforation. A name claim can in principle be correct while the score claim it would imply is false: a vision-language model could describe a defect correctly in an image that a detector flagged because of its background. And a detection can be correct with the name wrong, as in the boat example of Section 1.

Second, deviation and type are not tied one to one. Two inputs of different kinds can deviate from the reference in the same feature, and two inputs of the same kind can deviate in different features. The second has been observed: Jacob et al. (2021) report it in their benchmark for explainable anomaly detection over time series: for global explanations, the concordance of explanations across anomalies of the same type was harder to achieve than stability for all three explanation methods tested, so that a user is likely to be shown explanations resting on different features for anomalies of one type. A deviation explanation, however accurate, does not by itself identify a type.

Third, the referee for names can be moved but not removed. Open-world recognition as first formalised treats the step between detecting unknowns and adding classes as a labelling process applied to the collected unknowns, "e.g. by humans" (Bendale and Boulton, 2015). Every method in Sections 6 to 8 that supplies names does so by importing a vocabulary, from a label set, a text corpus, a language model or a person, and the correctness of the name then depends on that import. Table 1 summarises the three claims.

4. Score explanations explain the score, and that is what exposed the gap

4.1 Likelihood scores and the background

The clearest demonstrations that a detector's score and the novelty a person has in mind can diverge come from generative models used as density-based detectors. Nalisnick et al. (2019) report that flow-based models, variational autoencoders and PixelCNNs trained on CIFAR-10 assign higher likelihood to SVHN house numbers than to their own test data, and find the same pattern for several other dataset pairs, including Fashion-MNIST against MNIST. Kirichenko et al. (2020) trace the flow case to what flows learn: local pixel correlations and generic image-to-latent transformations that are not specific to the training dataset. Serra et al. (2020) attribute the problem to the excessive influence of input complexity on likelihoods and derive a complexity-corrected score.

The case most directly about explanation is Ren et al. (2019). With a PixelCNN++ trained on Fashion-MNIST and MNIST as the out-of-distribution set, the raw likelihood detects MNIST at an AUROC of 0.089 in their Table 1, far worse than chance, and the likelihood correlates with the proportion of background (zero-valued) pixels in an image with a Pearson coefficient of 0.85. A per-pixel decomposition shows the likelihood dominated by background pixels, while their likelihood ratio, which subtracts a background model, concentrates on the pixels that carry the object and raises the AUROC to 0.994. The per-pixel map in this study is a score explanation in the sense of Section 3: it says which pixels account for the score. It is correct about the score and it reveals that the score is about the background.

This is the first complication of the premise. The premise treats "attribution explains the score, not the novelty" as a limitation of attribution. In these cases it is the reason to run attribution. Without the per-pixel decomposition, the only evidence of the problem would be the AUROC, which says that something is wrong and not what.

4.2 Clever Hans in anomaly detection

Kauffmann, Ruff et al. (2020) apply an explanation procedure to three detector types (a kernel density estimator, an autoencoder and a deep one-class model) on corrupted MNIST and on MVTec-AD, and find the

detectors producing high scores for the wrong reasons across all three. For the kernel density model on the MVTec wood class, the heatmap shows high-frequency wood stripes appearing anomalous and completely dominating the small perforations that are the actual defects. The autoencoder, on a bottle with a large contamination at its centre, bases its score on fine elements at the bottle's outer edge. The deep one-class model, on a zipper, looks mainly at the boundary between the zipper fabric and the background rather than at the opening. The authors argue that these effects stem from the structure of the models rather than from spurious correlations in the data, and give an analysis in which the kernel density score approximates a difference-to-the-mean strategy in input space.

The later study by Kauffmann, Dippel et al. (2025) turns this into a deployment failure. Their explanation of a distance-based detector on five MVTec-AD categories shows it resting its scores largely on the defect pixels, as it should, but with a non-negligible share coming from irrelevant background pixels, which a joint pixel-frequency analysis traces mostly to high frequencies. When the image resizing step is switched from nearest-neighbour to antialiased resizing, the detector's F1 falls from 0.92 to 0.80 and its false negative rate rises from 4 to 26 percent. The authors call the result surprising because image quality improved; the high frequencies the detector partly depended on were removed, and the anomaly scores fell with them.

Liznerski et al. (2021) show the same effect in the setting where a one-class detector is trained with horses as the anomalous class and ImageNet as nominal: the explanation heatmaps highlight the watermarks found in the lower-left corner of about a fifth of PASCAL VOC horse images, and bars and fences, rather than horses. Ming, Yin and Li (2022) formalise the broader phenomenon for out-of-distribution detection, reporting that detection performance deteriorates severely as the correlation between spurious environmental features and labels grows in the training set.

4.3 Covariate change flagged as novelty

A score that responds to low-level statistics will flag inputs whose content is familiar but whose rendering is not. The full-spectrum evaluation in OpenOOD v1.5 measures exactly this: when covariate-shifted in-distribution images (corrupted or re-sampled versions of known classes) are counted as in-distribution, the near-OOD AUROC of most detectors on ImageNet-1K drops by more than 10 points, and the method with the smallest drop, a two-level semantic hierarchy, still falls from 72.85 to 66.17 (Zhang, J. et al., 2023). The authors describe the behaviour as not aligned with human perception: a person sorting dogs from cars would not call a covariate-shifted dog novel. Ahmed and Courville (2020) had argued earlier that the out-distributions of practical interest are semantic and context-dependent,

Table 1. The three claims an anomaly explanation can make, what each is relative to, what can check it, and examples of published checks. This decomposition is the paper's own; the rows name the sources whose checks illustrate each column.

Claim	Relative to	Referee	Published checks (examples)
Score	Detector and its reference	The detector, under intervention	Model-randomisation (Adebayo et al., 2018); reconstruction-replacement (Elrashid et al., 2026); counterfactual normality (Liznerski et al., 2026)
Deviation	Reference data	Separability from reference	Subspace separability (Micenkova et al., 2013); rules on held-out part (Jacob et al., 2021)
Name	A vocabulary	A label, description or person	Top-1 category accuracy (Wu et al., 2024); MCQ accuracy (Jiang, Li et al., 2025); semantic IoU (Conti et al., 2023)
Correspondence	All three	All three, same instances	None located

and Vaze et al. (2022) built a benchmark to isolate semantic novelty from other kinds of shift.

For the present argument, this is where the gap between "different" and "what" is widest. A flagged covariate-shifted dog is different; what it is, is a dog. A score explanation would point, correctly, to the corruption. A deviation explanation would say, correctly, that the texture statistics differ from the reference. Only a name claim, "this is a dog", would tell the user that nothing new has arrived, and the detector has no way to make it.

4.4 The reliability of score explanations themselves

Score explanations are not automatically faithful to the score. Adebayo et al. (2018) show that some saliency methods produce maps independent of both the model and the data-generating process, and conclude that such methods are unfit for tasks like finding outliers in data. Adebayo et al. (2022) test whether post hoc explanations can reveal a model's reliance on a spurious signal the user does not already suspect. Feature attributions and concept activation scores detected visible spurious signals (a text tag, a stripe pattern) when the user knew to look for them, were less effective for background blur, and could not conclusively detect reliance on a signal the user did not know about. Feature attributions also raised false alarms on models that did not rely on the spurious signal.

For anomaly detection there is an additional trap. The standard way to evaluate a feature-importance explanation is to remove the important features and watch the score change; Hase et al. (2021) argue that the inputs produced by removal are themselves out of distribution, which contaminates both the explanation and the metric. For an anomaly detector, whose whole function is to score out-of-distribution inputs highly, a removal-based faithfulness test is circular: the deletion that is meant to remove the anomaly adds a new one. Elrashid et al. (2026) address this for audio by replacing highlighted regions with the autoencoder's own reconstructions, so that the intervention moves the input toward normal rather than away from it, and report that a masked autoencoder gave more faithful and temporally precise explanations than a standard one at slightly lower detection performance.

4.5 What this section establishes

A score explanation is a claim about the detector. Its value is diagnostic: it shows what the detector is using, and the most cited demonstrations in the anomaly-explanation literature are demonstrations that the detector was using something other than the novelty. The premise is right that such an explanation does not say what the novelty is. It is wrong to present this as a shortfall of attribution. A score explanation that named the novelty instead of the background in Ren et al. (2019) would have been unfaithful to the detector.

5. Deviation explanations stay in the reference's vocabulary

5.1 Contrastive characterisation

A second family of methods describes how a flagged input differs from normal data. Micenkova et al. (2013) search for the feature subspaces in which a given outlier is separable from inliers, and present them as explanations that work with any detector. Macha and Akoglu (2018) note that anomaly description has a much sparser literature than anomaly detection and propose x-PACS, which groups known anomalies into patterns and finds, for each, a characterising subspace and feature rules that separate it from normal instances. Gupta et al. (2017) choose a small number of two-feature scatter plots that best explain why a detector flagged nodes in a time-evolving graph. Liu et al. (2018) combine an outlierness score, the attributes that contribute to it, and a description of the neighbourhoods the outlier is contrasted with. Antwarg et al. (2019) adapt Shapley values to autoencoder anomalies, separating features that raise the reconstruction error from those that offset it. Kauffmann, Mueller and Montavon (2020) decompose one-class support vector machine outputs into input-variable contributions. A 2026 detector for tabular data makes the explanation part of the scoring, flagging each point by the directions in which its projection escapes an extreme-value baseline, so that the flagging directions are the explanation (Diop, 2026).

Counterfactual methods state the deviation as the smallest change that would make the input normal. Sulem et al. (2022) generate diverse sets of perturbed time series that a detector would not flag. Liznerski et al. (2026) train generative models to produce several

alternative modifications of an anomalous image, each of which the detector perceives as normal, and describe the result as a semantic explanation of what triggered the detector; their evaluation covers more than 80 anomaly-detection setups and measures whether the counterfactuals look normal to the detector, how realistic they are, and whether they disentangle distinct concepts.

5.2 What deviation explanations can and cannot say

Every method in this section describes the flagged input in the features of the reference, or in images that a person must read. A subspace rule says that x has an unusual value of feature 7 given features 3 and 12. A counterfactual image says that x would be normal if this region looked like that. Neither says what x is. Liznerski et al. (2026) use "semantic" for explanations that change object-level content rather than pixel noise, which is a real advance over heatmaps, but the name of the change still has to be supplied by the person looking at the pair of images.

The time-series benchmark of Jacob et al. (2021) makes the limitation measurable. Its six anomaly types were injected deliberately into Spark cluster traces, and ground-truth root-cause intervals are provided. Yet the benchmark's "accuracy" for an explanation is defined by re-detection: the logical predicates an explanation method derives from part of an anomaly are evaluated for anomaly detection on the remaining part and nearby normal data. LIME, which returns feature importances rather than predicates, could not be scored for accuracy at all. The benchmark could have asked whether an explanation identified the anomaly's type; the metric it defined asks whether the explanation separates the anomaly from normal data, which is a deviation criterion. That choice is natural, and it shows how far the deviation framing extends into evaluation.

5.3 People supply the name

When deviation explanations are handed to people who already hold the relevant vocabulary, they help. Schemmer et al. (2023) ran a behavioural experiment in which participants had to decide whether anomalies flagged by an autoencoder in New York taxi data were extreme weather events or other events such as holidays and sporting events. With counterfactual explanations, mean accuracy was 70.24 percent against 55.11 percent without them (22 control and 21 treatment participants after exclusions; Mann-Whitney $u = 137.5$, $p = 0.021$); time per decision did not change significantly. The categories were given by the task. The explanation helped people choose between names they already had.

This is the configuration Sommer and Paxson (2010) describe for intrusion detection and Bendale and Boulton (2015) build into open-world recognition: the detector says different, an explanation says how, and a person says what. The evidence in this section supports it as a

working arrangement. It does not show that the naming step can be removed.

6. Naming: where the vocabulary comes from, and how often the name is right

A name claim needs a vocabulary. The methods that supply names differ mainly in where the vocabulary comes from: a fixed concept set, the text space of a vision-language model, or free-form generation by a multimodal language model. This section takes them in turn and collects what has been measured.

6.1 Fixed concept sets

Concept-based explanations describe a model's behaviour in terms of human concepts: TCAV measures the sensitivity of a prediction to a user-defined concept (Kim et al., 2018), and concept bottleneck models predict concepts supplied at training time and then predict the label from them (Koh et al., 2020). A concept bottleneck design has been extended to visual anomaly detection, with supervision regimes ranging from fully supervised concepts to synthetic anomalies only, and paired with a pixel-level branch for localisation (Stropeni et al., 2025).

For novelty, the difficulty is that the concept set is fixed before the novel input arrives. Choi et al. (2023) give the most careful treatment of concept explanations for out-of-distribution detectors. They define detection completeness, the extent to which a detector's decisions can be reproduced from concept scores, and learn concepts that raise it: on Animals with Attributes as in-distribution data and SUN as the out-of-distribution set, concepts from an earlier method gave a detection completeness of 0.782 and their learned concepts gave 0.961. The concepts are learned from the classifier's internal-layer representations without additional human annotation, and are interpreted by visualising them; for the predicted class Collie, the key concepts were read as "furry dog skin" and "oval dog face". In that example the vocabulary of the explanation is the vocabulary of the in-distribution class: it describes what a Collie looks like, against which the flagged input is judged. The authors note that a human-subject evaluation was not performed. Adebayo et al. (2022) supply the complementary negative: concept activation methods detected reliance on known spurious signals such as a tag, but when the signal was not among the concepts tested, reliance could not be detected in the concepts that were available.

6.2 Text space of a vision-language model

Contrastive vision-language models make an open vocabulary available at test time, which appears to dissolve the problem. It does not, for two measured reasons.

The first is that naming from a large open vocabulary is much harder than naming from the right small one. Conti et al. (2023) formalise vocabulary-free image classification and propose CaSED, which retrieves

captions from an external database and picks the best-matching candidate category. Averaged over ten datasets, CaSED reached a cluster accuracy of 43.1 percent, a semantic similarity of 50.4 (on a 0 to 100 scale) and a semantic word-overlap (IoU) of 17.7 percent, against 69.0, 82.0 and 69.4 for their upper bound, CLIP given the ground-truth class names (their Tables 1 to 3); they also report that simply running CLIP over a very large predefined vocabulary was not effective. Han et al. (2023) report a roughly 50 percent improvement over their baseline on ImageNet, in the unsupervised setting, when a vision-language model must assign names from an essentially unconstrained vocabulary, which shows progress without closing the gap. These measurements are on ordinary benchmark images, not novel ones; they measure how much of the naming ability of a vision-language model depends on being handed the right names.

The second is that novel inputs are named as the nearest familiar thing. Li, F. et al. (2025) name the failure "categorization confusion" in open-vocabulary video anomaly detection: novel anomalies are often misclassified as visually similar base categories. Section 6.4 gives the numbers.

6.3 Generative description by multimodal models

The newest systems describe anomalies in free text. AnomalyGPT fine-tunes a large vision-language model on simulated anomalous images paired with textual descriptions, and reports 86.1 percent accuracy on MVTec-AD with one normal example (Gu et al., 2024). Holmes-VAD builds a video instruction-tuning benchmark, VAD-Instruct50k, with a semi-automatic pipeline in which a captioner and a language model write the analyses (Zhang, H. et al., 2024). The CUVA benchmark annotates video anomalies with what happened, why, and how severe it was, and finds that most vision-language models describe anomalies well but perform poorly on causal analysis (Du et al., 2024). Vad-R1 trains explicit reasoning chains for video anomaly reasoning (Huang et al., 2025). For time series, AXIS conditions a frozen language model on numeric, encoder-derived and task-prior hints and evaluates explanations with language-model and human judges (Lan et al., 2025).

MMAD gives the most informative measurement for the present question, because it scores defect-related and object-related questions on the same images (Jiang, Li et al., 2025). Its 39,672 multiple-choice questions over 8,366 industrial images cover seven subtasks. For GPT-4o in the standard one-shot setting (their Table 2), accuracy was 68.63 percent on anomaly discrimination, 65.80 on defect classification, 55.62 on defect localisation, 73.21 on defect description, 83.41 on defect analysis, 94.98 on object classification and 82.80 on object analysis, for an average of 74.92. In a preliminary human evaluation on 177 sampled questions, three

industrial anomaly-detection researchers scored 95.24 on discrimination, 75.00 on defect classification, 92.31 on localisation and 86.11 on object classification. Random chance is 50 percent for discrimination and 25 for the others. AnomalyGPT, trained for discrimination in a fixed format, performed worst overall.

Two features of these numbers matter. The model named the object far more reliably than it classified the defect, which is the novelty-relevant part of the image. And the experts' own defect classification, 75.00, was their weakest score, which suggests that naming a defect type is hard for people as well; the ground truth for names is not free even when a person supplies it.

A further caution comes from using language models to enrich detectors. Dai et al. (2023) find that descriptions generated by a language model for in-distribution classes include hallucinated features, their example being a featherless head attributed to a hen, and that using such descriptions indiscriminately degraded out-of-distribution detection in four of five in-distribution datasets, which they attribute to the hallucinated descriptors ("We hypothesize this is because LLMs generate unfaithful descriptors due to hallucinations"); they then filtered descriptions by a consistency check.

6.4 Open-vocabulary anomaly detection: the cleanest comparison

Wu et al. (2024) define open-vocabulary video anomaly detection: a model is trained on normal videos and anomalies from base categories, and must at test time both detect anomalies and assign each anomalous video to a category from a set that includes novel categories. The names of the novel categories are part of the test vocabulary, and the model is fine-tuned on pseudo-anomalies synthesised from potential novel category names. The design separates class-agnostic detection from class-specific categorisation and reports each for base and novel categories, so detection and naming of the same novel inputs, by the same model, can be read side by side. Li, F. et al. (2025) adopt the same task and report the same measurements, which gives a second side-by-side comparison.

On UCF-Crime, the full model's frame-level detection AUC was 93.80 for base categories and 88.20 for novel ones (their Table 1). Its video-level top-1 categorisation accuracy was 49.02 for base and 37.08 for novel (their Table 7). On XD-Violence, detection AP was 57.10 for base and 76.03 for novel (their Table 2), while categorisation accuracy was 89.31 for base and 30.90 for novel. Without the synthesis module, novel-category accuracy on XD-Violence was 15.18. The model of Li, F. et al. (2025), Anomize, adds label-relation encodings designed to reduce categorisation confusion. On XD-Violence its detection AP was 57.37 for base and 84.24 for novel categories (their Table 1), and its top-1 accuracy was 92.37 for base and 87.43 for novel (their Table 2), a naming gap of about five points. On UCF-

Crime its detection AUC was 93.00 for base and 87.05 for novel, and its top-1 accuracy 56.86 for base and 41.57 for novel. UCF-Crime has thirteen anomaly types, so a uniform guess among them would be right about one time in thirteen; XD-Violence has six. Figure 1 shows the comparison.

Detection AUC, detection AP and top-1 accuracy are different metrics on different units (frames against videos), and they are not placed on one scale here. What the numbers support is a within-metric comparison. For Wu et al. (2024), on XD-Violence, novelty made anomalies easier to detect and much harder to name, and on both datasets moving from base to novel categories cost far more in naming accuracy than in the corresponding detection score, although the correct names were available. For Anomize the picture splits by dataset: on UCF-Crime the move to novel categories again cost more in naming (about 15 points) than in detection (about 6), while on XD-Violence novel anomalies were detected better than base ones and named almost as well. The qualitative observation that novel anomalies are pulled toward similar base categories is theirs, and they name a second failure for novel inputs alongside it: "detection ambiguity, where the model struggles to assign accurate anomaly scores to unfamiliar anomalies" (Li, F. et al., 2025).

6.5 What this section establishes

In the studies of Sections 6.2 and 6.4 and in MMAD, a name scored against ground truth is right much less often than the corresponding detection or recognition task in the same study, with one exception. The shortfall appears for fixed vocabularies with the novel names included (Wu et al., 2024, on both datasets; Li, F. et al., 2025, on UCF-Crime), for industrial defects against object identity (Jiang, Li et al., 2025), and for open vocabularies without the names (Conti et al., 2023). The exception is Anomize on XD-Violence, which named novel categories almost as often as base ones (Li, F. et al., 2025). CUVA, scoring free-text answers with other methods, concludes that "most VLMs excel in the description of anomalies" (Du et al., 2024), so the shortfall is not universal. The premise that a detector's own signal cannot supply names is borne out: every name in this section came from a vocabulary imported from outside the detector, and its accuracy depended on the import.

7. Detectors whose scores are made of names

7.1 The positive-signal family

The premise that detectors emit a purely negative signal does not hold for a large and growing family. Maximum Concept Matching scores an input by its best softmax-scaled similarity to the in-distribution class names, and gained 13.1 AUROC points over a purely visual baseline on a hard task with semantically similar classes (Ming, Cai et al., 2022). NegLabel draws a large number of

negative labels from a text corpus and scores inputs by their affinity to them (Jiang, Liu et al., 2024). EOE prompts a language model to envision outlier class labels based on visual similarity and penalises matches to them (Cao et al., 2024). ZOC trains a caption generator on top of CLIP to produce candidate names for the unknown for each test image, and scores the image by the accumulated probability of those candidates against the known names (Esmaeilpour et al., 2022). WinCLIP scores industrial anomalies by comparing image windows with a compositional ensemble of prompts built from state words and templates, reaching 91.8 percent zero-shot classification AUROC on MVTec-AD (Jeong et al., 2023). LAVAD goes furthest: it captions every video frame, asks a language model to score the anomaly from the captions, and so computes the score from a description (Zanella et al., 2024).

For these detectors the signal is positive in form. Some name comes attached to the score, and in LAVAD the score is derived from the name. A person reading a MCM score can see which in-distribution name the input matched best; a person reading a ZOC score can see which candidate unknown names carried it.

7.2 Why a name inside a score does not certify the name

Three observations limit what this buys.

First, the names inside the score serve the score. ZOC's score sums over candidate unknown names, and the authors' own illustration of the method is a boat image flagged correctly with "boat" absent from its candidates. Their error study shows the names working in both directions for images of espresso, which is one of the seen classes: for some, the generator proposes "coffee" rather than "espresso" and the score still correctly treats the image as known, while for another, "The set of generated labels and the label 'coffee' produce a high $S(x)$ ", so the generated names themselves produced a wrong detection (Esmaeilpour et al., 2022). The detection result, which is what their benchmarks score, does not require the names to be correct, so good detection numbers are not evidence about naming. NegLabel and EOE have the same structure: their labels are chosen because they help separate outliers from in-distribution inputs, not because they name the outliers.

Second, at least one strong zero-shot anomaly design has moved toward content-free words. Zhou et al. (2024) report that off-the-shelf vision-language models focus on the class semantics of foreground objects rather than on abnormality, and obtain better zero-shot detection across 17 datasets by learning object-agnostic prompts for generic normality and abnormality. The text side of the detector becomes a learned version of "abnormal". This is a positive signal in form and a negative one in content.

Third, where a score is derived from a description, as in LAVAD, the description is the thing that has to be

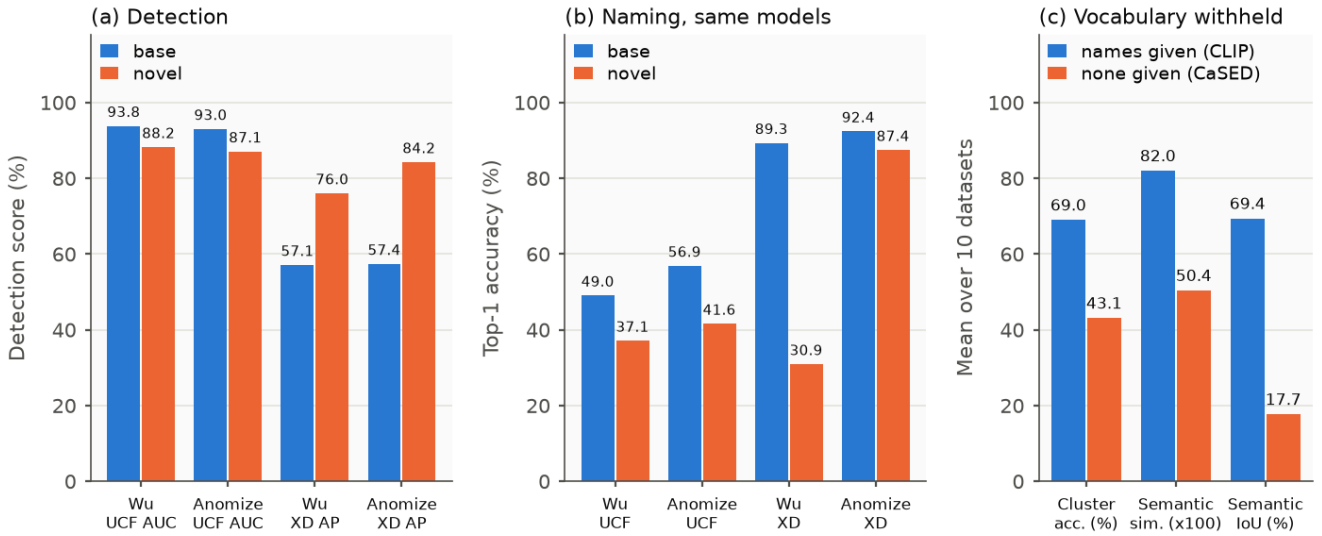


Figure 1. Detecting novelty is not naming it. (a) Detection scores of the open-vocabulary video anomaly models of Wu et al. (2024) and Li, F. et al. (2025, Anomize) for base and novel anomaly categories: frame-level AUC on UCF-Crime and frame-level AP on XD-Violence (Wu et al., Tables 1 and 2; Li, F. et al., Table 1). (b) Video-level top-1 categorisation accuracy of the same models on the same test sets (Wu et al., Table 7; Li, F. et al., Table 2), with the novel category names included in the test vocabulary. (c) Ten-dataset mean naming scores reported by Conti et al. (2023, Tables 1-3) for CLIP given the ground-truth class names and for CaSED given no vocabulary, on ordinary benchmark images; cluster accuracy and semantic IoU are percentages, semantic similarity is multiplied by 100. Each panel carries one metric and re-plots values reported by the cited sources.

checked, and its errors propagate to the score. Zanella et al. (2024) devote a component to cleaning noisy captions using cross-modal similarity, which is an acknowledgement that the captions are not reliably right. No located study reports, for LAVAD or for systems like it, how often the caption behind a high score correctly describes the anomalous event.

7.3 What this section establishes

The premise survives in a narrowed form. The signal from a vision-language detector can be positive in form, and the names involved are visible. But the detector is validated only as a detector, on whether its score separates normal from anomalous inputs, and that validation is compatible with wrong names. For these detectors, the name claim still needs a separate referee. The advantage they offer is that the name and the score share a representation, which makes the intervention test in Section 8 cheap to run. No located study runs it.

8. What would tie a name to the deviation

8.1 Four checks

A name claim for a flagged input can be checked in four ways, which test different things.

Correctness. Is the name right? This needs an external referee: a ground-truth label, a reference description, or a person. Sections 6.3 and 6.4 report such measurements.

Separation. Does the named property distinguish flagged inputs from the reference? This can be tested on

data: the fraction of flagged inputs with the property should exceed its fraction in the reference.

Score-faithfulness. If the named property is removed or suppressed, does the score fall? This ties the name to the detector's response. For a detector that computes its score from text, the test can be run in the text space, by removing the name from the candidate set or by projecting the concept out of the embedding; for others, it needs a generative edit, as in the counterfactuals of Liznerski et al. (2026), or a concept projection, as in the completeness measure of Choi et al. (2023).

Completeness. Does the set of named properties account for the whole score, or only part of it? Choi et al. (2023) define this for concept sets.

The correspondence claim of Section 2 needs correctness and score-faithfulness on the same instances, and completeness to rule out that the score was mainly raised by something unnamed.

8.2 Separation: the set-level literature

The most developed machinery for verifying that a name separates two groups is not in anomaly detection but in the literature on describing differences between datasets. Zhong et al. (2022) search for a natural-language description that is more often true of one text distribution than another: a language model proposes descriptions, and a trained verifier re-ranks them by checking how often each holds on a larger sample. On 54 real-world binary classification tasks with human-written reference descriptions, GPT-3 Curie produced a description similar to the human one 7 percent of the

time, fine-tuning and re-ranking raised this to 61 percent, and their best system, using GPT-3 Davinci, reached 76 percent. Zhong et al. (2023) extend the task to goal-driven discovery over 675 open-ended problems and propose four evaluation criteria (validity, relevance, novelty and significance).

For images, VisDiff captions two image sets, prompts a language model to propose differences, and ranks each proposal by the AUROC with which CLIP similarity to the proposal separates the two sets, discarding proposals that fail a significance test (Dunlap et al., 2024). On a benchmark of 150 paired image sets at three difficulty levels plus 37 pairs derived from ImageNet-R and ImageNet-star, the final system's top-1 score (a GPT-4 judgement of match to the ground-truth difference, validated against four annotators at a Pearson correlation of 0.80 on 200 proposals) was 0.88 on easy pairs, 0.75 on medium, 0.61 on hard, and 0.78 on the ImageNet-derived pairs (their Table 2). GSCLIP proposes a selector to identify which language explanations properly summarise a dataset shift (Zhu, Liang and Zou, 2022). B2T extracts keywords from captions of misclassified images and validates each by its similarity to those images under a vision-language scorer (Kim, Y. et al., 2024). DrML diagnoses classifier failures by probing a shared embedding space with language (Zhang, Y. et al., 2023).

Slice discovery gives a direct measurement of naming accuracy for automatically found groups. Domino identifies underperforming, coherent slices and names them by ranking phrases from a text corpus in a cross-modal embedding (Eyuboglu et al., 2022). Across 1,235 slice discovery settings it recovered 36 percent of the planted slices, and for natural images it ranked the exact slice name, or a WordNet synonym, first for 34.7 percent of rare-subclass slices, 41.0 percent of correlation slices and 39.0 percent of noisy-label slices, and in the top ten for 57.4, 55.4 and 48.7 percent respectively. At the level of individual neurons, MILAN searches for descriptions that maximise mutual information with the image regions where a neuron is active and obtains higher agreement with human descriptions than methods restricted to a fixed concept list (Hernandez et al., 2022).

8.3 What the set-level checks do and do not provide

These methods verify separation, and in their benchmarks they also measure correctness. That is more than anomaly naming has. Three limits prevent them from settling the correspondence question.

They operate on sets. VisDiff's proposer sees twenty images from each side and its PairedImageSets hold a hundred images per side; Domino's slices are groups; Zhong et al. (2022) re-rank on samples. A single flagged input cannot be separated from anything by frequency. Set-level description could name a cluster of flagged inputs; it does not name one.

Their verifier often shares the proposer's representation. VisDiff ranks with CLIP; B2T validates with a vision-language scorer; Dai et al. (2023) measure the consistency of language-model descriptions by the overlap of the images CLIP retrieves for them. A name that the shared embedding confuses with the true property will pass verification for the same reason it was proposed. The correctness numbers in these papers come from comparison with ground-truth descriptions, not from the verifier.

They verify that the name separates the groups, not that it is what a detector responded to. VisDiff's ranker is not the detector that flagged anything. Choi et al. (2023) is the one located method that ties a concept set to a specific detector's decisions, through detection completeness, and its concepts are learned from the classifier's representations and interpreted by people; it measures score-faithfulness without an external test of names, and its authors leave a human-subject evaluation outside that paper.

8.4 The missing measurement

Putting the pieces together: correctness of names has been measured for anomaly and novelty naming (Section 6). Score-faithfulness of explanations has been measured for pixels, features and concepts (Sections 4 and 6.1). Separation of names has been measured for sets (Section 8.2). No located study measures correctness and score-faithfulness of a name on the same flagged inputs. The closest is the open-vocabulary setting of Wu et al. (2024), which measures detection and correct naming on the same inputs, but not whether the correct name is what the detector's score was responding to; a video could be flagged for its camera motion and correctly categorised for its content. Until that joint measurement exists, the claim that a system explains the novelty it detects is not supported by the evidence, whichever way it would come out.

9. The numbers in one place

Table 2 gathers the measurements on which the argument depends, one published result per row and the source named in the first cell. Which of the three claims of Section 3 a row informs is given in the final column; that mapping is the author's interpretation, not the sources'.

Four features of the table deserve comment. The score rows show the size of the gap between what raises a score and what a person would call novel, measured in each case by explaining the score. The single deviation row shows people using explanations to choose between names they were given. In the name rows that report a detection or recognition figure beside the name (Wu et al., Li, F. et al., Jiang, Li et al. and Conti et al.), names are right substantially less often than the comparison task in the same study, except for Li, F. et al. on XD-Violence; the slice, text and image-set rows report

Table 2. Published measurements bearing on the score, deviation and name claims. All values are as reported by the named source and are consolidated here for comparison. AUC/AP = area under the ROC / precision-recall curve; MCQ = multiple-choice question accuracy; top-1 = top-ranked name matches ground truth.

Source	Setting	Quantity	Reported value	Claim
Ren et al. (2019), Table 1	PixelCNN + +, Fashion-MNIST vs MNIST	AUROC, likelihood vs ratio	0.089 vs 0.994; corr. with background 0.85	Score
Kauffmann, Dippel et al. (2025)	D2Neighbors, 5 MVTec-AD classes, resizing changed	F1; false negative rate	0.92 to 0.80; 4% to 26%	Score
Zhang, J. et al. (2023)	ImageNet-1K, full-spectrum	Near-OOD AUROC	Most methods down >10; MOS 72.85 to 66.17	Score
Schemmer et al. (2023)	Taxi anomalies, user study	Category accuracy	55.11% without vs 70.24% with explanations	Deviation
Choi et al. (2023)	AwA vs SUN	Detection completeness	0.782 prior concepts vs 0.961 learned	Score
Wu et al. (2024), T1, T7	UCF-Crime	AUC; top-1	Base 93.80, novel 88.20; base 49.02, novel 37.08	Name
Wu et al. (2024), T2, T7	XD-Violence	AP; top-1	Base 57.10, novel 76.03; base 89.31, novel 30.90	Name
Li, F. et al. (2025), T1, T2	UCF-Crime	AUC; top-1	Base 93.00, novel 87.05; base 56.86, novel 41.57	Name
Li, F. et al. (2025), T1, T2	XD-Violence	AP; top-1	Base 57.37, novel 84.24; base 92.37, novel 87.43	Name
Jiang, Li et al. (2025), T2	MMAD, GPT-4o	MCQ	Object class 94.98; defect class 65.80; defect loc. 55.62	Name
Jiang, Li et al. (2025), T2	MMAD, 3 experts	MCQ	Object class 86.11; defect class 75.00; defect loc. 92.31	Name
Conti et al. (2023), T1-3	10 datasets, no vocabulary	Cluster acc.; sem. IoU	43.1; 17.7 (69.0; 69.4 with names)	Name
Eyuboglu et al. (2022)	Planted slices, natural images	Top-1 name	34.7 / 41.0 / 39.0% (rare / correlation / noisy)	Name
Zhong et al. (2022)	54 text tasks	Match to human description	7% / 61% / 76% (Curie / tuned + reranked / Davinci)	Name
Dunlap et al. (2024), T2	VisDiffBench	Acc@1 (GPT-4 judged)	0.88 easy, 0.75 medium, 0.61 hard	Name

naming accuracy alone. And there is no correspondence row.

10. The decision structure

Algorithm 1 writes down what the argument asks of a system that must both flag inputs and say what they are. It is the author's synthesis of the evidence of Sections 4 to 8. Two features carry the argument: a name is never reported without the referee that checked it, and there is an explicit output for names no referee could check.

Three properties follow from the earlier sections. Step 1 checks the score claim with the detector as referee, and uses reconstruction-style interventions because deletion creates new anomalies (Hase et al., 2021; Elrashid et al., 2026). Step 3 is the score-faithfulness test of Section 8.1; for vision-language detectors it amounts to re-scoring with a candidate name removed, which is cheap. Step 4 separates what the system can check from what it cannot, so that a correct-looking name produced by a generative model is not reported with the same standing as a name confirmed by a label. Step 5 makes the covariate case explicit: when the score is carried by rendering rather than content, the most useful name may be the familiar one.

11. What would settle the open part

The open question is whether, for current detectors paired with current naming methods, the name given to a flagged input names what raised its score. The following studies would answer it with existing data.

A correspondence benchmark on planted anomalies.

Industrial and synthetic benchmarks already contain anomalies whose type is known: MVTec-AD's defect categories (Bergmann et al., 2019), the structural and logical anomalies of MVTec LOCO (Bergmann et al., 2022), and synthesised anomalies of named categories (Wu et al., 2024). For each flagged instance, report detection, the top-1 name from a fixed and from an open vocabulary, and the fall in score when the named property is suppressed. The prediction of this paper is that correct names and score-faithful names will overlap only partly, and that the overlap will be lowest where Section 4 found scores driven by texture or background.

The covariate control. Repeat the benchmark on covariate-shifted versions of normal inputs, as in full-spectrum evaluation (Zhang, J. et al., 2023). A naming step that works should return the familiar class name for these; the rate at which it instead invents a novel name is a direct measure of how far naming inherits the detector's confusion.

Logical anomalies. For anomalies that are violations of arrangement or count rather than local damage

Algorithm 1. Reporting a flagged input: score, deviation, name (conceptual synthesis by the author)

Input:	input x with score $s(x) > \text{threshold}$; detector D with reference R ; vocabulary V (label set, concept bank, text corpus or generator) with stated provenance; external referee E (labels, descriptions, person) or none
Output:	report with fields SCORE, DEVIATION, NAME, each tagged with the referee that checked it
Step 1:	SCORE. Attribute $s(x)$ to parts of x . Check the attribution by an intervention that moves x toward R (e.g. replace parts with reconstructions), not by deletion. If the attributed parts are background, texture or rendering, flag POSSIBLE-COVARIATE. # Sec 4
Step 2:	DEVIATION. State the respect in which x departs from R (subspace, rule or counterfactual). Check it on held-out reference data. # Sec 5
Step 3:	NAME. Propose candidates $T_{1..T_k}$ from V . For each: (a) score-faithfulness: suppress T_i in x (text-space removal, concept projection or generative edit); record the drop in $s(x)$; (b) keep T_i only if the drop is material and the remaining score is below threshold. # Sec 7, 8
Step 4:	CORRECTNESS. If E is available, check the kept names against E and report the result. If E is not available, report NAME as UNVERIFIED-BY-REFEREE. # Sec 3, 6
Step 5:	If no T_i survives Step 3, report NAME as NONE, with SCORE and DEVIATION; if POSSIBLE-COVARIATE was set, also report the nearest reference class name. # Sec 4.3
Step 6:	Log (x , report, referees). When E later labels x , score the log for correctness and correspondence.

(Bergmann et al., 2022), score explanations that highlight a region are ill-posed, since the defect may be an absence. Measure whether naming methods describe such anomalies correctly more or less often than local ones. If naming helps anywhere that localisation cannot, it should be here.

Names for vision-language detectors. For MCM, NegLabel, EOE, ZOC and WinCLIP-style detectors, the score-faithfulness test is a re-scoring. Report, for flagged inputs with ground-truth labels, how often the name that carries most of the score is correct. ZOC's own examples suggest the answer may be low without the detection rate being affected.

The human baseline. MMAD's expert scores (Jiang, Li et al., 2025) show that people are not perfect at naming defects. Any correspondence benchmark should report inter-annotator agreement on names, so that a model's shortfall is measured against the achievable ceiling, not against a ground truth of unknown quality.

What would count against this paper. A study showing that, on flagged inputs, names from an automatic system are correct at rates close to detection AUC and that suppressing the named property removes most of the score would show that the gap in Section 6 is a transitional shortfall of current models and not a structural feature of the task. The decomposition would survive; the practical conclusion that names need an external referee would weaken, because the referee could be sampled rather than consulted for every case.

12. The strongest case against this paper

The models are improving. MMAD's ordinary human participants outperformed the best model by about four points on average, and experts by about twelve (Jiang, Li et al., 2025). If multimodal models close that gap, the naming shortfall in Table 2 becomes a historical note. This paper accepts that the numbers will move. Its claim is not that names will stay wrong but that their correctness is checked by a different referee from the detector's, and that the correspondence between a name and a score is not measured. Better models would make the missing measurement more worth running, not less.

Set-level description already solves the problem. VisDiff, Domino and the distribution-difference methods verify names and score well against ground truth on their benchmarks. For clusters of flagged inputs, they may be the right tool. But anomaly detection often has one flagged input, not a set; their verifiers typically share the proposer's embedding; and they verify separation, not that a particular detector responded to the named property. They solve a neighbouring problem, and the machinery could be adapted. It has not been adapted in any study located here.

Naming is not the detector's job. One could argue that a detector should detect and leave interpretation to people, as open-world recognition does (Bendale and Boulton, 2015). That is a defensible design, and Section 5.3 shows it working. But it is not what systems that attach generated descriptions to detections present themselves as doing, and Sommer and Paxson (2010) argued that operators will not be served by a detector that leaves the interpretive gap unbridged. The paper's point is only that when names are offered, they should be reported with

the referee that checked them.

The decomposition is definitional. It is partly. The observation that a score is relative to a reference and a name is relative to a vocabulary follows from the definitions. What is not definitional is the evidence: that score explanations have repeatedly revealed scores driven by non-novel features, that names are measured to lag detection by large margins in most settings even with the correct names supplied, and that no study has measured correspondence. A definition that organises those findings and predicts where the next measurement should fail is doing work.

13. What is not known

It is not known how often the correct name for a flagged input also names what raised its score. No located study measures it.

It is not known whether the naming shortfall in open-vocabulary anomaly detection comes mainly from the vocabulary, from the visual representation, or from the pull of trained categories. Wu et al. (2024) and Li, F. et al. (2025) report improvements from synthesis and label-relation encodings without isolating the cause.

It is not known how accurate generated anomaly descriptions are, instance by instance, in the video and time-series systems of Section 6.3. Their evaluations use language-model judges, human preference or multiple-choice formats, and none located reports the rate at which a description names the event that the detector flagged.

It is not known what agreement people reach when naming novel anomalies. MMAD's three experts give one data point; no located study reports inter-annotator agreement on names for novel anomaly types.

It is not known whether score-faithfulness tests in text space, which are cheap for vision-language detectors, agree with generative or concept-projection tests for the same name.

14. Limitations

The paper's evidence for the naming gap comes mainly from a small number of studies: two open-vocabulary video papers on one task supply the only side-by-side measurements of detection and naming on the same novel inputs, and they disagree on one of its two datasets, one industrial benchmark supplies the defect-versus-object contrast, and one vocabulary-free study supplies the with-and-without-vocabulary contrast, on ordinary rather than novel images. The consistency of direction across them is the argument's support; any one of them could be unrepresentative of its field.

The metrics compared within studies differ in kind. Frame-level AUC and video-level top-1 accuracy measure different things on different units, and the paper has limited its claims to comparisons within a metric. A

reader who wants a single number for "the naming gap" will not find one here, because no source defines one.

The literature search centred on computer vision, with smaller coverage of tabular, time-series and network data. The score-driven failures of Section 4 are documented mostly for images; the extent to which tabular detectors show the same pattern was not established by the sources read.

The algorithm and the proposed studies are stated as designs for evaluation. The score-faithfulness test in Step 3 of Algorithm 1 assumes that a named property can be suppressed without introducing a new anomaly; for generative edits that assumption is itself open, as the out-of-distribution problem in feature removal shows (Hase et al., 2021).

Finally, the three-claim decomposition assigns referees in the idealised case. In practice the reference data may be contaminated, the detector may be retrained on flagged inputs, and people who supply names disagree. The decomposition says which kind of check each claim needs; it does not say that the checks are easy.

15. Conclusion

A detector reports that an input is different from a reference. An explanation of its score says what in the input produced that report, and is checked by the detector. An explanation of the deviation says how the input differs from the reference, and is checked by the reference data. A name says what the input is, and is checked only by something outside both. The literature on explaining anomalies has built strong tools for the first two and has shown, with them, that detectors often respond to backgrounds, textures and rendering rather than to the novelty a person would recognise. Detectors that compute scores from names have made the signal positive in form, without making detection results evidence about the names. Where naming has been scored beside detection, it has mostly lagged by a wide margin, even with the right names supplied; the best published open-vocabulary result shows a large naming gap on UCF-Crime and a small one on XD-Violence. Whether the name given to a flagged input describes what the detector responded to has not been measured. Until it is, a system that attaches a name to a detection should report which referee checked the name, and one that cannot should say so.

Acknowledgements

The author thanks the maintainers of the arXiv, Crossref and OpenAlex APIs, through which every citation in this paper was verified.

References

- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. & Kim, B. (2018). Sanity Checks for Saliency Maps. *NeurIPS* 2018. arXiv:1810.03292. doi:10.48550/arXiv.1810.03292.

- Adebayo, J., Muelly, M., Abelson, H. & Kim, B. (2022). Post hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation. ICLR 2022. arXiv:2212.04629. doi:10.48550/arXiv.2212.04629.
- Ahmed, F. & Courville, A. (2020). Detecting semantic anomalies. AAAI 2020. arXiv:1908.04388. doi:10.48550/arXiv.1908.04388.
- Antwarg, L., Miller, R. M., Shapira, B. & Rokach, L. (2019). Explaining Anomalies Detected by Autoencoders Using SHAP. arXiv preprint. arXiv:1903.02407. doi:10.48550/arXiv.1903.02407.
- Bendale, A. & Boulton, T. (2015). Towards Open World Recognition. CVPR 2015. arXiv:1412.5687. doi:10.48550/arXiv.1412.5687.
- Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D. & Steger, C. (2022). Beyond Dents and Scratches: Logical Constraints in Unsupervised Anomaly Detection and Localization. International Journal of Computer Vision. doi:10.1007/s11263-022-01578-9.
- Bergmann, P., Fauser, M., Sattlegger, D. & Steger, C. (2019). MVTec AD - A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. doi:10.1109/cvpr.2019.00982.
- Boulton, T. E., Grabowicz, P. A., Prijatelj, D. S., Stern, R., Holder, L. B., Alspector, J., Jafarzadeh, M., Ahmad, T., Dhamija, A. R., Li, C., Cruz, S., Shrivastava, A., Vondrick, C. & Scheirer, W. (2021). Towards a Unifying Framework for Formal Theories of Novelty. Proceedings of the AAAI Conference on Artificial Intelligence 35(17). doi:10.1609/aaai.v35i17.17766.
- Cao, C., Zhong, Z., Zhou, Z., Liu, Y., Liu, T. & Han, B. (2024). Envisioning Outlier Exposure by Large Language Models for Out-of-Distribution Detection. ICML 2024. arXiv:2406.00806. doi:10.48550/arXiv.2406.00806.
- Chadwick, T., Chao, J., Izumigawa, C., Galdorisi, G., Ortiz-Pena, H., Loup, E., Soultanian, N., Manzanares, M., Mai, A., Yen, R. & Lange, D. S. (2023). Characterizing Novelty in the Military Domain. arXiv preprint. arXiv:2302.12314. doi:10.48550/arXiv.2302.12314.
- Chandola, V., Banerjee, A. & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys 41(3). doi:10.1145/1541880.1541882.
- Choi, J., Raghuram, J., Feng, R., Chen, J., Jha, S. & Prakash, A. (2023). Concept-based Explanations for Out-Of-Distribution Detectors. ICML 2023. arXiv:2203.02586. doi:10.48550/arXiv.2203.02586.
- Conti, A., Fini, E., Mancini, M., Rota, P., Wang, Y. & Ricci, E. (2023). Vocabulary-free Image Classification. NeurIPS 2023. arXiv:2306.00917. doi:10.48550/arXiv.2306.00917.
- Dai, Y., Lang, H., Zeng, K., Huang, F. & Li, Y. (2023). Exploring Large Language Models for Multi-Modal Out-of-Distribution Detection. Findings of EMNLP 2023. arXiv:2310.08027. doi:10.48550/arXiv.2310.08027.
- Diop, L. (2026). Witnesses Explain Anomalies. ICDM 2026 (accepted). arXiv:2609.03826. doi:10.48550/arXiv.2609.03826.
- Du, H., Zhang, S., Xie, B., Nan, G., Zhang, J., Xu, J., Liu, H., Leng, S., Liu, J., Fan, H., Huang, D., Feng, J., Chen, L., Zhang, C., Li, X., Zhang, H., Chen, J., Cui, Q. & Tao, X. (2024). Uncovering What, Why and How: A Comprehensive Benchmark for Causation Understanding of Video Anomaly. CVPR 2024. arXiv:2405.00181. doi:10.48550/arXiv.2405.00181.
- Dunlap, L., Zhang, Y., Wang, X., Zhong, R., Darrell, T., Steinhardt, J., Gonzalez, J. E. & Yeung-Levy, S. (2024). Describing Differences in Image Sets with Natural Language. CVPR 2024. arXiv:2312.02974. doi:10.48550/arXiv.2312.02974.
- Ehrashid, M., Deschenes, A., Subakan, C., Ravanelli, M., Georges, R. & Morin, M. (2026). Toward Faithful Explanations in Acoustic Anomaly Detection. ICASSP 2026. arXiv:2601.12660. doi:10.48550/arXiv.2601.12660.
- Esmailpour, S., Liu, B., Robertson, E. & Shu, L. (2022). Zero-Shot Out-of-Distribution Detection Based on the Pre-trained Model CLIP. AAAI 2022. arXiv:2109.02748. doi:10.48550/arXiv.2109.02748.
- Eyuboglu, S., Varma, M., Saab, K., Delbrouck, J.-B., Lee-Messer, C., Dunnmon, J., Zou, J. & Re, C. (2022). Domino: Discovering Systematic Errors with Cross-Modal Embeddings. ICLR 2022. arXiv:2203.14960. doi:10.48550/arXiv.2203.14960.
- Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M. & Wang, J. (2024). AnomalyGPT: Detecting Industrial Anomalies Using Large Vision-Language Models. AAAI 2024. arXiv:2308.15366. doi:10.48550/arXiv.2308.15366.
- Guerin, J., Delmas, K., Ferreira, R. S. & Guiochet, J. (2023). Out-Of-Distribution Detection Is Not All You Need. AAAI 2023. arXiv:2211.16158. doi:10.48550/arXiv.2211.16158.
- Gupta, N., Eswaran, D., Shah, N., Akoglu, L. & Faloutsos, C. (2017). LookOut on Time-Evolving Graphs: Succinctly Explaining Anomalies from Any Detector. arXiv preprint. arXiv:1710.05333. doi:10.48550/arXiv.1710.05333.
- Han, K., Huang, X., Li, Y., Vaze, S., Li, J. & Jia, X. (2023). What's in a Name? Beyond Class Indices for Image Recognition. arXiv preprint (CVPR 2024 Workshop on Computer Vision in the Wild). arXiv:2304.02364. doi:10.48550/arXiv.2304.02364.
- Hase, P., Xie, H. & Bansal, M. (2021). The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations. NeurIPS 2021. arXiv:2106.00786. doi:10.48550/arXiv.2106.00786.
- Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A. & Andreas, J. (2022). Natural Language Descriptions of Deep Visual Features. ICLR 2022. arXiv:2201.11114. doi:10.48550/arXiv.2201.11114.
- Huang, C., Wang, B., Wen, J., Liu, C., Wang, W., Shen, L. & Cao, X. (2025). Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought. arXiv preprint. arXiv:2505.19877. doi:10.48550/arXiv.2505.19877.
- Jacob, V., Song, F., Stiegler, A., Rad, B., Diao, Y. & Tatbul, N. (2021). Exathlon: A Benchmark for Explainable Anomaly Detection over Time Series. arXiv preprint. arXiv:2010.05073. doi:10.48550/arXiv.2010.05073.
- Jafarzadeh, M., Dhamija, A. R., Cruz, S., Li, C., Ahmad, T. & Boulton, T. E. (2020). A Review of Open-World Learning and Steps Toward Open-World Learning Without Labels. arXiv preprint. arXiv:2011.12906. doi:10.48550/arXiv.2011.12906.
- Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A. & Dabeer, O. (2023). WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation. CVPR 2023. arXiv:2303.14814. doi:10.48550/arXiv.2303.14814.
- Jiang, X., Li, J., Deng, H., Liu, Y., Gao, B.-B., Zhou, Y., Li, J., Wang, C. & Zheng, F. (2025). MMAD: A Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection. ICLR 2025. arXiv:2410.09453. doi:10.48550/arXiv.2410.09453.
- Jiang, X., Liu, F., Fang, Z., Chen, H., Liu, T., Zheng, F. & Han, B. (2024). Negative Label Guided OOD Detection with Pretrained Vision-Language Models. ICLR 2024. arXiv:2403.20078. doi:10.48550/arXiv.2403.20078.
- Kauffmann, J., Dippel, J., Ruff, L., Samek, W., Mueller, K.-R. & Montavon, G. (2025). The Clever Hans Effect in Unsupervised Learning. Nature Machine Intelligence. arXiv:2408.08041. doi:10.48550/arXiv.2408.08041.
- Kauffmann, J., Mueller, K.-R. & Montavon, G. (2020). Towards Explaining Anomalies: A Deep Taylor Decomposition of One-Class Models. Pattern Recognition. arXiv:1805.06230. doi:10.48550/arXiv.1805.06230.
- Kauffmann, J., Ruff, L., Montavon, G. & Mueller, K.-R. (2020). The Clever Hans Effect in Anomaly Detection. arXiv preprint. arXiv:2006.10609. doi:10.48550/arXiv.2006.10609.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. & Sayres, R. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). ICML 2018. arXiv:1711.11279. doi:10.48550/arXiv.1711.11279.
- Kim, Y., Mo, S., Kim, M., Lee, K., Lee, J. & Shin, J. (2024). Discovering and Mitigating Visual Biases through Keyword

- Explanation. CVPR 2024. arXiv:2301.11104. doi:10.48550/arXiv.2301.11104.
- Kirichenko, P., Izmailov, P. & Wilson, A. G. (2020). Why Normalizing Flows Fail to Detect Out-of-Distribution Data. arXiv preprint. arXiv:2006.08545. doi:10.48550/arXiv.2006.08545.
- Koh, P. W., Nguyen, T., Tang, Y. S., Musmann, S., Pierson, E., Kim, B. & Liang, P. (2020). Concept Bottleneck Models. ICML 2020. arXiv:2007.04612. doi:10.48550/arXiv.2007.04612.
- Lan, T., Le, H. D., Li, J., He, W., Wang, M., Liu, C. & Zhang, C. (2025). AXIS: Explainable Time Series Anomaly Detection with Large Language Models. arXiv preprint. arXiv:2509.24378. doi:10.48550/arXiv.2509.24378.
- Li, F., Liu, W., Chen, J., Zhang, R., Wang, Y., Zhong, X. & Wang, Z. (2025). Anomize: Better Open Vocabulary Video Anomaly Detection. arXiv preprint. arXiv:2503.18094. doi:10.48550/arXiv.2503.18094.
- Li, Z., Zhu, Y. & van Leeuwen, M. (2023). A Survey on Explainable Anomaly Detection. ACM Transactions on Knowledge Discovery from Data. arXiv:2210.06959. doi:10.48550/arXiv.2210.06959.
- Liu, N., Shin, D. & Hu, X. (2018). Contextual Outlier Interpretation. arXiv preprint. arXiv:1711.10589. doi:10.48550/arXiv.1711.10589.
- Liznerski, P., Ruff, L., Vandermeulen, R. A., Franks, B. J., Kloft, M. & Mueller, K.-R. (2021). Explainable Deep One-Class Classification. ICLR 2021. arXiv:2007.01760. doi:10.48550/arXiv.2007.01760.
- Liznerski, P., Varshneya, S., Calikus, E., Wang, P., Bartscher, A., Vollmer, S. J., Fellenz, S. & Kloft, M. (2026). Reimagining Anomalies: What If Anomalies Were Normal? AAAI 2026. arXiv:2402.14469. doi:10.48550/arXiv.2402.14469.
- Macha, M. & Akoglu, L. (2018). Explaining Anomalies in Groups with Characterizing Subspace Rules. arXiv preprint. arXiv:1708.05929. doi:10.48550/arXiv.1708.05929.
- Micenkova, B., Ng, R. T., Dang, X.-H. & Assent, I. (2013). Explaining Outliers by Subspace Separability. 2013 IEEE 13th International Conference on Data Mining. doi:10.1109/icdm.2013.132.
- Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W. & Li, Y. (2022). Delving into Out-of-Distribution Detection with Vision-Language Representations. NeurIPS 2022. arXiv:2211.13445. doi:10.48550/arXiv.2211.13445.
- Ming, Y., Yin, H. & Li, Y. (2022). On the Impact of Spurious Correlation for Out-of-distribution Detection. AAAI 2022. arXiv:2109.05642. doi:10.48550/arXiv.2109.05642.
- Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D. & Lakshminarayanan, B. (2019). Do Deep Generative Models Know What They Don't Know? ICLR 2019. arXiv:1810.09136. doi:10.48550/arXiv.1810.09136.
- Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V. & Lakshminarayanan, B. (2019). Likelihood Ratios for Out-of-Distribution Detection. NeurIPS 2019. arXiv:1906.02845. doi:10.48550/arXiv.1906.02845.
- Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T. & Gehler, P. (2022). Towards Total Recall in Industrial Anomaly Detection. CVPR 2022. arXiv:2106.08265. doi:10.48550/arXiv.2106.08265.
- Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G. & Mueller, K.-R. (2021). A Unifying Review of Deep and Shallow Anomaly Detection. Proceedings of the IEEE. arXiv:2009.11732. doi:10.48550/arXiv.2009.11732.
- Schemmer, M., Holstein, J., Bauer, N., Kuehl, N. & Satzger, G. (2023). Towards Meaningful Anomaly Detection: The Effect of Counterfactual Explanations on the Investigation of Anomalies in Multivariate Time Series. arXiv preprint. arXiv:2302.03302. doi:10.48550/arXiv.2302.03302.
- Serra, J., Alvarez, D., Gomez, V., Slizovskaia, O., Nunez, J. F. & Luque, J. (2020). Input complexity and out-of-distribution detection with likelihood-based generative models. ICLR 2020. arXiv:1909.11480. doi:10.48550/arXiv.1909.11480.
- Sommer, R. & Paxson, V. (2010). Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. 2010 IEEE Symposium on Security and Privacy. doi:10.1109/sp.2010.25.
- Stropeni, A., Zaccaria, V., Borsatti, F., Dalle Pezze, D., Barusco, M. & Susto, G. A. (2025). Explainable Visual Anomaly Detection via Concept Bottleneck Models. arXiv preprint. arXiv:2511.20088. doi:10.48550/arXiv.2511.20088.
- Sulem, D., Donini, M., Zafar, M. B., Aubet, F.-X., Gasthaus, J., Januschowski, T., Das, S., Kenthapadi, K. & Archambeau, C. (2022). Diverse Counterfactual Explanations for Anomaly Detection in Time Series. arXiv preprint. arXiv:2203.11103. doi:10.48550/arXiv.2203.11103.
- Vaze, S., Han, K., Vedaldi, A. & Zisserman, A. (2022). Open-Set Recognition: a Good Closed-Set Classifier is All You Need? ICLR 2022. arXiv:2110.06207. doi:10.48550/arXiv.2110.06207.
- Wang, Y., Guo, D., Li, S., Camps, O. & Fu, Y. (2023). Unveiling the Unseen: A Comprehensive Survey on Explainable Anomaly Detection in Images and Videos. arXiv preprint. arXiv:2302.06670. doi:10.48550/arXiv.2302.06670.
- Wu, P., Zhou, X., Pang, G., Sun, Y., Liu, J., Wang, P. & Zhang, Y. (2024). Open-Vocabulary Video Anomaly Detection. CVPR 2024. arXiv:2311.07042. doi:10.48550/arXiv.2311.07042.
- Yang, J., Zhou, K., Li, Y. & Liu, Z. (2021). Generalized Out-of-Distribution Detection: A Survey. arXiv preprint. arXiv:2110.11334. doi:10.48550/arXiv.2110.11334.
- Zanella, L., Menapace, W., Mancini, M., Wang, Y. & Ricci, E. (2024). Harnessing Large Language Models for Training-free Video Anomaly Detection. CVPR 2024. arXiv:2404.01014. doi:10.48550/arXiv.2404.01014.
- Zhang, H., Xu, X., Wang, X., Zuo, J., Han, C., Huang, X., Gao, C., Wang, Y. & Sang, N. (2024). Holmes-VAD: Towards Unbiased and Explainable Video Anomaly Detection via Multi-modal LLM. arXiv preprint. arXiv:2406.12235. doi:10.48550/arXiv.2406.12235.
- Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Li, Y., Liu, Z., Chen, Y. & Li, H. (2023). OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection. Journal of Data-centric Machine Learning Research. arXiv:2306.09301. doi:10.48550/arXiv.2306.09301.
- Zhang, Y., HaoChen, J. Z., Huang, S.-C., Wang, K.-C., Zou, J. & Yeung, S. (2023). Diagnosing and Rectifying Vision Models using Language. ICLR 2023. arXiv:2302.04269. doi:10.48550/arXiv.2302.04269.
- Zhong, R., Snell, C., Klein, D. & Steinhardt, J. (2022). Describing Differences between Text Distributions with Natural Language. ICML 2022. arXiv:2201.12323. doi:10.48550/arXiv.2201.12323.
- Zhong, R., Zhang, P., Li, S., Ahn, J., Klein, D. & Steinhardt, J. (2023). Goal Driven Discovery of Distributional Differences via Language Descriptions. arXiv preprint. arXiv:2302.14233. doi:10.48550/arXiv.2302.14233.
- Zhou, Q., Pang, G., Tian, Y., He, S. & Chen, J. (2024). AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection. ICLR 2024. arXiv:2310.18961. doi:10.48550/arXiv.2310.18961.
- Zhu, Z., Liang, W. & Zou, J. (2022). GSCLIP: A Framework for Explaining Distribution Shifts in Natural Language. ICML 2022 DataPerf workshop. arXiv:2206.15007. doi:10.48550/arXiv.2206.15007.