WHERE WILL IT FAIL? DETECTOR-NATIVE CONCEPT BLINDSPOTS FOR FORECASTING ERRORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Object detection underpins numerous safety-critical applications such as autonomous driving and embodied AI, yet even the best detectors still make perception errors that pose serious risks. Understanding detector failure modes is therefore important for diagnosing model vulnerabilities, assessing risk, and guiding improvements. A detector's intrinsic concepts may play a fundamental and general role in determining its errors, but explicitly extracting and quantifying such concept-level vulnerabilities remains challenging. We introduce a framework to discover, validate, and leverage *detector-native concept blindspots* for failure diagnosis and risk screening. We develop DetSAE, our task-relevant sparse autoencoder, to extract sparse concepts from a fixed detector's object-level representations while preserving its classification and localization behavior. We then generate 20 style-transferred versions of the original dataset and track objects to identify blindspots associated with statistically supported excess detection degradation. Experiments demonstrate that suppressing excessive blindspot activations improves frozen-head classification across detector backbones, providing causal evidence of their contribution to errors. Using the same frozen blindspot representation, our separate false-negative and false-positive risk predictors achieve the best performance among compared methods across diverse benchmarks. These results show that seemingly isolated detection failures can share concept-level weaknesses that support both model interpretation and actionable risk assessment. Our code and generated datasets are publicly available on GitHub.

1 INTRODUCTION

Object detection supports numerous safety-critical applications, including autonomous driving and embodied intelligence. Extensive research has sought to improve the accuracy and robustness of object detectors (Ren et al., 2015; Zhang et al., 2022; Danish et al., 2024). Yet even state-of-the-art detectors still make perception errors, particularly on objects from unseen domains or under severe occlusion (Danish et al., 2024; Liu et al., 2024; Reddy et al., 2022). As deployment expands into safety-critical applications, such failures can pose serious risks. Accurately identifying and analyzing the causes of detector failures are therefore important for diagnosing detector vulnerabilities, assessing deployment risks, and guiding model improvement.

Existing approaches estimate uncertainty or detect distributional novelty to identify potentially unreliable inputs (Oksuz et al., 2023; Sastry & Oore, 2020; Sun et al., 2022). More direct failure-prediction methods learn to anticipate performance degradation from detector representations, sometimes augmented with external semantic knowledge (Rahman et al., 2021; Yatbaz et al., 2024; Keser & Knoll, 2026; Nakamura et al., 2024; Zimmermann et al., 2026). These methods estimate when a detector is likely to fail, but generally map reliability cues or latent features directly to failure risk, without identifying the detector's underlying concept-level vulnerabilities. Systematic-error discovery takes a further step by identifying groups of inputs on which a model consistently underperforms; for example, Eyuboglu et al. (2022) discover coherent data slices sharing human-interpretable attributes. Such approaches reveal what kinds of inputs tend to fail, but leave open whether these recurring failures can be traced to systematic weaknesses in the detector's own concept representations. We investigate this possibility by focusing on intrinsic vulnerabilities in how a detector represents visual concepts, which we call *detector-native concept blindspots*.

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.

Concept blindspots have been studied in generative and classification tasks. Early work revealed systematic semantic omissions in generative models (Bau et al., 2019), while Bohacek et al. (2026) identify persistent concept omissions and overrepresentations using sparse representations. In classification, Plumb et al. (2022) characterize low-performing semantic subgroups as blindspots, while Konforti et al. (2026) explain systematic failures through concept-prediction errors in concept bottleneck models. Applying this perspective to object detection is non-trivial: conventional detectors do not explicitly output concepts, while aggregate metrics obscure their underlying concept blindspots. Therefore, the key challenges are to explicitly extract and quantify detector-native concept blindspots, and to see whether they can improve failure prediction and risk assessment.

We therefore propose a framework for discovering, validating, and leveraging detector-native concept blindspots. We diagnose perception failures by extracting model concepts from learned object-level representations and consolidating them into concept clusters. Under appearance changes, we identify clusters whose associated objects exhibit statistically supported excess detection degradation, and use them to investigate the systematic failures. Grounded in the detector's own learned representations, our framework provides a further explanation of its model-specific failure vulnerabilities. Our main contributions are as follows:

- **Concept extraction and intervention for object detection.** We introduce DetSAE, which decomposes object-level features into sparse concepts while preserving the detector's classification and localization behavior, and we intervene on blindspot activations to validate their causal contribution to classification errors.
- **Detector-native concept blindspot analysis.** Our analysis makes concept-related weaknesses obscured by aggregate performance identifiable, quantifiable, and comparable, providing a fine-grained and testable account of systematic detection failures.
- **Demonstrating the value of concept blindspots for object detection.** The same discovered blindspots support both false-negative and false-positive risk ranking, achieving the best performance among the compared methods and extending detector failure attribution to actionable risk assessment.

2 Related Work

Concept-based interpretation. Concept methods expose semantic structure in learned representations. TCAV measures sensitivity to predefined concepts (Kim et al., 2018); ACE and CRAFT discover concepts through clustering or factorization (Ghorbani et al., 2019; Fel et al., 2023); Concept Bottleneck Models use supervised concepts as intermediate variables (Koh et al., 2020). Sparse autoencoders extract sparse latent features (Gao et al., 2025a), while output-preserving dictionary learning improves their functional relevance (Braun et al., 2024). DetSAE instead learns object-level sparse concepts while preserving detector classification and localization behavior.

Systematic failures and concept blindspots. Systematic-error methods organize failures into semantic groups, including Domino (Eyuboglu et al., 2022), classifier blindspot discovery (Plumb et al., 2022), HiBug2 (Chen et al., 2025), and grounded instance-level failure hypotheses (Zhang et al., 2025). Concept-level failures have also been studied in generative models (Bau et al., 2019; Bohacek et al., 2026) and Concept Bottleneck Models (Konforti et al., 2026). We instead identify blindspots from a detector's own object-level concepts using statistically supported excess detection degradation, and validate their causal contribution through intervention.

Failure prediction and detector introspection. Detection risk has been estimated from uncertainty and calibration (Oksuz et al., 2023), internal activations and frame-level monitors (Rahman et al., 2021; Yatbaz et al., 2024; Keser & Knoll, 2026), vision-language or spatial uncertainty (Lafon et al., 2025; Guarino et al., 2026), detector–foundation-model disagreement (Zimmermann et al., 2026), and OOD scores (Sastry & Oore, 2020; Sun et al., 2022; Wang et al., 2022). These methods infer risk from predictive or feature-level signals. We first identify and validate detector-native concept blindspots, then use the same vulnerability information for both false-negative and false-positive risk screening.

3 METHOD

Given a labeled reference dataset $\mathcal{D} = \{(x_n, Y_n)\}$ and a frozen object detector F, we seek detector-native concepts that expose systematic model vulnerabilities. As illustrated in Figure 1, our framework extracts task-relevant concepts, identifies concept blindspots across controlled style changes, and uses blindspot-aware concept representations to train detection-error risk predictors. We further validate the identified blindspots through causal intervention in Section 3.3.

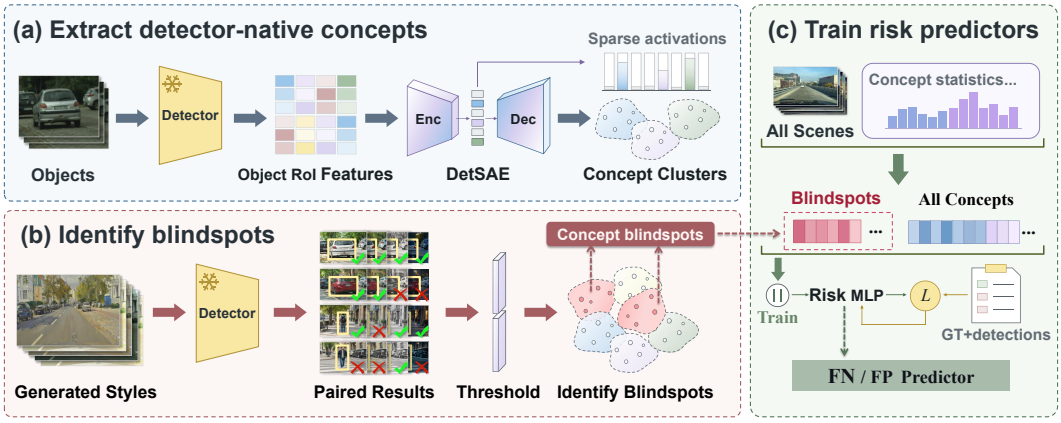

Figure 1: **Framework.** (a) DetSAE extracts sparse concepts from object-level detector representations and consolidates them into concept clusters. (b) The same objects are evaluated across generated styles to identify concept blindspots associated with excess detection degradation. (c) All-concept statistics together with explicit blindspot channels are used to train separate FN and FP risk predictors under detector-error supervision.

3.1 DETECTOR-NATIVE CONCEPT EXTRACTION

For object i of class c, let $h_i \in \mathbb{R}^d$ denote its RoI feature. Using reference statistics (μ, σ), a class-specific DetSAE normalizes $\widetilde{h}_i = (h_i - \mu)/\sigma$ and computes

$$a_i = \mathrm{TopK}_k\Big(\mathrm{ReLU}(W_e\widetilde{h}_i + b_e)\Big), \qquad \widehat{h}_i = \sigma \odot Da_i + \mu, \tag{1}$$

where W_e, b_e are encoder parameters, $a_i \in \mathbb{R}^K$ is the sparse concept-activation vector, and $D \in \mathbb{R}^{d\times K}$ contains unit-norm concept directions; TopK_k retains the k largest activations.

Feature reconstruction alone need not preserve the detector behavior being interpreted. With frozen class-probability head p and class-specific box-regression head b_c, DetSAE minimizes

$$\mathcal{L}_{\mathrm{DetSAE}} = \mathrm{MSE}(Da_i, \widetilde{h}_i) + \lambda_{\mathrm{cls}} D_{\mathrm{KL}}\Big(p(h_i)\|p(\widehat{h}_i)\Big) + \lambda_{\mathrm{box}}\,\mathrm{SL1}\Big(b_c(h_i), b_c(\widehat{h}_i)\Big) + \lambda_{\mathrm{aux}}\mathcal{L}_{\mathrm{aux}}. \tag{2}$$

Here $\lambda_{\mathrm{cls}}, \lambda_{\mathrm{box}}, \lambda_{\mathrm{aux}}$ weight the classification, localization, and auxiliary objectives. The auxiliary loss $\mathcal{L}_{\mathrm{aux}} = \mathrm{MSE}(Da_i^{\mathrm{aux}}, \mathrm{sg}(\widetilde{h}_i - Da_i))$ uses the top-4 ReLU responses among directions inactive in the current batch to reconstruct the detached residual, encouraging underused directions to remain trainable; sg denotes stop-gradient (Gao et al., 2025a).

Directions with sufficient held-out support, frozen-head sensitivity, and stability across independently trained DetSAEs are merged within each class into concept clusters G_g using activation, top-object, dictionary, and feature similarities. Their activation shares are

$$s_{ig} = \frac{\sum_{j\in G_g} a_{ij}}{\sum_j a_{ij}}. \tag{3}$$

Clusters with $s_{ig} \geq 0.05$ are ordered by share, and the largest are marked $m_{ig} = 1$ until at least half of the eligible activation mass is covered. The resulting membership m_i is defined on the original image and fixed across style variants.

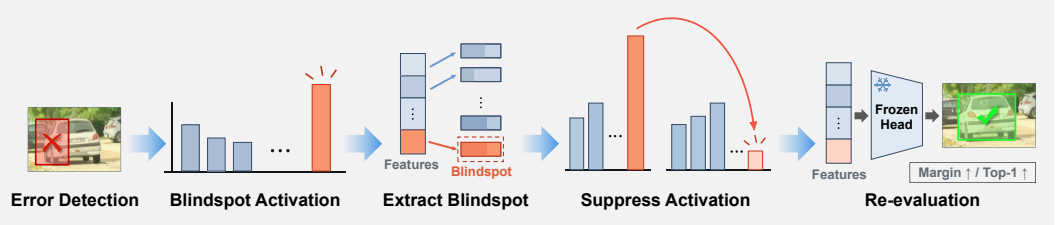

Figure 2: **Causal evidence by blindspot intervention.** Excessive blindspot activations are restored to their original-image levels and re-evaluated by the frozen detection head.

3.2 Detector-Native Concept Blindspot Identification

We generate $S = 20$ curated style variants of each reference image using Seedream models (Gao et al., 2025b; Seedream et al., 2025; ByteDance Seed Team, 2026). Every variant was manually reviewed against its original image to verify object count, shape, position, and annotation consistency, allowing the same objects to be tracked across styles. The specific generator is not essential: our framework only requires paired style-transferred images preserving object content and annotations, and the same blindspot-identification procedure applies to any generator satisfying these requirements.

For the same object i under style s, detection quality is

$$Q_i^{(s)} = \begin{cases} p_i^{(s)} \ \mathrm{IoU}_i^{(s)}, & \text{one-to-one correct-class match with IoU} \geq 0.5, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $s = 0$ denotes the original image and $p_i^{(s)}$ is the matched detection confidence.

Let $\Delta_{is} = Q_i^{(0)} - Q_i^{(s)}$ denote the detection degradation of object i from the original image to style s. Low performance alone does not imply a blindspot: an object may already be difficult, or a style may degrade an entire class. We therefore remove the class–style mean degradation $\overline{\Delta}_{c_i,s}$:

$$y_i = \frac{1}{|\mathcal{S}_i|} \sum_{s \in \mathcal{S}_i} \left(\Delta_{is} - \overline{\Delta}_{c_i,s} \right), \tag{5}$$

where $\mathcal{S}_i$ contains the valid styles for object i, and $\overline{\Delta}_{c,s} = |\mathcal{I}_{c,s}|^{-1} \sum_{i \in \mathcal{I}_{c,s}} \Delta_{is}$, with $\mathcal{I}_{c,s} = \{i : c_i = c,\ s \in \mathcal{S}_i\}$. For each class, we model excess degradation as $y_i = \alpha + \phi_i^\top \theta + m_i^\top \beta + \epsilon_i$ and jointly estimate the coefficients by

$$(\widehat{\alpha}, \widehat{\theta}, \widehat{\beta}) = \arg\min_{\alpha,\theta,\beta} \frac{1}{n_c} \sum_{i:c_i=c} \left(y_i - \alpha - \phi_i^\top \theta - m_i^\top \beta \right)^2 + \lambda \|\beta\|_2^2, \tag{6}$$

where α is the intercept, θ and β are coefficients for the covariates ϕ_i and concept-membership vector m_i, respectively, and $(\widehat{\alpha}, \widehat{\theta}, \widehat{\beta})$ denote their fitted estimates. Here n_c is the number of evaluable class-c objects, ϕ_i contains original quality and geometry (area, aspect ratio, position, border distance, and squared geometry terms), and λ is selected by scene-grouped cross-validation. Since y_i measures excess degradation, $\widehat{\beta}_g$ is the fitted additional degradation associated with membership in cluster g, after controlling for ϕ_i and co-occurring concepts.

We assess each cluster using scene-level bootstrap confidence intervals, Benjamini–Hochberg-corrected one-sided tests for $\beta_g > \delta$, and leave-one-style-out refits to verify stability across styles. We define the set of concept blindspots as

$$\mathcal{B} = \left\{ g : \widehat{\beta}_g > \delta,\ L_g > \delta,\ q_g \leq 0.05,\ \min_s \widehat{\beta}_g^{(-s)} > \delta \right\}, \qquad \delta = 0.05, \tag{7}$$

where L_g is the lower bound of the 95% bootstrap interval, q_g is the adjusted p-value for $H_0 : \beta_g \leq \delta$, and $\widehat{\beta}_g^{(-s)}$ is the leave-one-style-out estimate. Clusters outside $\mathcal{B}$ are treated as normal.

3.3 Causal Validation by Concept Intervention

We further test whether excessive blindspot activations causally contribute to classification errors, as illustrated in Figure 2. For a ground-truth RoI with $m_{ig} = 1$ for blindspot cluster g, define

$$u_{ig}^{(s)} = \sum_{j \in G_g} D_j a_{ij}^{(s)}, \qquad r_{ig}^{(s)} = \frac{\left\langle u_{ig}^{(s)}, u_{ig}^{(0)} \right\rangle}{\max\left(\|u_{ig}^{(0)}\|_2^2, 10^{-10} \right)}. \tag{8}$$

When $r_{ig}^{(s)} \geq 1.5$, activations in G_g are reset to their original-image values: $a'_{ij} = a_{ij}^{(0)}$ for $j \in G_g$ and $a'_{ij} = a_{ij}^{(s)}$ otherwise. Thus, each selected activation changes by $a_{ij}^{(0)} - a_{ij}^{(s)}$, with no fixed attenuation factor. Let $e_i^{(s)} = \widetilde{h}_i^{(s)} - D a_i^{(s)}$ denote the residual under style s. The intervened feature is

$$h_i^{\text{int}} = \sigma \odot \left(D a'_i + e_i^{(s)} \right) + \mu. \tag{9}$$

The residual, other activations, and frozen head remain unchanged. We evaluate true-class margin change and net Top-1 recovery, with matched-magnitude random, opposite-direction, and normal-concept interventions as controls.

3.4 Blindspot-Aware Risk Predictor Training

For image x, let $\mathcal{P}_c(x)$ denote the retained proposals for class c. For proposal $r \in \mathcal{P}_{c_g}(x)$ and concept cluster g of class c_g, we define

$$t_{rg} = \frac{\sum_{j \in G_g} a_j(h_r)}{\rho_g}, \qquad t_{rg}^{\text{w}} = p(c_g \mid h_r) t_{rg}, \tag{10}$$

where ρ_g is the median positive cluster activation on original reference objects and $p(c_g \mid h_r)$ is the detector confidence. Let t_g and t_g^{w} collect these values over $\mathcal{P}_{c_g}(x)$. With M and T_3 denoting maximum and top-three mean,

$$v_g(x) = \left[M(t_g), T_3(t_g), M(t_g^{\text{w}}), T_3(t_g^{\text{w}}) \right], \qquad u(x) = [v(x); v_{\mathcal{B}}(x)]. \tag{11}$$

Here $v(x)$ concatenates all concept clusters, including blindspots, while $v_{\mathcal{B}}(x)$ repeats the blindspot subset as dedicated channels.

Keeping the detector, DetSAE, and $\mathcal{B}$ frozen, we train separate FN and FP predictors under two training variants: one uses only original reference images and the other one additionally uses the 20 style variants; all other components remain identical. For image n, class c, and error type $\tau \in \{\text{FN}, \text{FP}\}$, let $N_{n,c}^{\tau}$ be the error count, $b_{n,c}^{\tau} = \mathbf{1}[N_{n,c}^{\tau} > 0]$, and $\ell_{n,c}^{\tau} = \log(1 + N_{n,c}^{\tau})$, with FN/FP labels defined in Section 5.1. Each MLP predicts occurrence and log-count. The FN predictor additionally predicts $\nu_n = \log(1 + U_n)$, where $U_n = \sum_{i \in Y_n} (1 - Q_i)$ measures the aggregate detection-quality deficit in the current image. We optimize

$$\mathcal{L}_{\text{risk}}^{\tau} = \text{BCE}_w(\widehat{b}^{\tau}, b^{\tau}) + \text{SL1}(\widehat{\ell}^{\tau}, \ell^{\tau}) + 0.2\, \mathcal{L}_{\text{rank}}^{\tau} + 0.5\, \mathbf{1}[\tau = \text{FN}]\, \text{SL1}(\widehat{\nu}, \nu), \tag{12}$$

where BCE_w is positive-class-weighted binary cross-entropy, SL1 is Smooth L1, and 0.2 and 0.5 are fixed loss weights. For within-batch pairs with $N_n^{\tau} \neq N_m^{\tau}$,

$$\mathcal{L}_{\text{rank}}^{\tau} = \mathbb{E}_{(n,m)} \log\left(1 + \exp\left[-\text{sgn}(N_n^{\tau} - N_m^{\tau})(\widehat{\ell}_n^{\tau} - \widehat{\ell}_m^{\tau}) \right] \right), \tag{13}$$

which encourages images with more errors to receive higher predicted risk.

At inference, classwise counts are recovered as $[\exp(\widehat{\ell}_{n,c}^{\tau}) - 1]_+$ and summed over Car and Person. For each error type, predictions from three independently trained MLPs are averaged to obtain the final image-level risk score.

4 Validating Detector-Native Concept Blindspots

We validate the framework in three stages: whether DetSAE preserves the detector behavior it represents, whether objects activating blindspots are systematically harder to detect, and whether intervening on blindspot activations causally changes frozen-head classification errors.

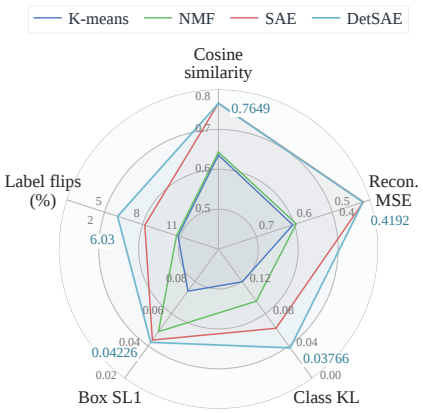

Figure 3: **Task fidelity after reconstruction.** Object-weighted held-out means over three seeds; outward is better.

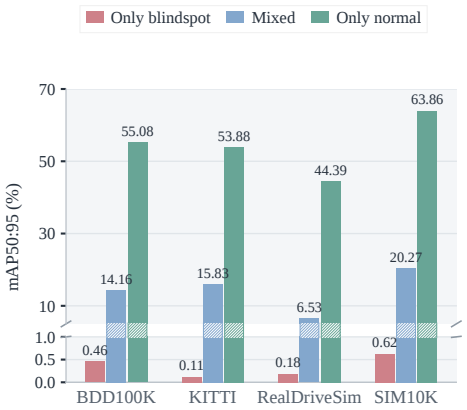

Figure 4: **Detection difficulty by active concept composition.** Group-conditional $\text{mAP}_{50:95}$ (%) on external datasets.

4.1 Experimental Setup

We use a common experimental setup for the validation below and the risk-screening experiments in Section 5. Our main detector is Faster R-CNN (Ren et al., 2015) with a ResNet-101-FPN (R101) backbone trained on Cityscapes (Cordts et al., 2016) (2,975 training and 500 scene-disjoint calibration images). Evaluation covers three benchmark groups: KITTI (Geiger et al., 2012) and 1,764 daytime-clear BDD100K images (Yu et al., 2020) for camera changes; SIM10K (Johnson-Roberson et al., 2016) and RealDriveSim (Jadon et al., 2025) for simulation; and Rainy-Cityscapes (Hu et al., 2019) at $\alpha = 0.02$ and Foggy-Cityscapes (Sakaridis et al., 2018) at $\beta = 0.02$ for weather changes. For RealDriveSim, we use a fixed 6,000-image subset: 2,520 randomly sampled Day images and all 2,580 Adverse-A and 900 Adverse-B images. Unless noted otherwise, evaluation covers Car and Person; analyses without sufficient Person support use Car only.

DetSAE uses $K = 64$, $k = 8$, 2,500 updates, batch size 512, $\lambda_{\text{cls}} = 0.1$, $\lambda_{\text{box}} = 0.025$, and $\lambda_{\text{aux}} = 0.03125$. For R101, Sections 3.1–3.2 yield **101 concept clusters, including 17 blindspots (15 Car and 2 Person)**; these clusters and their blindspot labels remain frozen for subsequent analyses and risk screening.

4.2 Task Fidelity of DetSAE

Figure 3 evaluates whether the learned concept representations preserve the detector behavior they explain. We compare DetSAE with K-means, NMF, and SAE on identical source RoIs, 64 coordinates, three seeds, and the same scene-disjoint split. K-means activates one coordinate; the others retain the top eight. Cosine similarity and reconstruction MSE compare original and reconstructed RoI features (with MSE computed in standardized feature space); Class KL and Box SL1 compare their frozen-head class probabilities and box-regression outputs, while label flips measure the fraction of objects whose predicted class changes after reconstruction.

At nearly identical cosine similarity to SAE (0.7649 vs. 0.7655), DetSAE reduces Class KL, Box SL1, and label flips by 39.2%, 3.3%, and 26.3%, respectively. Thus, DetSAE better preserves detector behavior without sacrificing feature-level fidelity.

4.3 Detection Difficulty by Active Concept Composition

We evaluate the frozen source-discovered clusters and blindspot labels on external datasets. Target GT objects are grouped by their active DetSAE clusters under the same activation-share criterion: *only blindspot*, *mixed* (blindspot and normal), or *only normal*. We compute group-conditional COCO-style $\text{mAP}_{50:95}$ (Lin et al., 2014) using full-image detections.

Figure 4 shows a consistent mAP ordering of

$$\textit{only blindspot} < \textit{mixed} < \textit{only normal}$$

in all four datasets. This pattern reveals a strong negative association between activated blindspot concepts and detection performance. The same ordering is preserved in all 100 resamples for each dataset, using the Car category with matched images and equal numbers of objects per group. This control rules out unequal group sizes or the concentration of blindspot-only objects in intrinsically difficult images as simple explanations for the observed gap. These results show that blindspot activations are consistently associated with markedly greater object-level detection difficulty.

4.4 Causal Validation Across Detector Backbones

We next evaluate the causal intervention in Figure 2 across detector backbones. We apply paired restoration and the matched-magnitude controls from Section 3.3 to independently trained R101-FPN and R50-FPN Faster R-CNN detectors. The normal-concept control uses a concept from the same object. Concepts and blindspots are discovered separately for each detector; ResNet-50-FPN (R50) yields 13 Car and 3 Person blindspots.

Table 1: **Causal intervention results.** Margin denotes the change in true-class logit margin; Top-1 denotes net classification recovery in percentage points. Higher is better.

	R101-Car		R101-Person		R50-Car		R50-Person	
Intervention	Margin	Top-1	Margin	Top-1	Margin	Top-1	Margin	Top-1
Random	-0.001	-0.12	+0.004	0.00	-0.000	+0.12	+0.003	+0.16
Opposite	-2.373	-4.57	-1.230	-1.60	-0.652	-2.18	-0.389	-0.47
Normal concept	+0.120	+0.31	-0.002	-0.20	+0.025	-0.05	-0.004	-0.32
Paired restoration	**+2.326**	**+8.24**	**+1.228**	**+1.60**	**+0.649**	**+2.57**	**+0.388**	**+1.11**

Paired restoration yields the largest margin and Top-1 gain in every detector–class setting (Table 1). In contrast, reversing the intervention consistently degrades classification, while normal-concept and random interventions produce only minor changes. These controlled interventions demonstrate a causal role of excessive blindspot activations in frozen-head classification errors, suggesting a potential route for modulating detector failures through the corresponding blindspot activations.

5 Blindspot-Aware Detection-Error Screening

We next test whether the same frozen blindspot representation can rank images by false-negative (FN) and false-positive (FP) risk.

5.1 Evaluation Protocol

Error definitions. An FN is a ground-truth object without a correct-class detection at confidence ≥ 0.05 and IoU ≥ 0.5. For FP evaluation, we retain at most the 100 highest-confidence detections per image and class with confidence ≥ 0.5. We match these in descending confidence order to same-class ground truth at IoU ≥ 0.5, with each object matched at most once. Unmatched detections are FPs. On KITTI, detections matched to `Van` or `Person_sitting`, or with at least 50% of their area inside `DontCare`, are ignored in FP evaluation.

Risk-ranking metrics. For error type $\tau \in \{\mathrm{FN}, \mathrm{FP}\}$, let N_n^τ denote the error count in image n and σ rank M images from highest to lowest risk:

$$\mathrm{Capture@5\%} = \frac{\sum_{j=1}^{\lceil 0.05M \rceil} N_{\sigma_j}^\tau}{\sum_{n=1}^{M} N_n^\tau}. \tag{14}$$

AUROC and AUPRC (computed as average precision) evaluate image-level error occurrence $b_n^\tau = \mathbf{1}[N_n^\tau > 0]$ (Fawcett, 2006; Davis & Goadrich, 2006). NAURC measures the error rate remaining as high-risk images are rejected (Cattelan & Silva, 2023): $(A - A_*)/(\pi - A_*)$, where A is the mean accepted-prefix error rate, π is error prevalence, and $A_* = \pi + (1-\pi)\log(1-\pi)$ is the oracle reference. Higher Capture@5%, AUROC, and AUPRC are better; lower NAURC is better.

5.2 Implementation Details and Baselines

Training and controls. FN and FP predictors are trained separately with AdamW (2.5×10^{-4} learning rate, 10^{-4} weight decay). We compare two risk-predictor training variants. *Ours (OG)* uses only original reference images, whereas *Ours (MS)* additionally uses their 20 style variants. Both share the same frozen detector, DetSAE, and blindspot set identified using style variants. The OG–MS comparison therefore isolates the effect of style augmentation during risk-predictor training, with all other settings unchanged. Each gradient-trained method is trained for 20,000 updates per candidate with three seeds, learning-rate multipliers $\{0.5, 1, 2\}$, and effective batch size 64. Checkpoints are selected on the 500 original calibration images using task-specific source Capture@5%, with AUROC and earlier updates breaking ties. No evaluation-set labels are used for training or model selection. All methods share detector outputs, evaluation images, category mappings, and evaluation code.

Baselines. We compare SF, LFR, LF-ASH, and LFA (Rahman et al., 2021; Yatbaz et al., 2024; Keser & Knoll, 2026); SAOD, ViLU, BTA, and KGFP (Oksuz et al., 2023; Lafon et al., 2025; Guarino et al., 2026; Zimmermann et al., 2026); and GRAM, KNN-5, and Residual (Sastry & Oore, 2020; Sun et al., 2022; Wang et al., 2022). All trainable baselines use only Cityscapes.

5.3 Detection-Error Risk Screening

Table 2: **Camera benchmark.** FN/FP risk screening on KITTI and BDD100K.

	KITTI				BDD100K			
Method	C@5	AUROC	AUPRC	NAURC	C@5	AUROC	AUPRC	NAURC
SF	11.3/5.5	63.8/48.5	38.2/86.4	67.8/95.9	5.5/4.5	54.1/49.7	84.8/92.8	94.6/98.4
LFR	5.0/8.9	55.9/63.7	28.5/93.0	83.1/80.4	5.2/5.8	52.7/57.8	84.2/94.3	94.5/91.4
LF-ASH	9.7/7.8	62.6/74.3	35.4/94.8	67.0/63.4	5.7/6.1	51.3/59.5	83.5/95.1	100.0/89.6
LFA	**14.8**/4.5	**72.0**/66.6	**48.0**/92.6	51.3/82.2	5.8/6.2	54.3/50.9	85.5/93.7	93.0/100.1
SAOD	0.5/0.6	30.0/22.2	18.0/77.6	149.0/125.5	5.0/1.0	47.6/27.6	80.4/86.1	101.9/117.3
ViLU	12.7/7.0	58.4/69.7	37.1/94.0	87.1/76.2	6.3/5.7	58.8/57.1	86.7/94.3	84.5/87.8
BTA	10.3/6.7	69.4/57.7	41.7/90.6	55.0/92.2	3.7/5.1	48.0/54.9	81.5/93.1	103.6/91.5
KGFP	9.5/8.3	71.2/73.7	42.6/94.9	49.8/61.9	**7.3**/6.8	62.0/64.3	88.8/95.8	**81.7**/85.9
GRAM	7.2/5.5	43.8/34.6	25.0/84.2	121.2/116.7	4.5/4.5	45.3/52.6	81.2/93.3	106.6/96.5
KNN-5	14.4/9.2	67.6/49.3	43.0/88.7	55.1/100.0	4.3/5.5	48.4/59.1	81.9/94.6	104.2/86.5
Residual	5.4/5.1	48.4/35.7	24.8/83.3	104.7/116.6	5.5/4.7	48.1/48.9	82.6/93.1	103.5/103.9
Ours (OG)	14.0/**10.4**	71.9/**89.7**	46.3/**98.2**	**48.9**/**31.1**	7.2/**10.5**	**63.6**/**87.3**	**89.1**/**98.9**	83.4/**39.9**
Ours (MS)	**15.0**/**10.2**	**75.3**/**91.0**	**49.7**/**98.4**	**39.5**/**23.5**	**7.8**/**10.9**	**64.8**/**89.6**	**89.2**/**99.1**	**73.0**/**34.1**

Table 3: **Simulation benchmark.** FN/FP risk screening on SIM10K and RealDriveSim.

	SIM10K				RealDriveSim			
Method	C@5	AUROC	AUPRC	NAURC	C@5	AUROC	AUPRC	NAURC
SF	4.9/2.6	51.0/50.2	72.4/69.3	90.6/96.6	5.8/5.1	57.0/58.2	91.7/86.1	89.4/86.5
LFR	4.4/6.9	49.6/56.5	72.1/74.9	102.2/90.7	6.0/6.4	60.4/62.9	92.5/88.4	82.9/76.1
LF-ASH	4.1/5.9	49.5/57.9	71.3/73.5	97.2/83.0	5.4/6.8	57.9/67.9	91.7/90.2	80.6/68.3
LFA	6.3/8.7	53.2/56.8	74.8/75.3	96.2/91.4	**8.7**/8.2	73.4/59.5	96.1/88.1	64.4/87.7
SAOD	3.2/0.1	41.6/22.6	67.4/53.2	110.4/138.1	1.1/0.9	29.5/35.3	81.8/74.8	118.5/116.9
ViLU	**7.3**/5.0	57.3/58.9	76.9/73.9	**84.7**/78.4	5.4/5.1	71.2/61.5	94.9/87.1	68.5/77.3
BTA	3.3/2.2	46.1/38.7	69.6/61.9	105.0/117.1	3.2/4.5	48.5/54.8	88.5/85.1	102.8/96.3
KGFP	5.7/8.5	56.4/60.6	77.1/77.4	88.5/83.1	7.6/8.0	**78.1**/65.5	96.7/90.5	**55.7**/76.9
GRAM	5.2/5.3	55.3/51.3	75.0/69.5	90.8/94.3	4.2/3.8	42.0/48.2	86.7/81.3	107.1/101.1
KNN-5	**7.5**/4.2	55.0/49.1	75.2/68.7	89.5/98.2	4.6/6.5	46.6/57.0	88.4/86.3	103.1/90.3
Residual	4.4/4.2	49.9/40.3	71.3/63.6	98.5/118.2	4.2/5.8	34.6/45.6	85.2/81.2	115.9/106.8
Ours (OG)	5.7/**12.4**	**57.5**/**84.0**	**77.7**/**90.9**	89.3/**31.1**	8.6/**11.6**	76.6/**86.6**	**96.8**/**96.7**	72.1/**33.0**
Ours (MS)	5.4/**14.3**	**60.1**/**86.6**	**77.9**/**92.6**	**80.6**/**24.0**	**9.1**/**11.5**	**82.9**/**90.5**	**97.5**/**97.7**	**46.1**/**21.7**

Tables 2–4 report FN/FP scores multiplied by 100; C@5 denotes Capture@5%. Bold-underlined and bold values denote the best and second-best results, respectively, according to the unrounded scores.

Table 4: **Weather benchmark.** FN/FP risk screening on Rainy- and Foggy-Cityscapes.

	Rainy-Cityscapes ($\alpha = 0.02$)				Foggy-Cityscapes ($\beta = 0.02$)			
Method	C@5	AUROC	AUPRC	NAURC	C@5	AUROC	AUPRC	NAURC
SF	3.8/5.7	60.9/69.0	98.4/91.3	98.0/72.3	6.8/9.9	61.8/67.3	94.2/83.3	84.5/65.6
LFR	6.8/8.4	40.1/58.8	97.1/89.5	113.2/96.6	**6.9**/11.6	**75.4**/69.3	**96.5**/86.2	59.2/65.5
LF-ASH	5.1/5.1	69.1/69.2	98.9/91.6	93.9/79.5	**6.9**/8.3	75.4/62.7	**96.5**/79.7	61.9/69.6
LFA	6.8/7.1	55.7/57.5	97.7/87.4	97.6/89.8	6.2/10.2	65.7/63.6	94.8/83.2	71.7/78.6
SAOD	3.7/3.1	70.7/46.7	98.9/82.2	92.3/105.3	1.3/0.6	31.2/40.2	83.1/64.4	115.4/112.0
ViLU	4.5/7.0	**81.1**/53.5	**99.4**/84.8	**81.4**/87.2	6.1/7.5	**78.5**/63.3	96.4/80.6	**51.0**/74.1
BTA	4.7/3.9	56.3/56.2	97.3/86.6	98.6/99.1	5.3/4.1	52.3/61.7	90.2/78.5	96.7/83.2
KGFP	7.0/7.7	70.6/51.4	98.9/85.7	91.7/103.0	6.2/8.3	74.1/61.6	95.4/81.1	**52.0**/84.1
GRAM	4.1/2.2	32.7/41.6	96.1/77.5	115.6/97.4	4.8/6.7	45.8/52.1	89.7/74.6	107.0/93.6
KNN-5	4.6/2.4	66.8/43.3	98.6/79.6	89.7/110.7	5.2/8.8	45.7/58.4	90.8/79.6	109.3/86.7
Residual	5.0/2.4	29.1/35.3	94.6/76.7	115.4/114.8	4.9/6.6	47.5/54.0	90.6/74.6	105.5/86.7
Ours (OG)	**7.1**/**13.6**	65.0/**71.9**	98.6/**92.0**	95.4/**57.5**	6.3/**15.3**	63.7/**77.1**	94.9/**89.9**	90.2/**46.3**
Ours (MS)	**7.6**/**16.3**	**85.9**/**73.6**	**99.5**/**93.4**	**68.7**/**68.6**	**6.9**/**16.1**	69.2/**82.3**	95.1/**92.8**	64.6/**35.2**

Both OG and MS outperform all external baselines in all 24 FP dataset–metric comparisons. MS ranks first in 21 FP and 20 FN comparisons, while OG leads the remaining three FP comparisons. Relative to OG, predictor-stage style augmentation improves 23 of 24 FN metrics and 21 of 24 FP metrics.

5.4 Risk-Representation Ablation

We ablate the risk representation on BDD100K and KITTI, with FN screening as a representative task. Table 5 compares random normal concepts, blindspots only, all concepts, and all concepts with explicit blindspot channels (our full representation). Random-normal subsets contain 17 clusters matched to the blindspots in class composition, with results averaged over five independently trained draws; all other training settings are fixed.

Table 5: **Risk-representation ablation.** FN screening on BDD100K and KITTI.

Domain	Input concepts	C@5	AUROC	AUPRC	NAURC
BDD100K	Random normal concepts	7.3	57.8	86.3	86.0
	Blindspots only	**8.4**	61.9	87.8	79.5
	All concepts	7.1	64.0	89.0	74.9
	All+blindspots (Ours)	7.8	**64.8**	**89.2**	**73.0**
KITTI	Random normal concepts	12.0	67.7	40.8	55.0
	Blindspots only	13.3	71.3	44.1	48.2
	All concepts	14.4	74.4	48.7	40.6
	All+blindspots (Ours)	**15.0**	**75.3**	**49.7**	**39.5**

Blindspots outperform matched random normal concepts on every metric in both datasets. Adding explicit blindspot channels to all concepts further improves all four metrics.

6 Conclusion

We introduced detector-native concept blindspots to study systematic object detection failures. By connecting blindspot discovery, causal validation, and FN/FP risk screening, our framework shows that seemingly isolated errors can reflect shared concept-level vulnerabilities.

AI use statement

In this work, we used generative AI tools to generate controlled style-transfer variants used as synthetic data for concept-blindspot discovery and risk-predictor training (Section 3.2) . We have not used generative AI tools for problem formulation, hypothesis development, research methodology or

experimental design, method implementation, translation, dataset cleaning or reformatting, or result interpretation; tasks involving theoretical models, mathematical claims or proofs, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for wording refinement and language editing of the manuscript. We have reviewed all AI-assisted work. Generated style variants were manually checked against their source images for object count, shape, position, and annotation consistency, and all AI-assisted text was reviewed by the authors. We take responsibility for the final content of this work, including text, claims, data, and artifacts produced with the aid of generative AI.

REFERENCES

David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4502–4511, 2019.

Maty Bohacek, Thomas Fel, Maneesh Agrawala, and Ekdeep Singh Lubana. Uncovering conceptual blindspots in generative image models using sparse autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 25715–25766, 2026.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying functionally important features with end-to-end sparse dictionary learning. *Advances in Neural Information Processing Systems*, 37:107286–107325, 2024.

ByteDance Seed Team. Seedream 5.0 Lite. Official model documentation, 2026. URL `https://seed.bytedance.com/en/seedream5_0_lite`. Accessed September 12, 2026.

Luís Felipe P Cattelan and Danilo Silva. How to fix a broken confidence estimator: Evaluating post-hoc methods for selective classification with deep neural networks. *arXiv preprint arXiv:2305.15508*, 2023.

Muxi Chen, Chenchen Zhao, and Qiang Xu. Hibug2: Efficient and interpretable error slice discovery for comprehensive model debugging. In *International Conference on Learning Representations*, volume 2025, pp. 102566–102590, 2025.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Muhammad Sohail Danish, Muhammad Haris Khan, Muhammad Akhtar Munir, M Saquib Sarfraz, and Mohsen Ali. Improving single domain-generalized object detection: A focus on diversification and alignment. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17732–17742. IEEE, 2024.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022.

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

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.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025a.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025b.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Vanessa Emanuela Guarino, Claudia Winklmayr, Jannik Franzen, Josef Lorenz Rumberger, Manuel Pfeuffer, Sonja Greven, Klaus Maier-Hein, Dagmar Kainmueller, Christoph Karg, and Carsten T Lüth. Better than average: Spatially-aware aggregation of segmentation uncertainty improves downstream performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13145–13156, 2026.

Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8014–8023. IEEE, 2019.

Arpit Jadon, Haoran Wang, Phillip Thomas, Michael Stanley, S Nathaniel Cibik, Rachel Laurat, Omar Maher, Lukas Hoyer, Ozan Unal, and Dengxin Dai. Realdrivesim: A realistic multi-modal multi-task synthetic dataset for autonomous driving. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pp. 316–323. IEEE, 2025.

Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016.

Mert Keser and Alois Knoll. Lfa: Layer feature attention for run-time introspection of 2d object detectors in automated driving. *arXiv preprint arXiv:2606.00372*, 2026.

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.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. Pmlr, 2020.

Yael Konforti, Mateo Espinosa Zarlenga, Elaf Almahmoud, and Mateja Jamnik. Cb-slice: Concept-based interpretable error slice discovery. *arXiv preprint arXiv:2605.29836*, 2026.

Marc Lafon, Yannis Karmim, Julio Silva-Rodríguez, Paul Couairon, Clément Rambour, Raphaël Fournier-Sniehotta, Ismail Ben Ayed, Jose Dolz, and Nicolas Thome. Vilu: Learning vision-language uncertainties for failure prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17807–17817, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yajing Liu, Shijun Zhou, Xiyao Liu, Chunhui Hao, Baojie Fan, and Jiandong Tian. Unbiased faster r-cnn for single-source domain generalized object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28838–28847. IEEE, 2024.

Yuzuru Nakamura, Yasunori Ishii, and Takayoshi Yamashita. Active domain adaptation with false negative prediction for object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28782–28792. IEEE, 2024.

Kemal Oksuz, Tom Joy, and Puneet K Dokania. Towards building self-aware object detectors via reliable uncertainty quantification and calibration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9263–9274. IEEE, 2023.

Gregory Plumb, Nari Johnson, Ángel Alexander Cabrera, and Ameet Talwalkar. Towards a more rigorous science of blindspot discovery in image classification models. *arXiv preprint arXiv:2207.04104*, 2022.

Quazi Marufur Rahman, Niko Sunderhauf, and Feras Dayoub. Per-frame map prediction for continuous performance monitoring of object detection during deployment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 152–160, 2021.

N Dinesh Reddy, Robert Tamburo, and Srinivasa G Narasimhan. Walt: Watch and learn 2d amodal representation from time-lapse imagery. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9346–9356. IEEE, 2022.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International conference on machine learning*, pp. 8491–8501. PMLR, 2020.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning*, pp. 20827–20840. PMLR, 2022.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4911–4920. IEEE, 2022.

Hakan Yekta Yatbaz, Mehrdad Dianati, Konstantinos Koufos, and Roger Woodman. Run-time introspection of 2d object detection in automated driving systems using learning representations. *IEEE Transactions on Intelligent Vehicles*, 9(6):5033–5046, 2024.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2633–2642. IEEE, 2020.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Wei Zhang, Chaoqun Wang, Zixuan Guan, Ping Sheng Kao, Pengfei Zhao, Peng Wu, and Sifeng He. Gh-esd: Grounded hypothesis-driven error slice discovery for instance-level vision tasks. *arXiv preprint arXiv:2512.24592*, 2025.

Jakob Zimmermann, Gerrit Holzbach, and David Lerch. Knowledge-guided failure prediction: Detecting when object detectors miss safety-critical objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4312–4321, 2026.