Knowledge-Guided Failure Prediction: Detecting When Object Detectors Miss Safety-Critical Objects

Jakob Paul Zimmermann
Fraunhofer HHI
Berlin, Germany
jakob.zimmermann@campus.tu-berlin.de

Gerrit Holzbach
Fraunhofer IOSB
Karlsruhe, Germany
gerrit.holzbach@iosb.fraunhofer.de

David Lerch
Fraunhofer IOSB
Karlsruhe, Germany
david.lerch@iosb.fraunhofer.de

Abstract

Object detectors deployed in safety-critical environments can fail silently, e.g. missing pedestrians, workers, or other safety-critical objects without emitting any warning. Traditional Out-of-Distribution (OOD) detection methods focus on identifying unfamiliar inputs, but do not directly predict functional failures of the detector itself. We introduce Knowledge-Guided Failure Prediction (KGFP), a representation-based monitoring framework that treats missed safety-critical detections as anomalies to be detected at runtime. KGFP measures semantic misalignment between internal object detector features and visual foundation model embeddings using a dual-encoder architecture with an angular distance metric. A key property is that when either the detector is operating outside its competence or the visual foundation model itself encounters novel inputs, the two embeddings diverge, producing a high-angle signal that reliably flags unsafe images. We compare our novel KGFP method to baseline OOD detection methods. On COCO person detection, applying KGFP as a selective-prediction gate raises person recall among accepted *images from 64.3% to 84.5% at 5% False Positive Rate (FPR), and maintains strong performance across six COCO-O visual domains, outperforming OOD baselines by large margins. Our code, models, and features are published at* https://gitlab.cc-asp.fraunhofer.de/iosb_public/KGFP

1. Introduction

Object detectors power safety-critical applications from autonomous driving to video surveillance, yet they fail unpredictably when encountering novel or challenging visual conditions [10, 11]. As automation increases, such perception components are deployed more widely in operational decision loops and increasingly fall under safety certification expectations [27].

A pedestrian detection system trained on clear weather may miss occluded pedestrians in fog; a construction site monitor [7] may fail to detect workers in unusual poses. These **silent failures**—where detectors confidently return empty results despite objects being present—can have catastrophic consequences.

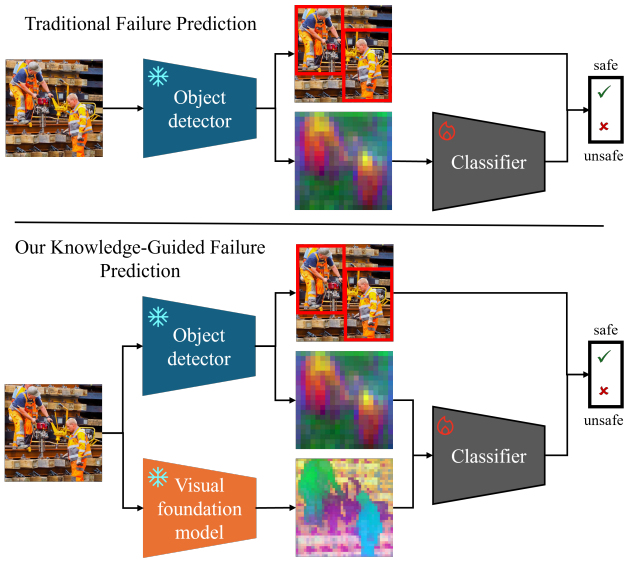

Figure 1. Overview of our knowledge-guided failure prediction approach. Unlike traditional methods that rely solely on internal detector features, we leverage visual foundation model features to predict whether an image is safe for person detection.

Traditional approaches to detector robustness focus on *improving detection accuracy* through domain adaptation [49], data augmentation, or architecture refinement [19]. However, even state-of-the-art detectors exhibit failure modes including missing objects in complex contexts [37] and under novel visual conditions [46]. For high-assurance settings, achieving the evidence and risk reduction expected at Safety Integrity Level 4 [12] cannot rely on test set accuracy alone; it also requires mechanisms that detect and control unsafe behavior at runtime when assumptions are violated. Rather than attempting to eliminate failures entirely, **runtime monitoring** [15, 33] aims to detect anomalous situations in which the perception system is likely to miss safety-critical objects and to issue timely alerts. Such alerts are essential for the system to **adapt and operate** safely under uncertainty, enabling downstream controllers to trigger fallback policies (e.g., initiating a minimal risk maneuver or querying a human operator) when functional failures are predicted.

We formulate **failure prediction for safety-critical objects** as a supervised binary classification task: given an image and a detector's predictions, determine whether the detector has missed any safety-critical objects. Crucially, we focus on a **safety-critical subset of object classes**—for pedestrian detection systems, persons are the primary safety concern; objects like small birds or distant vehicles matter less for immediate safety. This differs fundamentally from the typically unsupervised binary decision problem of Out-of-Distribution (OOD) detection [45].

Knowledge-Guided Failure Prediction (KGFP) is trained to predict the safety of In-Distribution (ID) data, and demonstrates generalization capabilities to OOD data.

A key challenge in anomaly-based monitoring is alarm fatigue. Unlike methods that flag any novel input, KGFP focuses on functional failures, significantly reducing false alarms by ignoring benign novelty that does not impact the detection of safety-critical objects. Moreover, unlike conventional OOD detection, it flags ID images where the detector is struggling.

Recent foundation models trained on billions of images [28] capture rich semantic knowledge beyond task-specific detectors. Self-supervised vision transformers like DINO [3] learn particularly powerful semantic representations. Our key insight: **semantic similarity between learned encodings of detector internal features and foundation model embeddings indicates detection reliability**. When encoded representations of YOLOv8's internal activations align with encoded DINO semantic embeddings, the detector operates within its competence zone. Misalignment signals potential failure.

Our main contributions are as follows:

1. **Safety-focused failure prediction**: To the best of our knowledge, we introduce the first framework that explicitly predicts detector failures with respect to safety-critical object classes (persons) rather than generic distribution membership, enabling targeted monitoring of safety-critical detections. We propose a novel evaluation metric that quantifies the percentage of ground-truth persons detected in accepted images (Person Recall) at a 5% false-alarm rate (False Positive Rate (FPR)) of the KGFP module.
2. **Foundation model integration**: We demonstrate that self-supervised world-knowledge (DINO) improves failure prediction over detector features alone through multi-scale fusion with cross-attention. We propose a dual-encoder architecture that uses cosine similarity between encoded representations of YOLOv8 internal features and DINO embeddings, where angular divergence directly signals unsafe images in which safety-critical objects will be missed.
3. **Comprehensive evaluation**: We perform systematic ablations across architecture, training, and foundation model choices with strong baselines on both ID data and novel visual domains.

2. Related Work

2.1. Out-of-Distribution Detection

OOD detection aims to identify inputs from unknown distributions [44, 45]. Classification-based methods leverage model outputs: Maximum Softmax Probability (MSP) [9], ODIN [18] with temperature scaling, Energy scores [21], and activation-based methods like ReAct [5, 38]. Distance-based approaches measure feature space deviations: Mahalanobis distance [17], K-Nearest Neighbors (KNN) [39], and recent improvements via feature normalization [23, 26]. GRAM [32] computes Gram matrices (channel covariance) from YOLOv8 feature maps at each scale using orders 1–5, fits min/max statistics per scale on training data, and flags test images as unsafe when Gram matrix elements fall outside these bounds. We use no spatial pooling to preserve full feature map statistics. KNN [39] stores spatially-pooled YOLOv8 features from training images. At test time, it computes the Euclidean distance to the 5th nearest training neighbor per scale; high distances indicate anomalous inputs and potential failure. We use L2 normalization for stable distance computation. Virtual-logit Matching (ViM) [43] fits Principal Component Analysis (PCA) with 100 components per YOLOv8 scale on training features and projects test features onto the residual subspace (orthogonal to the principal components). Large residual norms indicate OOD samples, i.e. features that deviate from the low-dimensional ID manifold. We also trained a DINO-ViM variant, where we use DINO embeddings instead of internal YOLOv8 activations.

Although all of these methods excel at image classification OOD detection, **object detection presents unique challenges**: (1) spatial localization requirements, (2) multi-

ple objects per image, (3) partial detection scenarios. Recent work adapts object detectors for OOD tasks [51], but focuses on flagging novel concepts in the scene rather than predicting safety-critical failures (e.g., missed pedestrians).

Angular margins have proven effective for learning discriminative embeddings in face recognition [1, 4, 22]. Recent work extends angular margins to OOD detection [29], showing that cosine-based similarity naturally separates ID from OOD samples when features are normalized. Our approach builds on these insights, using cosine similarity between dual-encoder projections as a safety metric, where angular distance directly measures semantic alignment between detector and foundation model representations.

2.2. Runtime Monitoring for Safety

Runtime monitoring provides safety assurance for deployed ML systems [33]. Comprehensive frameworks have been developed for perception systems: Ferreira et al. [6] survey threats (ID errors, novel visual conditions, adversarial attacks) and detection mechanisms, while Tran et al. [41] demonstrate simulation-based verification using Linear Temporal Logic for autonomous driving. Model assertions [15] check intermediate activations against expected distributions. Recent work by Torpmann-Hagen et al. [40] proposes a paradigm shift from binary OOD classification to regression-based loss prediction, training Generalized Additive Models to directly predict task loss (cross-entropy, Jaccard index) from OOD-ness features (MSP, Energy, KNN distance) for continuous risk assessment. While loss prediction provides continuous risk quantification, it relies on proxy regression rather than directly predicting safety-critical events (missed persons). Our binary formulation enables end-to-end training with explicit safety labels, learning semantic features that directly indicate when safety-critical objects are missed.

For object detection specifically, Yang et al. [46] predict false negatives using detector uncertainty, while Yatbaz et al. [47] monitor early layer patterns in 3D detectors. He et al. [8] address YOLOv8 hallucinations on OOD inputs through proximal OOD fine-tuning. Most closely related, DECIDER [36] trains a separate classifier to predict whether a detector will fail on a given input, but does not incorporate foundation model knowledge or angular distance metrics. However, these approaches rely solely on detector-internal signals. Our work differs by incorporating external semantic knowledge from foundation models, enabling detection of failures where internal features appear normal but semantic misalignment indicates unreliable predictions.

Unlike DECIDER [35], which requires computing similarity against class-specific textual attribute embeddings during inference to generate an auxiliary model's predictions, our framework relies solely on the angular alignment with general-purpose visual foundation model features, eliminating the dependency on predefined text definitions.

2.3. Foundation Models for Robustness

Self-supervised image and video models like DINO [3] and DINOv2 [24] or vision-language models like CLIP [28] learn transferable representations from massive unlabeled data. Recent work applies these to zero-shot anomaly detection: WinCLIP [13], AnomalyCLIP [50], and DINO prototypes [34]. P et al. [25] fuse DINO with YOLOv8 for data-efficient detection. Broader surveys [2] examine foundation models for visual anomaly detection across industrial and medical domains. Our work differs critically: while Wang et al. enhance detection accuracy, **we leverage foundation models for failure prediction**.

Modern detectors employ multi-scale feature pyramids [19]: YOLO [30] predicts at multiple resolutions, Faster R-CNN [31] uses Region Proposal Networks, and DETR variants [3] apply transformers. Hoiem et al. [11] analyze detector failures systematically, finding size and occlusion as dominant factors.

3. Method

We present KGFP, a supervised monitoring framework that predicts when an object detector will miss safety-critical objects. Unlike standard OOD detection, which flags distributional novelty in an unsupervised manner, KGFP is trained with explicit safe/unsafe labels derived from detector performance, enabling it to distinguish harmful failures from benign distribution shifts.

3.1. Problem Formulation

Given an image $\mathbf{x}$ and object detector D, let $\mathcal{Y}_{\text{safe}} = \{y_1, \ldots, y_n\}$ denote ground-truth bounding boxes for **safety-critical object classes**. For example, in pedestrian detection systems, $\mathcal{Y}_{\text{safe}}$ contains only persons, ignoring all other object classes. The detector produces predictions $\hat{\mathcal{Y}} = D(\mathbf{x})$ across all classes. Define the **failure label**:

$$z(\mathbf{x}) = \begin{cases} 1 & \text{(unsafe) if any } y_i \in \mathcal{Y}_{\text{safe}} \\ & \text{undetected (IoU} < 0.5\text{)} \\ 0 & \text{(safe) if all safety-critical} \\ & \text{objects matched} \end{cases} \tag{1}$$

The failure prediction task: predict $z(\mathbf{x})$ from $(D(\mathbf{x}), \mathbf{x})$ to identify unsafe images where D will miss safety-critical objects. This formulation differs from OOD detection by directly measuring *functional failure on safety-critical objects* rather than distributional novelty.

The objective of this study is to develop a model capable of predicting the failure of the object detector on a test sample. The technical approach entails the identification of this issue as OOD detection. In this context, the ID samples are correctly classified, while the detector exhibits failure on OOD data. Our training and evaluation is exclusively

focused on the person class, as pedestrian detection serves as the primary motivation for safety-critical applications.

Conceptually, the default output of a safety monitor for random input should be a safety warning. As two high dimensional random vectors are close to orthogonal with high probability [42], this expected behavior is built into the concept of KGFP.

3.2. Architecture: Dual-Encoder with Angular Metric

In order to evaluate a specialized model like an object detector we leverage the visual foundation model DINO [3]. We introduce a dual-encoder architecture that processes multi-scale YOLOv8 [14] Feature Pyramid Network (FPN) activations through cross-scale fusion and transformer blocks with self-attention and cross-attention to foundation model embeddings (see Figure 2). The resulting detector and world-knowledge representations are projected onto a shared embedding space, where their angular similarity provides a direct measure of detection reliability. In our ablation studies we show that our designed angular similarity metric performs on par on OOD and outperforms the MLP baseline on ID setting. This underscores the effectiveness of our angular failure metric.

Multi-Scale Feature Extraction For each image $\mathbf{x}$, we extract two complementary representations:

- **Predictor features**: YOLOv8l [14] (large model) produces multi-scale internal features from its FPN at levels $\{P3, P4, P5\}$ with channel dimensions $\{256, 512, 512\}$ and spatial resolutions 80×80, 40×40, 20×20 respectively (at 640×640 input). These scales correspond to detection of small (P3), medium (P4), and large (P5) objects. All scales are projected to a common channel dimension and upsampled to the largest spatial resolution (80×80, matching P3).
- **World-knowledge features** $f_{wk} \in \mathbb{R}^{768}$: DINO [3] Vision Transformer (ViT) global [CLS] embeddings extracted from the input image resized to 518×518 [24].

The overall architecture is shown in Figure 2.

Pre-Fusion Cross-Scale Attention Before fusing YOLO's multi-scale features, we apply cross-scale attention to allow information exchange between pyramid levels. Each scale attends to features from other scales, where each scale is treated as a single token. This allows the model to create scale-aware representations where, for example, P3 features (small objects) can leverage context from P5 features (scene-level patterns). The three scales are fused by element-wise addition, patchified, and finally processed through transformer blocks with self-attention.

Post-Fusion Transformer We apply transformer blocks with both self-attention and cross-attention mechanisms. The fused YOLOv8 features are patchified (4×4 patches) and processed through 2 self-attention blocks followed by 2 cross-attention blocks.

$$h'_{pr} = \text{SelfAttn}^{(2)}(f_{pr}) \tag{2}$$

$$h''_{pr} = \text{CrossAttn}^{(2)}(Q = h'_{pr}, K = f_{wk}, V = f_{wk}) \tag{3}$$

In the latter the queries are projections of the predictor embeddings, whereas the keys and values are projections of the DINO embedding. This procedure allows the detector representations to query foundation model semantic knowledge.

Self-attention refines YOLOv8 features by capturing long-range spatial dependencies across patches, while cross-attention allows the detector's representations to query DINO's semantic knowledge. This two-stage refinement creates semantically-grounded, spatially-coherent representations. We use 8 attention heads, 2 self-attention blocks, and 2 cross-attention blocks.

Dual Encoders Separate encoder heads map the refined features of the predictor and world-knowledge model to a shared 64-dimensional embedding space $\mathbb{R}^{64}$, which we denote as e_{pr} and e_{wk} respectively:

$$e_{pr} = E_{PR}(h''_{pr}; \theta_{pr}) \in \mathbb{R}^{64} \tag{4}$$

$$e_{wk} = E_{WK}(f_{wk}; \theta_{wk}) \in \mathbb{R}^{64} \tag{5}$$

The YOLOv8 encoder processes the post-fusion transformer output through global pooling and projection. The DINO encoder is a deep 5-layer Multi-Layer Perceptron (MLP): $768 \to 1024 \to 768 \to 640 \to 512 \to 64$, with LayerNorm and Gaussian Error Linear Unit (GELU) activations.

Angular Failure Metric We measure this angle between the two embedings e_{pr} and e_{wk} of the two encoders via cosine similarity

$$s_{\text{safety}}(\mathbf{x}) = \frac{e_{pr} \cdot e_{wk}}{\|e_{pr}\|_2 \|e_{wk}\|_2}. \tag{6}$$

High similarity (small angle) indicates semantic similarity and predicted safety, while low similarity (large angle) signals misalignment and likely detector failure.

3.3. Training Objective

We train the model end-to-end using Binary Cross-Entropy (BCE) loss. The cosine similarity score $s_{\text{safety}}(\mathbf{x}) \in [-1, 1]$ is mapped to safety probability $p_{\text{safe}}(\mathbf{x}) \in [0, 1]$ via $(1-s)/2$, where -1 ($\pm 180°$ angle) maps to 1 (unsafe) and +1 (0° angle) maps to 0 (safe). BCE loss is then applied with safety labels $y \in \{0, 1\}$ (0 = safe, all persons detected; 1 = unsafe, persons missed). With this encoding high cosine similarity (close

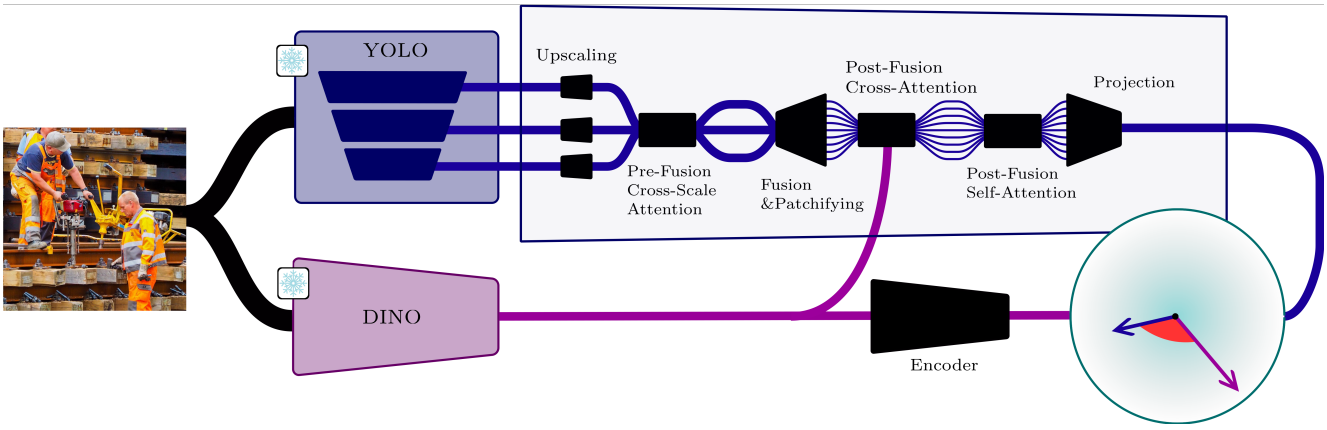

Figure 2. The dual-encoder architecture of our KGFP. We use pretrained DINO and YOLO models as backbones. For our KGFP we freeze the pretrained backbones and fine-tune a fusion framework. During evaluation the distance between DINO features and the fused features serves as the measure for our failure prediction.

embeddings) indicates safety, while low similarity (distant embeddings) indicates failure.

For model training and threshold tuning respectively we split COCO train 2017 (64,115 images) into training (90%) and validation (10%) sets [20]. COCO val2017 (2,693 images) serves as our held-out test set for final evaluation.

At test time, we compute the safety score $s_{\text{safety}}(\mathbf{x})$. The final prediction is based on a threshold, that is tuned on a disjoint validation set to 5% FPR. In our experiments of the full KGFP, this threshold is set to 0.843. That is, the probability for two random vectors in 64 dimensions to have a cosine similarity above the threshold is negligible [42].

4. Experimental Setup

We evaluate KGFP and established baselines on their ability to predict the correctness of YOLOv8's bounding boxes (IoU > 0.5). Both the object detector and visual foundation model backbones are frozen during all experiments.

4.1. Datasets

COCO 2017 We use only the person class from MS COCO [20]: 64,115 training images and 2,693 validation images with 262,465 and 10,777 person annotations respectively. All other object classes are ignored, as we focus exclusively on detecting failures for safety-critical objects (persons). YOLOv8l is trained on COCO train split. Each image is labeled safe/unsafe based on whether YOLOv8l correctly detects all persons (IoU threshold 0.5, certainty threshold 0.5).

COCO-O Visual Domains Following [10], we evaluate on 6 OOD domains: *cartoon* (artistic renderings), *sketch* (line drawings), *painting* (classical art), *handmake* (crafts/toys), *tattoo* (body art), and *weather* (rain/snow/fog corruptions). These represent diverse visual anomalies ranging from stylistic variations (cartoon, sketch) to environmental corruptions (weather).

4.2. Implementation Details

Model Configuration KGFP uses frozen YOLOv8l [14] and DINO (ViT-B) [3] as feature extractors. YOLOv8l internal features are extracted from FPN levels $\{P3/8, P4/16, P5/32\}$ at 640×640 resolution. DINO [CLS] tokens (768D) are extracted from 518×518 crops. The fusion architecture uses 8 attention heads with 2 self-attention blocks followed by 2 cross-attention blocks in the post-fusion transformer. Dual encoders project refined YOLOv8 and DINO features to a shared 64-dimensional embedding space (ablations test 128D, 256D, 512D). For efficiency, DINO embeddings are pre-computed to Hierarchical Data Format 5 (HDF5) cache, while YOLOv8 features are computed on-the-fly during training.

KGFP is trained end-to-end using Layer-wise Adaptive Rate Scaling (LARS) optimizer [48] with learning rate 0.00095, momentum 0.9, weight decay 0.0009, and LARS eta 0.001. Training runs for 60 epochs with cosine annealing (T_max=60, eta_min=5e-7) and gradient clipping (max norm 1.0) for stability. We use batch size 6.

4.3. Evaluation Metrics

Person Recall @ 5% FPR: *Primary safety metric* - percentage of ground-truth persons detected when accepting images where the safety score exceeds a threshold calibrated to yield 5% FPR on the ID validation split. Crucially, this single threshold is then applied unchanged to all COCO-O domains, simulating realistic deployment where no target-domain labels are available for recalibration. This metric directly quantifies safety: higher person recall means fewer missed persons in accepted images. Remark that if we de-

Table 1. Person Recall [%] among *accepted* images at 5% FPR (selective prediction). Each method acts as a gate that rejects a fraction of images deemed unsafe; Person Recall is measured only on the images the gate accepts. *YOLOv8 (base)* accepts all images (no gating). Best results per column in **bold**.

Method	COCO	COCO-O (Distribution Shift)						COCO-O
	Val	Cartoon	Sketch	Painting	Handmake	Tattoo	Weather	Avg
KGFP (Ours)	**84.5**	**19.3**	29.5	**36.2**	37.0	**11.9**	**71.4**	**34.2**
GRAM	65.4	13.2	26.0	31.0	31.5	9.5	67.8	29.8
KNN	65.1	14.3	32.7	32.8	**38.5**	10.1	68.3	32.8
ViM	65.5	14.9	**34.1**	32.3	34.5	11.2	67.3	32.4
DINO-MLP	69.1	13.6	26.3	33.4	33.1	8.2	65.2	30.0
DINO-ViM	65.2	15.4	34.0	32.1	34.8	7.5	66.6	31.7
YOLOv8 (base)	64.3	13.2	24.9	29.7	31.5	9.5	66.1	29.1

ploy a random selective-prediction gate we expect the Person Recall @ 5% FPR to match the person recall of the original object detector, which serves as a random baseline for the metric.

True Positive Rate (TPR) @ 5% FPR: True positive rate (correct safety predictions on image level) at 5% FPR. Unlike Person Recall, this counts images correctly classified as safe/unsafe, not individual persons detected.

Area Under ROC Curve (AUROC): Area under Receiver Operating Characteristic (ROC) curve for binary safety classification (safe vs. unsafe images). **Rec Area Under Curve (AUC)** and **Prec AUC**: Area under curve between the KGFP module's FPR and YOLO's Person Recall and Person Precision respectively on images predicted to be safe.

4.4. Baselines

In order to demonstrate the effectiveness of visual foundation models in object detector failure prediction, we compare our KGFP against representative OOD detection methods adapted for object detection in safety monitoring. Since these methods were originally designed for image classification, we adapt them to work with YOLOv8l internal features extracted at multiple scales (P3, P4, P5). We train the OOD baseline methods GRAM, KNN and ViM using only safe images from the ID training set, treating unsafe images (where persons are missed) as anomalies in the classical OOD detection framing.

In order to support our proposed angular failure metric, we also compare our KGFP to an MLPs baseline. Therefore, we train an ensemble of MLPs on the DINO embeddings. In our ablations we also train an MLPs head on DINO and YOLO embeddings as a baseline to our KGFP (see Table 3).

Table 2. Comparison with OOD detection baselines. **Safety AUROC** measures binary classification (safe vs unsafe). **Person Rec. AUC** measures YOLOv8 Recall averaged over the FPR of KGFP. **Person Prec. AUC** measures YOLOv8 precision averaged over the FPR of KGFP.

Method	**COCO Val**			**COCO-O Avg**		
	Safety AUROC	Rec. AUC	Prec. AUC	Safety AUROC	Rec. AUC	Prec. AUC
KGFP	**92.9**	**90.4**	**98.2**	**80.3**	**55.8**	**94.7**
GRAM	52.5	65.6	94.1	52.4	31.1	93.4
VIM	53.3	67.5	89.4	66.2	44.4	90.5
KNN	54.3	66.6	90.0	68.9	47.5	92.7

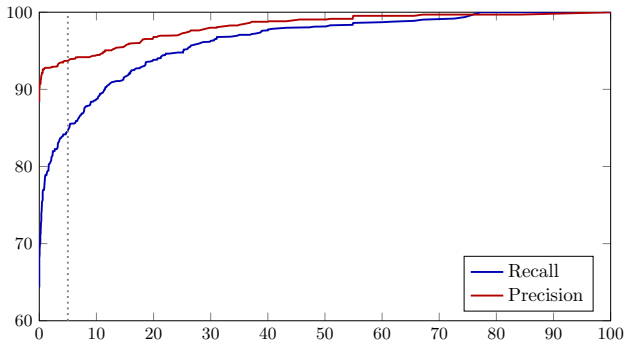

Figure 3. KGFP performance on COCO Val. We plot person recall (blue) and precision (red) on the y-axis [in %] versus KGFP FPR on the x-axis [in %].

5. Results

5.1. Main Results

Table 1 presents our main results. KGFP acts as a selective-prediction gate: it rejects images deemed likely to contain missed persons and does not modify YOLOv8's detections themselves. Among the *accepted* images, 84.5% of ground-

truth persons are correctly detected, compared to 64.3% when accepting all images without gating. On the COCO-O weather domain, gating raises person recall among accepted images from 66.1% to 71.4% at 5% FPR. This corresponds to a relative reduction of person oversights (YOLOv8 false negatives) of 56.6% on COCO Val and 15.6% on the COCO-O weather split at 5% FPR compared to 0% FPR.

We adapt OOD detection methods for failure prediction by training them on safe images only, treating unsafe images as OOD. On ID data, GRAM, KNN, and ViM achieve 65.1–65.5% recall, only slightly above the YOLOv8 baseline (64.3%). DINO-based variants (DINO-MLP: 69.1%, DINO-ViM: 65.2%) achieve higher recall but remain below KGFP. KGFP outperforms the best baseline by +15.4 percentage points, indicating that the dual-encoder architecture with explicit failure supervision is more effective than treating failures as OOD samples.

On COCO-O, KGFP achieves the **highest average person recall (34.2%)** across all six COCO-O domains. Note that OOD baselines were trained to treat unsafe images as OOD samples (see Section 4.4), so they face a more challenging task on COCO-O: distinguishing unsafe from safe images under genuinely novel visual conditions. KGFP performs notably better on *weather* corruptions (71.4%), maintaining near-ID performance. On *painting* (36.2%) and *handmake* (37.0%), KGFP leads or matches the best baselines. These results indicate that the learned alignment between YOLOv8 and DINO representations generalizes to novel visual domains.

Table 2 provides complementary area-under-curve metrics. KGFP achieves 92.9% Safety AUROC for binary safe/unsafe classification and 90.4% Person Rec. AUC on COCO Val, compared to 54.3% Safety AUROC for the best baseline (KNN). On COCO-O, KGFP maintains 80.3% Safety AUROC and 55.8% Person Rec. AUC on average across all novel visual domains.

5.2. Ablation Studies

Table 3 (left) analyzes architectural components. Removing all attention modules reduces ID recall by 2.3% (84.5% $\rightarrow$ 82.3%). Pre-fusion attention and post-fusion cross-attention each contribute $\sim$1% to ID performance, while post-fusion self-attention has negligible impact. Replacing cosine similarity with an MLP similarity head degrades Rec@5%FPR by 0.8%. Mean cosine similarity remains below 0.9 throughout training, indicating no embedding collapse; deeper DINO embedding MLPs (7+ layers) collapsed in preliminary experiments.

Table 3 (right) reports results across embedding sizes $d \in \{64, 128, 256, 512\}$ and optimizers. The 64D embedding performs best on both COCO Val (84.5%) and COCO-O (34.2%) with only 2.6M parameters. Larger embeddings show diminishing returns, with clear overparameterization at 512D. LARS slightly outperforms Adam [16] at 64D, while Adam performs poorly at 256D. Table 3 (center) compares world-knowledge encoders. DINO ViT-B/16 (86M parameters) achieves the highest ID recall (85.1%) with the fewest parameters, outperforming CLIP ViT-L/14 (84.5%, 427M parameters), DINOv2 ViT-L/14 and SigLIP ViT-B/16. We attribute this to DINO's self-distillation objective producing spatially coherent attention maps that capture fine-grained visual cues as partial occlusions, atypical poses, unusual lighting [3], whereas CLIP's language-aligned features are shaped by caption-level semantics [28] and thus less sensitive to these sub-textual failure patterns.

5.3. Discussion

Supervised vs. unsupervised. KGFP substantially outperforms all baselines on ID data. However, KGFP is supervised (trained with safe/unsafe labels) while the OOD baselines are unsupervised (fitted on safe images only). A fairer comparison is with the supervised DINO-MLP and the MLP-head ablation (Table 3), which use the same labels. KGFP outperforms both, indicating that the dual-encoder cosine-similarity formulation provides a stronger inductive bias for failure prediction than direct classification, despite having fewer trainable parameters.

Role of DINO features. Methods based solely on DINO features (DINO-MLP, DINO-ViM) perform worse than KGFP, confirming that DINO embeddings alone cannot reliably predict YOLOv8 failures—the fusion of both feature sources is essential. DINO provides complementary semantic context (scene-level patterns, occlusion cues) that is correlated with detector failure but not directly accessible from YOLOv8's task-specific features.

OOD generalization. The unsupervised baselines show near-baseline performance on ID data, which is expected: failures on ID images do not necessarily correspond to distributional novelty. On COCO-O, unsupervised baselines achieve competitive or superior performance on specific domains—KNN on handmake (38.5% vs. 37.0%) and ViM on sketch (34.1% vs. 29.5%)—because genuine distribution shift correlates with detector failure in these stylistically extreme domains. However, KGFP achieves the highest average recall across all six domains, demonstrating more consistent generalization.

Embedding dimensionality. The degradation at 256D and 512D (Table 3, right) is attributable to overparameterization: larger embedding spaces require more data to learn meaningful angular structure, and cosine similarity becomes less discriminative at high dimensionality because random vectors concentrate around orthogonality [42]. The 64D space provides sufficient capacity while maintaining a well-structured angular decision boundary.

Table 3. Ablation studies: (left) architecture components, (center) foundation model comparison, (right) embedding dimensions and optimizer. All values are Person Recall @ 5% FPR. ID: COCO Val; OOD: average over six COCO-O domains. Architecture abbreviations: pre-fn = pre-fusion, post-fn = post-fusion, attn = attention.

Attention Ablation Head Ablation	ID	OOD
Full KGFP	**84.5**	**34.2**
No pre-fn attn	83.4	31.0
No post-fn cross-attn	83.4	32.5
No post-fn self-attn	84.2	34.1
No attn	82.3	31.1
MLP head	83.7	34.2

World-Knowledge Model (Parameter)	ID	OOD
DINO (86M)	**85.1**	**34.2**
SigLIP (87M)	84.8	35.2
DINOv2 (304M)	84.4	33.9
CLIP (427M)	84.5	32.6

Embedding Dimension, Optimizer	ID	OOD
64D, Lars	**84.5**	**34.2**
64D, Adam	84.3	32.8
128D, LARS	83.1	33.1
256D, LARS	83.5	31.1
256D, Adam	65.1	29.3
512D, LARS	67.1	29.1

6. Limitations and Future Work

KGFP requires frozen foundation model embeddings. Changes to the foundation model or encounters with visual domains far beyond its training data could degrade performance. Future work should investigate continual adaptation of world-knowledge encoders. The all-or-nothing safety label (all persons detected vs. any missed) may be too coarse for some applications. Extensions could predict expected miss counts or provide spatial failure localization. We focus on person detection for clear safety implications. Generalization to multi-class scenarios requires rethinking the safety formulation—which objects are safety-critical? KGFP requires computing both YOLOv8 and DINO forward passes, which introduces significant overhead. Latency is critical in real-time systems. Future work should investigate the latency of our architecture, how it could be reduced, and what the trade-offs would be. The Latency is mainly caused by the backbones and the classification head introduces negligible latency. Frame-skipping strategies (running safety checks every N frames) can amortize costs for video streams, or deployment may benefit from model compression, distillation, or efficient foundation model variants. We evaluate on naturally occurring visual anomalies. Adversarial perturbations designed to fool both YOLOv8 and DINO could evade safety monitoring. Adversarial training or certified defenses warrant investigation.

7. Conclusion

We introduced KGFP, a runtime monitoring framework that detects unsafe images where a specific object detector misses safety-critical objects by measuring semantic alignment between learned encodings of object detector's internal activations and DINO embeddings. The architecture fuses multi-scale YOLOv8 features via pre-fusion cross-scale attention, then applies post-fusion cross-attention with DINO for semantic alignment measurement. KGFP provides actionable failure signals for safety-critical deployment. Focusing on safety-critical object classes (persons for pedestrian detection), KGFP achieves 85% person recall at 5% FPR on both ID data and novel visual domains, substantially outperforming traditional OOD detection methods and our angular failure metric outperforms an MLP baseline with fewer parameters. Comprehensive ablations reveal that foundation model world-knowledge, cross-attention fusion, and angular metric learning are critical for robust failure prediction. As object detectors proliferate in autonomous vehicles, surveillance, and healthcare, explicit failure monitoring becomes essential. KGFP demonstrates that foundation models trained on billions of images can serve as semantic "sanity checks" for task-specific detectors, identifying unsafe images before they cause harm.

Broader Impact and Disclosures

Ethics Statement. KGFP is designed to enhance safety by detecting object detector failures. It functions as an additional safety layer, not a replacement for robust model development. Limitations include potential false negatives or positives; users must carefully consider recall-precision trade-offs and evaluate performance across relevant demographic groups to mitigate potential dataset biases.

LLM Usage. We acknowledge the use of Claude 3.5 Sonnet and GitHub Copilot for writing refinement, coding assistance, and literature summarization. All novel algorithmic components, experimental designs, and scientific conclusions are the sole work of the human authors. The authors take full responsibility for all content, including any errors or inaccuracies that may have been introduced during LLM-assisted editing.

References

[1] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition Workshops (CVPRW), pages 1578–1587, 2022. 3

[2] Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. A survey on visual anomaly detection: Challenge, approach, and prospect, 2024. 3

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 2, 3, 4, 5, 7

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 3

[5] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[6] Raul Sena Ferreira, Frédéric Guérin, Karine Delmas, Jérémie Guiochet, and Hélène Waeselynck. Safety monitoring of machine learning perception functions: A survey. *arXiv preprint arXiv:2412.06869*, 2024. 3

[7] Raphael Hagmanns, Peter Mortimer, Miguel Granero, Thorsten Luettel, and Janko Petereit. Excavating in the wild: The goose-ex dataset for semantic segmentation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025. 1

[8] Weicheng He, Changshun Wu, Chih-Hong Cheng, Xiaowei Huang, and Saddek Bensalem. Mitigating hallucinations in yolo-based object detection models: A revisit to out-of-distribution detection. *arXiv preprint arXiv:2503.07330*, 2025. 3

[9] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017. 2

[10] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, 2021. 1, 5

[11] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, page 340–353, Berlin, Heidelberg, 2012. Springer-Verlag. 1, 3

[12] International Electrotechnical Commission. IEC 61508: Functional safety of electrical/electronic/programmable electronic safety-related systems. Standard IEC 61508, International Electrotechnical Commission, Geneva, Switzerland, 2010. Parts 1-7. 2

[13] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19606–19616, 2023. 3

[14] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8, 2023. Version 8.0.0. 4, 5

[15] Daniel Kang, Deepti Raghavan, Peter Bailis, and Matei Zaharia. Model assertions for monitoring and improving ml models. In *Proceedings of Machine Learning and Systems*, pages 481–496, 2020. 2, 3

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015. 7

[17] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 7167–7177, Red Hook, NY, USA, 2018. Curran Associates Inc. 2

[18] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks, 2020. 2

[19] Tsung Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 936–944, 2017. 2, 3

[20] 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 (ECCV)*, pages 740–755, 2014. 5

[21] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 2

[22] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14234, 2021. 3

[23] Maximilian Müller and Matthias Hein. Mahalanobis++: Improving OOD detection via feature normalization. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 45151–45184. PMLR, 2025. 2

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. 3, 4

[25] Malaisree P, Youwai S, Kitkobsin T, Janrungautai S, Amorndechaphon D, and Rojanavasu P. Dino-yolo: Self-supervised pre-training for data-efficient object detection in civil engineering applications, 2025. 3

[26] Jaewoo Park, Jacky Chen Long Chai, Jaeho Yoon, and Andrew Beng Jin Teoh. Understanding the feature norm for out-of-distribution detection. In *Proceedings - 2023 IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 1557–1567, United States, 2023. Institute of Electrical and Electronics Engineers Inc. 2

[27] Julius Pfrommer, Thomas Usländer, and Jürgen Beyerer. Ki-engineering–ai systems engineering: Systematic development of ai as part of systems that master complex tasks. *at-Automatisierungstechnik*, 70(9):756–766, 2022. 1

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2, 3, 7

[29] Deepak Ravikumar, Efstathia Soufleri, and Kaushik Roy. Improved out-of-distribution detection with additive angular margin loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3464–3471, 2025. 3

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 3

[32] Chandramouli S. Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *Proceedings of the 37th International Conference on Machine Learning*. JMLR.org, 2020. 2

[33] Albert Schotschneider, Svetlana Pavlitska, and J. Marius Zöllner. Runtime safety monitoring of deep neural networks for perception: A survey, 2025. 2, 3

[34] Poulami Sinhamahapatra, Franziska Schwaiger, Shirsha Bose, Huiyu Wang, Karsten Roscher, and Stephan Gunnemann. Finding dino: A plug-and-play framework for zero-shot detection of out-of-distribution objects using prototypes. In *Proceedings - 2025 IEEE Winter Conference on Applications of Computer Vision, WACV 2025*, pages 8474–8483. Institute of Electrical and Electronics Engineers Inc., 2025. 3

[35] Rakshith Subramanyam, Kowshik Thopalli, Vivek Narayanaswamy, and Jayaraman J. Thiagarajan. Decider: Leveraging foundation model priors for improved model failure detection and explanation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXIX*, page 465–482, Berlin, Heidelberg, 2024. Springer-Verlag. 3

[36] Rakshith Subramanyam, Kowshik Thopalli, Vivek Sivaraman Narayanaswamy, and Jayaraman J. Thiagarajan. DECIDER: leveraging foundation model priors for improved model failure detection and explanation. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXIX*, pages 465–482. Springer, 2024. 3

[37] Jin Sun and David W. Jacobs. Seeing what is not there: Learning context to determine where objects are missing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[38] Yiyou Sun, Chuan Guo, and Yixuan Li. React: out-of-distribution detection with rectified activations. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2021. Curran Associates Inc. 2

[39] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 2

[40] Birk Torpmann-Hagen, Michael A. Riegler, Pål Halvorsen, and Dag Johansen. Runtime verification for visual deep learning systems with loss prediction. *IEEE Access*, 13:48502–48519, 2025. 3

[41] Duong Dinh Tran, Takashi Tomita, and Toshiaki Aoki. Safety analysis of autonomous driving systems: A simulation-based runtime verification approach. *IEEE Transactions on Reliability*, 74(4):4574–4588, 2025. 3

[42] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. 4, 5, 7

[43] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[44] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 2

[45] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024. 2

[46] Qinghua Yang, Hui Chen, Zhe Chen, and Junzhe Su. Introspective false negative prediction for black-box object detectors in autonomous driving. *Sensors*, 21(8):2819, 2021. 2, 3

[47] Hakan Yekta Yatbaz, Ennio Poli, Sergio Capobianco, and Giorgio Di Natale. Run-time monitoring of 3d object detection in automated driving systems using early layer neural activation patterns. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 3

[48] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. 5

[49] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6995–7003, 2018. 2

[50] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[51] Alon Zolfi and Shai Avidan. Yolood: Utilizing object detection concepts for multi-label out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3603–3612, 2024. 3