Robust Onion: Peeling Open Vocab Object Detectors Under Noise

Priyank Pathak*, Mukilan Karuppasamy*†, Aaditya Baranwal,
Shruti Vyas, and Yogesh S Rawat

UCF Institute of Artificial Intelligence, University of Central Florida (UCF)
{priyank,aaditya.baranwal,shruti,yogesh}@ucf.edu, mukilan.nitt@gmail.com

Project Page: https://ucf-crcv.github.io/RobustOnion/
Github: https://github.com/ppriyank/RobustOnion

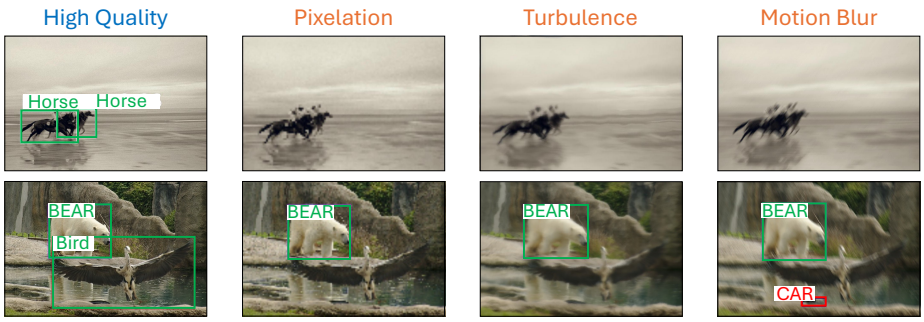
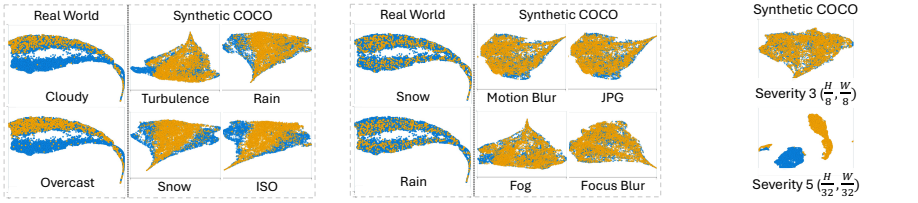


Fig. 1: Effect of Noise: GLIP [28] (above) & MM-GDINO [69] (bottom) performance on COCO [30] for noises like turbulence, pixelation, and motion blur.

Abstract. The impact of real-world noise on **O**pen **V**ocabulary **O**bject **D**etectors (OV-ODs) remains poorly understood due to their architectural complexity. We present our comprehensive analysis, **Robust Onion**, an empirical study that uses controlled synthetic visual degradations to *peel* OV-ODs layer-by-layer, revealing **how, why**, and **where** robustness degrades, systematically analyzing *feature collapse*. Our findings reveal that models with similar vision backbones exhibit comparable robustness, driven by similar feature collapse at similar layers, while factors such as pretraining strategy, architectural nuances, and caption supervision contribute little. Robustness is primarily governed by the image domain rather than annotations, explaining the similar robustness impact on COCO and LVIS, and why datasets like ODinW-13 can give an impression of inflated robustness due to large, isolated objects. Finally, we validate our insights by *improving robustness* on real-world BDD-100K, WiderFace, and VisDRONE via our lightweight plug-and-play **NN & TK0** approach, using 96× fewer trainable parameters than end-to-end training. We also **explain the prior works'** robustness observations.

Keywords: Explainability · Interpretability · Object Detectors · Noises

*Equal contribution † work done as intern



(a) Observable Collapse **(b) Minimum** collapse **(c)** Pixelation Severity

Fig. 2: Real-World & Synthetic: GLIP-T synthetic COCO noisy features collapse against clean image *aligns* with BDD-100K real-world collapse. BDD-100K doesnt have explicit clean images, hence all noisy categories are in blue, except the highlighted.

1 Introduction

Vision Language Models (VLMs) have shown strong generalization in tasks like image-to-text retrieval [43,48], open-vocabulary classification [1], image captioning [11], visual-question-answering [24] *etc.* The ability to adapt without fine-tuning makes VLMs highly lucrative for applications where *zero-shot* is not just a convenience but a necessity. VLM based **O**pen **V**ocabulary **O**bject **D**etectors (**OV-ODs**) have gained attention for their easy adaptability in security [23], medical imaging [63], environmental monitoring [57], and self-driving cars [53].

Real-world deployment of OV-ODs requires a critical understanding of their robustness against visual distortions and noise. However, their enormous complexity from various moving parts like vision backbone, text backbones, fusion network, box predictors, alignment networks *etc.*, makes them one of the most obscure deep learning models. Additionally, studying and isolating the true impact of low quality (LQ) noise requires matching high quality (HQ) counterpart images; and such *well-annotated 'noise-paired' real-world datasets* are nearly impossible to find. As a result, despite their growing use, the effect of out-of-distribution noise on VLM based OV-ODs remains largely underexplored.

Addressing the gap in robustness against real-world distortions (*e.g.* BDD-100K [62]), our novel analysis, **Robust Onion** broadly categorizes noise-induced **feature (variance) collapse** [3, 8, 31, 41] into two categories: **1. Observable** (Fig. 2a): noisy features form distinct clusters separate from clean HQ features (*e.g.* cloudy, overcast), **2. Minimal** (Fig. 2b): little or no observable collapse (*e.g.* snow, rain). By carefully tuning synthetic noises, we mimic these feature collapses, providing a *practical proxy* for real-world degradations (*e.g.* turbulence for observable, motion blur for minimal). Increasing noise severity can change the collapse from minimal to observable (Fig. 2c). Robust Onion then empirically *peels* each component of OV-OD under these controlled noises.

To answer: *"How do visual distortions impact complex models such as OV-ODs?"* and *"What are the most effective directions to improve their robustness?"*, We present an interpretability-driven analysis of robustness in OV-ODs. Rather than simply evaluating performance, our goal is to explain why failures occur, where they originate in the architecture, and how robustness can be improved.

Our analysis revolves around five key questions: **(1)** Do bells and whistles like pretraining, finetuning, and architecture modules impact the robustness? **(2)** Which part of the model architecture (Backbone, FPN, RPN, etc) induces the robustness? **(3)** Are larger models inherently more robust, or are other factors decisive? **(4)** Is robustness solely determined by the model, or do input images play a role? **(5)** Does language prompting improve robustness under visual noise?

Our analysis frames robustness as an explainable phenomenon rather than a black box outcome. We peel the layers of OV-ODs to reveal the following:

- **Vision Backbone drives robustness:** Similar backbones exhibit comparable robustness due to resembling feature collapses at *similar depths*, regardless of bells and whistles of pretraining or overall architecture.
- **Layer-wise robustness:** Shallow layers are adversely affected by the noise. Further, *cross-exchanging the backbone layer* can improve robustness.
- **Dataset bias:** ODinW-13 can give a false impression of robustness because it features large and isolated objects.
- **Domain over annotation:** Detecting *on* what (domain) matters more than detecting *what* (annotation), thus similar robustness of COCO & LVIS.
- **Minimal language influence:** Once the visual features are degraded, language/captions contribute little to recover lost robustness.

Each analysis includes **[Takeaways & Model Design]** highlighting key actionable insights for designing **robust real-world OV-OD**. We conclude by validating our analysis on real-world autonomous driving datasets like BDD-100K, DAWN, Foggy Cityscapes, and Virtual KITTI 2, WiderFace, and VisDRONE via our **NN & TK0** approach, a lightweight technique to induce robustness in object detectors (open vocabulary). **NN & TK0** uses 96× fewer trainable parameters than end-to-end training with similar robustness. Our analysis also explains several of the previously published robustness-related observations.

2 Related Work

VLMs and OV-ODs: VLM generalizability [38, 47] has progressed into object detection [19, 51, 65]. Advances in large-scale multimodal training, have further enhanced open-vocabulary object detectors (OV-OD) [2, 9, 39, 54, 68]. Versatility of OV-ODs [16, 27] makes them ideal for real-world (*e.g.* surveillance [15], industrial-inspection [4], satellite imagery [44], autonomous-driving [53], *etc*; thus understanding their limitations [3, 6, 66] against noise is crucial.

Robustness against Noise: Weather (*e.g.* rain, fog, snow) & common artifacts (*e.g.* jpg compression) pose significant challenges in object detection [13, 34, 46, 61, 67]. These distortions lose discriminative features, fundamentally affecting almost all models [52]. Robustness methods often use synthetic noises to improve on real-world tasks, like person re-id [42], self-driving in fog/rain [21]. Despite the prevalence of noises, their impact on VLMs remains largely unexplored [12,29], with most recently LR0.FM [41] analyzing robustness of VLMs in

image classification. On the contrary, OV-ODs are far more complex than simple VLMs. Here, we present a comprehensive analysis of SOTA OV-ODs, revealing bottlenecks and critical factors in their robustness. More works in Tab. 3.

3 Analysis Setup

Models: We analyze 6 publicly available OV-ODs: RegionCLIP [70] (RC, RCx4), GLIP [28], FIBER [17], MM-Grounding-DINO [69] (MM-GDINO), GLEE [56], and YOLO-World [61] (YOLO). Fig. 3 shows robustness of all models. CNN-based ones (YOLO & RegionCLIP) are not as robust as transformer ones (unless fine-tuned), hence, **we mainly focus of our analysis on transformers**.

Datasets: Robustness is evaluated on 3 benchmarks: COCO [30] (val2017), LVIS [20] (miniVal) and ODinW-13 [28] (set of 13 datasets). COCO (80 categories) and LVIS (1,203 categories) have *same images, but different annotations*. Language analysis uses RefCOCO/+/g [25, 33, 64], and Flickr30k [45]. Model-related insights are validated on **real-world** driving datasets: BDD-100K [62], DAWN [26], Foggy Cityscapes [14, 49], Virtual KITTI 2 [7], WiderFace [59], and VisDRONE [72].

Framework: Fig. 4 shows the general framework of OV-ODs. Image and captions are processed through a pyramidal multi-scale vision encoder (*e.g.* ResNet, Swin Transformer) and a text encoder (*e.g.* BERT, CLIP), respectively. Vision features are enhanced via FPN (or pixel-decoder) followed by cross-self-attention fusing text and vision embedding. These fused features predict bounding boxes, confidence scores, and class labels.

Noises: Analyzing noise requires measuring the drop in performance *relative* to clean features *i.e.* analysis of LQ-HQ image pairs (absent in real-world). Instead, we sample one controlled synthetic noise from each feature collapse

Models	HQ	Pix.	Robustness =1−(HQ−Pix)/HQ
GLEE-Lite-pretrain-Stage 1	42.59	23.00	0.54
GLEE-Plus-pretrain-Stage 1	44.00	36.17	0.82
GLEE-Pro-pretrain-Stage 1	50.83	44.73	0.88
GLEE-Lite-joint-Stage 2	54.96	28.39	0.52
GLEE-Plus-joint-Stage 2	60.44	41.88	0.69
GLEE-Pro-joint-Stage 2	61.96	52.82	0.85
GLEE-Lite-scaleup-Stage 3	53.70	26.90	0.50
GLEE-Plus-scaleup-Stage 3	60.34	42.20	0.70
GLEE-Pro-scaleup-Stage 3	61.71	51.80	0.84
GDINO-T Swin-T (O_G_CAP4)	48.50	29.30	0.60
MM-GDINO-T (O_G)	50.40	29.20	0.58
MM-GDINO-T (O_G_GR)	50.50	30.60	0.61
MM-GDINO-T (O_G_V)	50.60	29.30	0.58
MM-GDINO-T (O_G_GR_V)	50.40	30.10	0.60
MM-GDINO-B (O_G_V)	52.50	35.70	0.68
MM-GDINO-B* - ALL	59.50	40.10	0.67
MM-GDINO-L	53.00	37.60	0.71
MM-GDINO-L* - ALL	60.30	41.70	0.69
FIBER-B	49.47	30.80	0.62
FIBER-B*-COCO-FT	58.40	36.60	0.63
FIBER-B*-LEVIS-FT	50.70	31.70	0.63
FIBER-B*-RefCOCO	15.50	10.90	0.70
FIBER-B*-RefCOCO+	18.00	12.30	0.68
FIBER-B*-RefCOCOg	22.70	15.60	0.69
GLIP-T (A)	42.90	20.20	0.47
GLIP-T (B)	44.90	22.30	0.50
GLIP-T (C)	46.70	25.70	0.55
GLIP-T [5]	46.60	26.20	0.56
GLIP-L [7]	51.23	34.20	0.67
YOLO-Worldv2-S-640	37.50	5.30	0.14
YOLO-Worldv2-M-640	42.80	9.20	0.21
YOLO-Worldv2-L-640	45.40	10.10	0.22
YOLO-Worldv2-L-640-LITE	45.10	8.50	0.19
YOLO-Worldv2-L (CLIP-L🔥)-640	46.00	11.70	0.25
YOLO-Worldv2-X-640	46.70	9.60	0.21
YOLO-Worldv2-XL-640	47.50	12.60	0.27
RegionCLIP R50 (RC)	58.18	20.10	0.35
RegionCLIP R50x4 (RCx4)	62.40	28.90	0.46
RC-COCO-FT	75.30	41.60	0.55
RCx4-COCO-FT	80.00	49.80	0.62
RCx4 Fully80-COCO-FT	88.77	56.86	0.64
RC-LVIS-FT	80.00	43.00	0.54
RCx4-LVIS-FT	82.70	47.80	0.58
RCx4 Fully123-LVIS-FT	82.38	47.55	0.58

Fig. 3: Models vs Pixelated COCO (mAP): Shade ∝ robustness. Fine-tuned (COCO, LVIS, RefCOCO) in **bold**.

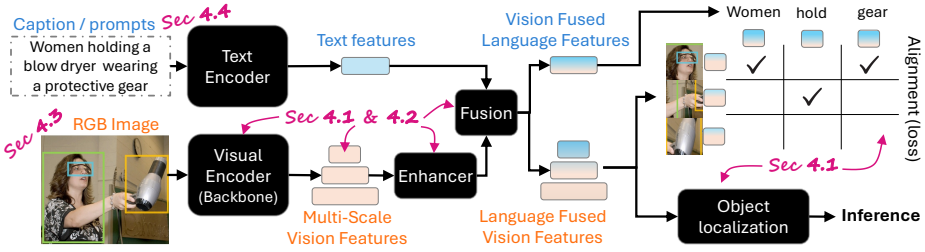


Fig. 4: OV-OD Overview: Architecture (black), may include additional components and losses. Fusion of text features with multi-scale vision features via self-attention, cross-exchanges text-vision modality information. The role of each component in robustness against noise is described in listed sections. The vision feature enhancer (neck) is commonly referred as FPN / pixel decoder. Image modified from GLIP [28].

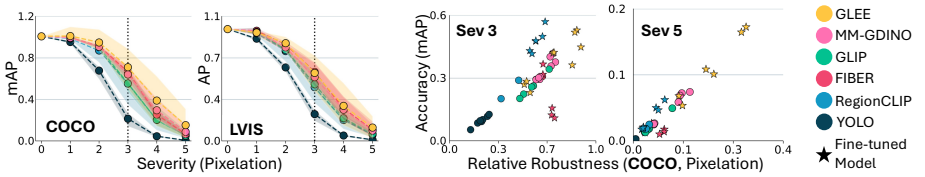


Fig. 5: *(Left)* **Performance decline w/Severity:** Models start dropping performance around severity 3. Shade shows all model variants, with solid indicating mean accuracy. *(Right)* **Accuracy and Robustness linearity:** Accuracy (zero-shot) forms an $\sim$ linear relationship with robustness; preserving the relative ranking of models.

category (Fig. 2): Turbulence (*observable* collapse), and Motion Blur (*minimal* collapse). We also use Pixelation (*e.g.* compression, distant objects) for severity (intensity) analysis. Pixelation is simulated [41] by downsampling the image and upscaling it back via bicubic interpolation[1]. Turbulence (*e.g.* hot air) is simulated via pre-trained GAN [36]. Motion Blur [22] simulates motion (*e.g.* videos). These noises are only applied to the input image, leaving *textual captions unchanged*.

Evaluation: Metrics include AP (LVIS), mAP (COCO), AP_{avg} (ODinW-13), and Recall@1 (Flickr30K). We shall use relative robustness (denoted as 'robustness') [10,50] as the key metric for measuring a model's robustness against noise. Relative Robustness $= 1 -$ (Drop in Accuracy/Accuracy) $= 1 - (Acc_{Clean} - Acc_{Noise})/Acc_{Clean}$. Here, Acc_{Clean} and Acc_{Noise} denote accuracy on original and noisy images. Relative Robustness is independent of absolute performance, enabling cross-model, and cross-dataset comparisons. Higher severity (4, 5) risks random predictions ($Acc_{Noise} \simeq 0$, Fig. 5 (left)) We also observe a linear relationship between absolute accuracy and relative robustness (Fig. 5 (right)), with outliers being mostly fine-tuned models, *e.g.* RegionCLIP (★, accuracy ↑ & robustness ↑), FIBER-B (★ accuracy ↓ & constant robustness).

[1] Severity 's': $(H, W) \rightarrow (\frac{H}{2^s}, \frac{W}{2^s}) \rightarrow (H, W)$. Pixelated images $\neq$ low resolution (*e.g.* $(\frac{H}{2^3}, \frac{W}{2^3})$ can still be high resolution $(256, 256)$).

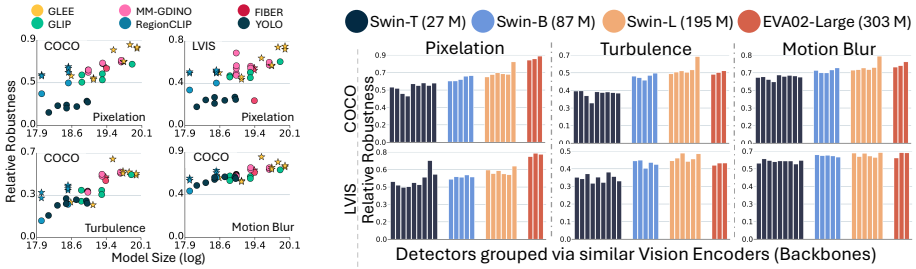


Fig. 6: *(Left)* **Robustness vs Size:** Larger models are more robust (+ve correlation with size); GLEE transformers > GLEE ResNet robustness. *(Right)* **Similar vision backbones:** Performance remains relatively consistent across models with similar backbones and depths. EVA-02 (303M) $\simeq$ Swin-L (195M) robustness, both 24-blocks.

4 Analysis and Explainability (Where, Why, What)

In our analysis visuals, y-axis will represent relative robustness (unless indicated otherwise); insights in highlight; fine-tuned model as ★; **[Takeaways & Model Design]** describes practical robust OV-OD design using insights.

4.1 WHERE: Model-Based Analysis

Fig. 6 left) shows a strong positive correlation between robustness and model size (entire detector + text backbone) with pearson correlation on COCO / LVIS: 0.68 / 0.66 for pixelation, 0.78 / 0.72 for turbulence, and 0.77 / 0.70 for motion blur[2]. Large Transformer detectors consistently outperform CNNs (< 19.4 size).

Fig. 6 right) groups detectors on similar vision backbones (irrespective of other architectural components like enhancers, fusion, decoders, text backbones *etc.*), revealing: 1) Models with **similar backbones show similar robustness** despite different architectures ($\frac{3}{23}$ exceptions). 2) Larger backbones are almost always robust (*e.g.* EVA-02). However, in general (turbulence & motion blur), deeper transformers are slightly more robust than shallow ones, while size **not** playing a crucial role, *i.e.* ResNet < Swin-T (12 blocks, *27M*) < Swin-B (12 blocks, *87M*) $\leq$ Swin-L (24 blocks, *195M*) $\simeq$ EVA-02 (24 blocks, *303M*), where $\simeq$ indicates similar robustness. Similar trend observed on other noises like Pixel drop, ISO, Salt-pepper, JPG compression, and Fog (*Supplementary*).

Fig. 7 left) shows pre-training dataset size weakly affects the zero-shot robustness; indirectly meaning different pre-training datasets (& their count). For example, GLIP shows steady robustness across different pretraining sizes (green dots). Pearson's correlation on COCO / LVIS: 0.35 / 0.23 for pixelation, 0.41 / 0.39 for turbulence, and 0.42 / 0.22 for motion blur.

Fig. 7 right) shows clean COCO and LVIS fine-tuning improves (noisy COCO / LVIS) RegionCLIP (ResNet) robustness over zero-shot, while FIBER-B (transformer) is minimally impacted. Overall, fine-tuning is **not** a universal robustness

[2] Pearson correlation < 0.3 is none / weak and moderate for $[0.3, 0.7]$

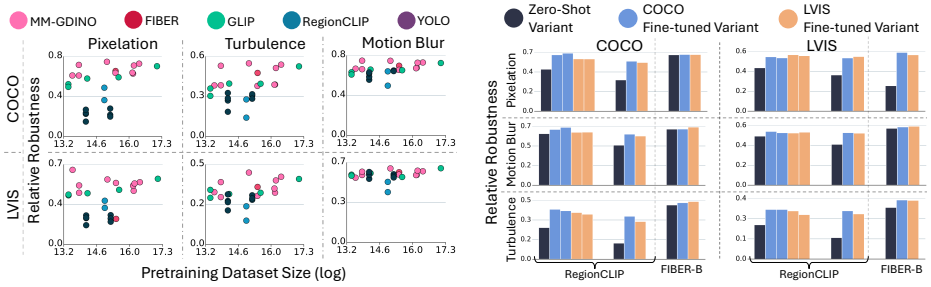


Fig. 7: *(Left)* **Zero-shot Pretraining**: GLIP (green) consistent robustness regardless of pretraining size (no clear correlation). *(Right)* **Fine-tuning Impact**: Identical impact of COCO and LVIS finetuning on COCO with synthetic noise evaluation.

boost [10]. Interestingly, COCO and LVIS (same images, different annotations) have a similar impact on robustness, *i.e.* impact of fine-tuning is mostly governed by **domain of images**, with annotation playing a minimal role. This importance of image domain (over annotation) will re-emerge in Sec. 4.3.

[Takeaways & Model Design] We observe that robustness is driven primarily by backbone architecture and depth. Large models like Swin-L and EVA-02-L are more robust than smaller ResNets, even under identical training and frameworks (GLEE transformer > GLEE ResNet). Models sharing similar backbones exhibit similar robustness, indicating that backbone choice matters far more than *auxiliary modules* (*e.g.* MM-GDINO/GLEE decoder, neck/FPN network, *etc.*), and *extensive pre-training* (*e.g.* MM-GDINO pre-trained on 9 datasets, and GLEE on 18 in three stages of pre-training, and GLIP via two stages *etc.*). Explicit noise-robust training is likely to have a stronger impact on robustness than simple pretraining and fine-tuning. Given resource constraint environments, Swin-B can be an amazing alternative to EVA-02 (4x bigger), and Swin-L (2x bigger), with *size playing a limited role* for sufficiently large backbones.

4.2 WHY: Insight into Transformers: GLIP, MM-GDINO & GLEE

Fig. 6 showed consistent robustness for Swin-L in GLIP, MM-GDINO, and GLEE, with a maximum Δ in robustness of 0.15/0.08 in COCO/LVIS across all noises. This narrows the analysis to the vision backbone. Fig. 8 illustrates the UMAP [37] of the last 4 layers of the backbone, last layer for the feature enhancer, and fusion network. Ideally, robust models shouldn't distinguish between severities, *i.e.* sev 5 noisy features identical to sev 0 (HQ), *i.e. robustness* $\approx$ *features overlap*. *Supplementary* details further on t-SNE plots and other noises.

1) Backbone (Layers): Deeper layers (#3,#4) have a higher features overlap across severities that the shallow ones (#1,#2), *i.e.* shallow layers are more vulnerable to noise. Additionally, 2$^{\text{nd}}$ and 4$^{\text{nd}}$ block (#1,#2) for all models have similar feature overlap (despite architectural differences), *i.e.* **same depth** have

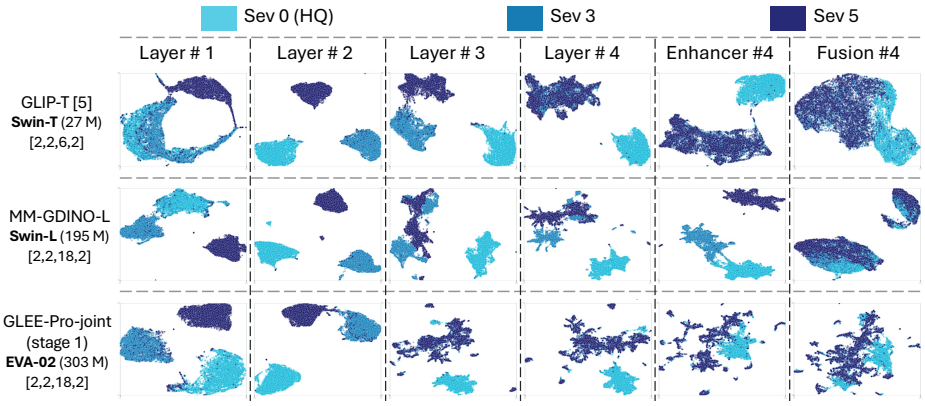


Fig. 8: Pixelation UMAP: Vision backbone, enhancer, and language fused features shown for sev 5 (dark) and sev 0 (lighter). n-th layer features represented by '#n', *e.g.* layer #4 is 24-th block in Swin-L & EVA-02, and 12-th in Swin-T (blocks in each layer shown on left). Feature overlap between sev 5 and sev 0 (similar) implies robustness.

similar feature collapse. This helps explain why similar-depth backbones exhibit similar robustness (Swin-L $\simeq$ EVA-02-L, both have 24-block depth).

2) Enhancer: Enhancers are simple conv block on top of vision backbone (Fig. 4); produces similar feature overlap across severities as that of backbone features (layers #4). This indicates a minimal contribution to robustness.

3) Fusion: Fusion cross-exchanges multi-scale vision features ($192 \times 192, 96 \times 96, 48 \times 48, 24 \times 24$) with language. Although language does not significantly affect robustness (Sec. 4.4), **cross exchanging information across all vision layers induces robustness**, as evidenced by the significant overlap of features.

[Takeaways & Model Design] Robustness in OV-ODs, which typically *consist of 3-5 transformers, can be reliably enhanced by simply focusing on the vision backbone.* Techniques that enhance robustness, particularly via shallow layer modifications [41], can be viable, as early layer features are more vulnerable to noise. Encouraging cross-layer information exchange within the vision backbone, such as self-attention or non-local [55] modules, can further enhance robustness. Although, this insight is derived from synthetic noise, **we show its real-world impact** in Sec. 5. Layers at **similar depth tend to exhibit similar feature collapse**, which partly explains why similar backbones have similar robustness.

4.3 WHAT: Robustness as a function of Dataset

Fig. 9 a) illustrates larger ($\geq 96 \times 96$) object detection is more robust to noise than the smaller ones ($\leq 32 \times 32$). Fig. 9 b) illustrates that almost all **detectors are highly robust when there is only one object to detect**. As the number of objects/image goes above 3, robustness starts to see a drop, eventually saturating around 10 or more objects/image. The jitter in robustness after 25

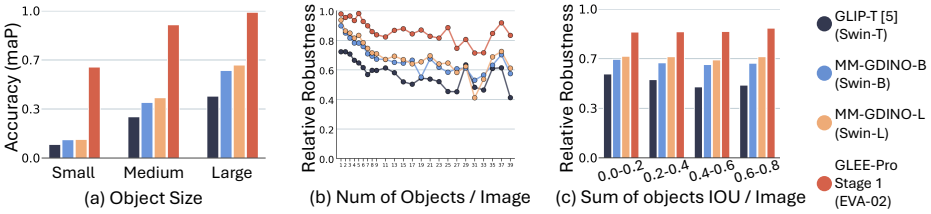


Fig. 9: *(a)* **Object size:** Larger objects are more robust. *(b)* **Object Count:** Models are very robust when they have to detect <3 objects. Jumping accuracy after >25 objects is likely due to very few samples in that range for averaging to smooth out. *(c)* **Occlusion:** Robustness pretty much is unaffected with degrees of overlap between objects. All experiments on COCO for pixelation. Other noises in *Supplementary*.

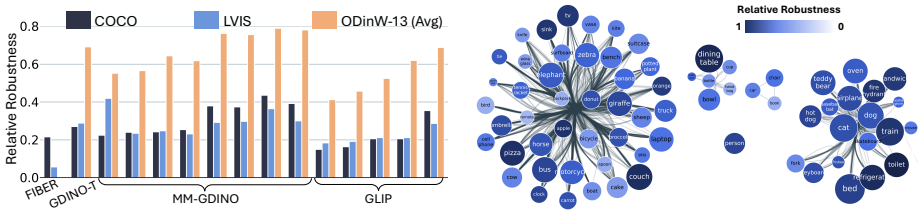


Fig. 10: *(Left)* **Dataset Dependency:** ODinW-13 is more immune to noise than COCO & LVIS (both with comparable robustness) on sev 4 pixelation. *(Right)* **Class-wise robustness:** Certain COCO classes (grouped by bins of log frequency) are more robust (shade of blue) for GLIP-T. Moderate correlation with object size (dot size).

objects likely stems from the small sample size in that bin. Fig. 9 c) illustrate the IOU of overlapping objects *i.e.* occlusion, has hardly any impact on robustness (*counter-intuitive*). Cluttering (or business) in an image is a function of both the number and occlusion of objects. Hence, it's safe to say, despite occlusion, cluttered image with few objects will still be relatively easier to detect.

Fig. 10 (left) shows ODinW-13 maintains high robustness scores of $\simeq 0.6$ across models, nearly twice that of LVIS and COCO (both exhibiting similar robustness). This supports *Fig. 7 (right)*, robustness relies more on the image domain (same images for COCO & LVIS, different for ODinW-13) rather than annotation. Differences between ODinW-13 and COCO/LVIS include *i) Object Size:* Only $\approx 10\%$ of ODinW-13 objects are very small ($\leq 32 \times 32$) versus $\simeq 42\%$ in COCO and $\simeq 58\%$ in LVIS. *ii) Object Density:* ODinW-13 has $\simeq 50\%$ images with single object, compared to $\simeq 12\%$ for COCO and $\simeq 38\%$ for LVIS.

Fig. 10 (right) shows robustness varying by class of object, moderately reflected by average object size (size of dot). For pixelation, robustness correlates with mean class size for COCO / ODinW-13 as 0.52 / 0.45. On COCO, categories like *parking meter*, *stop sign*, and *toilet* are easiest to detect, while for ODinW-13 classes like *lobster*, *jellyfish*, and *hand* are easiest *(Supplementary)*. In contrast, robustness shows almost no correlation with class frequency ($\simeq 0.02$

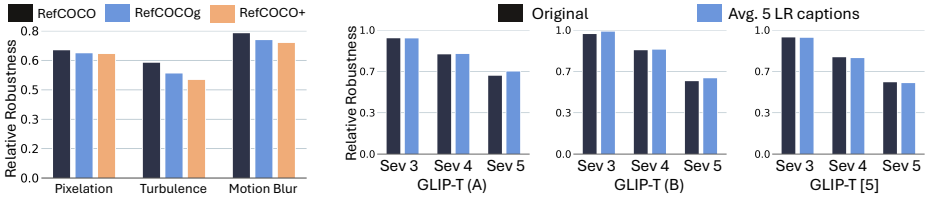


Fig. 11: *(Left)* **Train Captions:** REC fine-tuned FIBER evaluated on COCO. Despite *training on captions* with different degrees of expressiveness (RefCOCOg is most descriptive), robustness varies slightly, indicating limited impact on robustness. *(Right)* **Prompt Engineering:** *Evaluation* on Flickr30k with *test captions* modified with textual context of pixelation (light) has minimal impact on robustness of GLIP variants.

for COCO, $\simeq$ 0.021 for ODinW-13), suggesting **how often a class appears does not drive robustness**. This explains why LVIS's long-tail distribution, adding rare classes to COCO, exhibits similar robustness to COCO. A possible explanation for the disproportionate robustness of certain objects is that they usually appear alone (*e.g.* single traffic signal), and may be easier to detect.

[Takeaways & Model Design] Models are inherently robust for large, singular objects. Conversely, **datasets like ODinW-13 can overstate robustness**, giving false impression of high robustness than reflecting true measure. Robustness seems to largely depend on diverse image domains rather than annotation (detecting *on* what is more important than *what*, COCO $\simeq$ LVIS).

4.4 WHERE: Expressiveness of Captions and Prompt Engineering

Fig. 11 left) shows that fine-tuning on the REC datasets (RefCOCO, RefCOCO+, and RefCOCOg) results in similar robustness with a minor differences in turbulence. RefCOCO contains simple expressions, while RefCOCO+ has appearance-based prompts, and RefCOCOg includes more elaborate and detailed language; all based COCO with differently phrased captions. Empirically, this implies the descriptive nature (expressiveness) of captions or text prompts used in training has seemingly weak/limited impact on robustness. This yet another instance of annotation paying negligible influence on robustness.

Fig. 11 right) evaluates GLIP variants on Flickr30k on 2 sets of captions: 1) *Original (dark)* captions from the dataset , and 2) *LLM-modified (light)* captions obtained from Granite LLM [40], generating five augmented captions infused with context of pixelation/low-resolution. For example, "*A boy smiles in front of a stony wall in a city*" becomes "*In a low-resolution cityscape, a boy's smile is captured against a rough stone wall*". Evaluation occurs on the average of vision embedding fused with 5 LR captions; exhibiting consistent robustness across severities across models. Thus, it's safe to say prompt engineering to introduce noise-awareness in test captions has marginal impact.

ODinW-13 is *evaluated* for robustness using descriptive fine-grained prompts and coarse-grained superclass prompts as shown in Fig. 12. Superclass prompts

groups multiple fine-grained classes under one parent superclass (and reworded if the dataset has only 1 class). This shows a small difference in robustness, with mean Δ of $\simeq 0.14$ for pixelation, $\simeq 0.17$ for turbulence, $\simeq 0.12$ for motion blur, and negligible Δ overall. Lower performance on coarse-grained superclass prompts can explain lower robustness (linearity between robustness & accuracy Fig. 5 (right)). On average, superclass annotations have limited impact on robustness.

[Takeaways & Model Design] The above findings suggest that expressive captions (during fine-tuning), and prompts engineered with the context of degradations (during evaluation) do not significantly improve robustness. Combining these, with previously observed results like the marginal impact of text backbone, (Fig. 6 (right)), and role of annotations (Sec. 4.3), its safe to say that **language expressiveness and distortion-awareness has minimal influence on 'visual' robustness**. To our knowledge, we are among the first to demonstrate this limited role of language, hinting that future robustness efforts targeting vision modality is likely to yeild more success. Another potential direction is to *re-train* entire models with prompts injected with noise context, rather than at evaluation. This would require a new dataset with noise-aware captions and annotations for caption–image alignment and is *left as future work*.

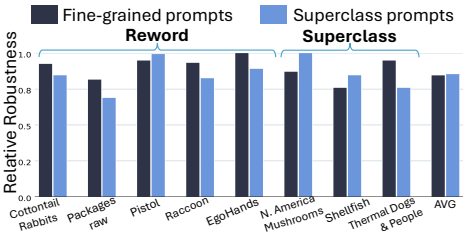


Fig. 12: Fine-grained vs Superclass: Similar robustness on ODinW-13 datasets with pixelation (GLIP-T). Datasets ($\frac{5}{13}$) with accuracy $\simeq 0$ (random prediction) not shown, 1-class datasets 'reworded' & multiple classes clubbed as 'superclass'.

5 HOW: Empirical Validation of Our Analysis

Next, we focus on the empirical validation of our findings and try to explain observations from previously published works.

5.1 Validating Model Design on Real-world datasets

Sec. 4.2 & Fig. 6 (right) in our analysis has highlighted **four** key model designs:

1) *Visual backbone* is the primary determinant of robustness. Improving it w/o modifying other components (neck, fusion) should suffice.
2) *Shallow layers* are most adversely affected by visual noise.
3) *Cross-exchanging* information across backbone layers improves feature overlap across severities, *i.e.* features from pixelated images behave like HQ images.
4) *Language Features* minimal impact on visual robustness (need not be trained).

Table 1: Robust Onion Empirical Validation: BDD-100K trained GLIP-T zero shot evaluation on DAWN, Foggy Cityscapes (Fog City), Virtual KITTI 2, and COCO. E2E and Fuse modify pretrained weights (baselines). PC=Partly Cloudy; O=Overcast; S=Snow; R=Rain; F=Fog; Sa=Sand; All=Overall across all categories. Green/red indicates performance ↑/↓ relative to Zero-shot (0-shot). Our contribution in orange.

Model	Training						Zero-Shot Evaluation									
	BDD-100K						DAWN					Fog	Virtual KITTI 2				COCO
	PC	S	R	F	O	All	F	R	S	Sa	All	City	F	O	R	All	
0-shot	43.4	43.7	41.7	46.6	47.4	31.4	32.6	40.5	38.5	49.8	33.8	21.4	15.6	11.2	11.5	20.9	46.6
E2E	71.3	72.7	71.0	69.5	77.4	66.4	31.6	36.4	38.5	45.2	36.0	26.9	16.7	12.0	12.6	31.6	3.0
	27.9	29.0	29.3	22.9	30.0	35.0	-1.0	-4.1	0.0	-4.6	2.2	5.5	1.1	0.8	1.1	10.7	-43.6
Fuse	69.4	72.1	70.3	67.7	76.6	65.1	30.0	35.9	36.3	42.5	35.6	26.6	17.2	12.3	12.7	29.8	3.8
	26.0	28.4	28.6	21.1	29.2	33.7	-2.6	-4.6	-2.2	-7.3	1.8	5.2	1.6	1.1	1.2	8.9	-42.8
TK0	61.5	63.0	61.0	59.2	68.4	45.1	35.3	42.3	40.6	50.1	32.5	25.5	16.9	12.2	12.6	26.1	43.2
	18.1	19.3	19.3	12.6	21.0	13.7	2.7	1.8	2.1	0.3	-1.3	4.1	1.3	1.0	1.1	5.2	-3.4
NN	61.8	64.0	61.6	63.0	68.5	45.0	33.3	41.6	38.3	49.7	33.4	25.8	16.9	12.0	12.4	28.7	22.1
	18.4	20.3	19.9	16.4	21.1	13.6	0.7	1.1	-0.2	-0.1	-0.4	4.4	1.3	0.8	0.9	7.8	-24.5
NN &	66.8	68.6	66.1	65.5	73.1	55.2	34.0	41.3	39.8	49.5	36.1	28.2	17.3	12.4	13.1	29.9	33.4
TK0	23.4	24.9	24.4	18.9	25.7	23.8	1.4	0.8	1.3	-0.3	2.3	6.8	1.7	1.2	1.6	9.0	-13.2

Table 2: More Real World Dataset: X (Train) → Y (zero-shot evaluation). WiderFace has blurry faces of people; UAVDT & VisDRONE are drone-captured satellite images of streets. Our Contribution in orange. Best performance in **bold**.

Method	# of Trainable	Flickr30k → **WiderFace**		UAVDT → **VisDRONE**	
GLIP-T	Parameters (M)	AP ↑	AP-50 ↑	AP↑	AP-50↑
Zero-Shot	-	11.98	26.49	**20.44**	27.09
E2E	231.76	10.86	24.69	18.68	30.56
Fuse	91.85	10.55	24.92	20.07	29.89
NN & TK0	2.41	**13.23**	**28.71**	19.34	**30.97**

Dataset: Design choices are validated on *real-world driving dataset* BDD-100K [62] (AP50); learning traffic related classes like pedestrian, cars, and bikers. *Zero-shot* noise robustness is evaluated on *noisy weather* driving datasets, including DAWN [26], Foggy Cityscapes [14, 49] (amodal annotation, fog intensity 0.02), and Virtual KITTI 2 [7], all evaluated using mAP. COCO is used to measure the change in original zero-shot performance after adapting the model. We also show results on WiderFace [59] (trained on Flickr30k), and VisDRONE [72] (trained on UAVDT [18]).

Architecture: *Visual backbone and Shallow layers* We extend LR-TK0 [41] method to hierarchical transformers (*e.g.* Swin) for object detection. LR-TK0 originally added trainable tokens on top of a frozen transformer; trained via distillation for low-resolution classification. Here, we remove the expensive teacher student design and retain only the cost efficient trainable tokens, applied *only to the visual backbone while keeping the entire OV-OD frozen*. At every layer, *par-*

ticularly shallow layers, we insert a fixed 32×32 set of trainable spatial tokens. These tokens are interpolated to match the layer spatial resolution and added to the frozen feature maps. Interpolating a small number of spatial tokens provides two advantages: *1) Flexible hierarchy*: Original LR-TK0 attaches a fixed number of prompts to every ViT layer and cannot handle varying $H \times W$, while interpolation adapts to different spatial resolutions across layers. *2) Lower Overhead*: For a 600×600 Swin-T input, our 32×32 tokens add only 5.7% parameters, compared to 22.5% for the fixed token design. We denote call extension as **TK0**.

Cross-exchanging layers For cross-exchanging information across layers, we introduce a simple trainable non-local block [55], implemented as a single head self attention module. The query, key, and value are formed by concatenating spatial tokens from all layers. The self attention operation enables tokens to share spatial information across layers. We denote this method as **NN**. *Supplementary* has a visual representation of this framework.

Implementation GLIP-T is trained using its default training configuration, on the BDD-100K training set (698,630 image text captions). Baselines include *'E2E'* , which trains the entire GLIP model end-to-end (including text backbone), and *'Fuse'*, which freezes the model and trains only the fusion transformer. The fusion network is the default OV-OD design of cross-exchanging information across vision backbone layers. However, both baselines modify pretrained weights and risk degrading zero shot performance of VLM.

Results & Discussion Tab. 1 shows zero-shot performance of BDD-100K trained GLIP-T on real-world driving datasets. Tab. 2 shows an additional zero-shot evaluation on two challenging real-world benchmarks: *WiderFace*, which contains unconstrained face images with significant blur, scale variation, and occlusion, and *VisDrone*, which consists of aerial images captured by drones under various weather, lighting, and environmental conditions.

-Baselines: E2E & Fuse perform poorly on vanilla COCO dataset, highlighting the importance of preserving pretrained weights. *Fuse*, which keeps the language and vision backbones frozen and **trains only cross-layer fusion**, achieves robustness comparable to end-to-end training, with overall Δ of 1.3% on BDD-100K, 0.5% on DAWN, 0.3% on Foggy Cityscapes, and 1.8% on Virtual KITTI 2, despite improving accuracy by about 33%, 2%, 5%, and 9% respectively.

-TK0 and NN: Training our modified TK0 improves robustness while causing minimal drop in zero shot COCO performance since the OV-OD remains entreily frozen. However, its robustness is still below E2E and Fuse. A similar trend is observed for NN. While TK0 lacks cross exchange, NN is a 1-head 1-layer self-attention while fusion network is an entire multi-layered multi-head transformer decoder. Combining both insights through *'NN & TK0'* yields improvements comparable to the baselines, and even surpasses them on DAWN and Foggy Cityscapes; suggest fusing spatial vision features, while modifying improving robustness on shallow layers of vision backbone can improves robustness.

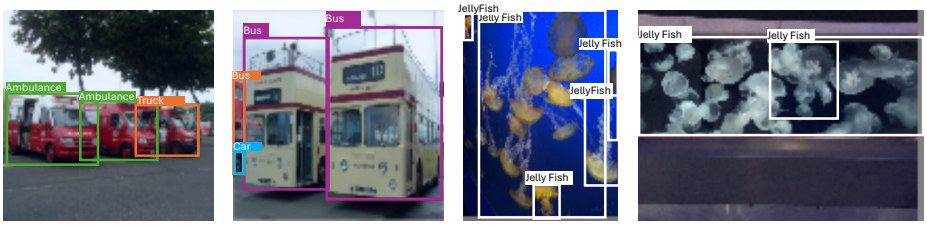
-Compute Cost: For GLIP-T (Swin-T), E2E trains the entire OV-OD model, including the vision backbone, fusion decoder, and language transformer, total-

Table 3: Explanation of previous works with our findings. Reference (Refer) refers to the relevant section/figures in our work, hosting our explanation.

Previous Work Observation		Our Explanation	Refer
Mao *et al.* [35]	High-accuracy detectors are often more robust.	Robustness is proportional to clean accuracy.	Fig. 5
	Robustness improvements are due to the vision backbone, rather than the neck or detection head.	Simple Conv / MLP blocks (neck & head) have a limited impact on feature collapse. Cross-layer feature exchange (*'fusion'*) between vision features improves this overlap. OV-OD has it by design.	Sec. 4.2 & Enhancer in Fig. 8
Yao *et al.* [60]	Simply replacing the vision backbone from Swin-T with Swin-L improved robustness.	Vision backbone depth impacts robustness; Swin-L (24 blocks) > Swin-T (12 blocks).	Sec. 4.1 & Fig. 6 (right)
Zhou *et al.* [71] & Pathak *et al.* [41]	Vision model shallow-layer features exhibit greater degradation than deeper-layer features under noise.	Shallow layers have significant feature collapse (distinct non-overlapping clusters of noisy-clean feats), compared to deeper ones.	Layer #1 and #2 in Fig. 8
Yutaro and Mayu [58]	Robustness contribution from the Swin architecture is more significant than the pretraining strategy & robustness techniques.	Backbone seems to have a stronger influence on robustness than the choice of pretraining strategy and datasets, and other bells and whistles.	Fig. 7 (left) & Fig. 6 (right)
Chhipa *et al.* [13]	GroundingDINO is more robust than Owl-ViT overall.	GroundingDINO clean accuracy is better than Owl-ViT; (robustness $\propto$ clean accuracy).	Fig. 5 (right)
Liu *et al.* [32]	DETR and Deformable DETR have similar robustness, violating the accuracy-robustness linearity.	Both models share the same vision backbone, which likely leads to similar robustness.	Fig. 6 (right)
Pathak [42]	External datasets improve the target dataset's robustness (*'task-agnostic robustness'*).	Detecting '*on* what' is more important than '*what*' (annotations' minimal role; images domain makes the difference).	Sec. 4.1, Sec. 4.3
Bhojanapalli *et al.* [5]	Self-attention is more important than MLP for robustness.	Features exchange (self-attention) reduces feature collapse, while Conv, MLP (feature transformation) do not affect feature collapse.	Sec. 4.2

ing about 231.76 M trainable parameters. Compared to E2E, *Fuse* trains only the fusion transformer with 91.85 M parameters, making it about 2.5× cheaper, gaining almost similar robustness. *TK0* operates only on the vision backbone with **1.57 M** trainable parameters, about **147.4**× lighter. *NN* introduces just **0.84 M** parameters, about **275.7**× fewer compared to end-to-end. In totality, insights from our analysis (*'NN & TK0'*) achieve similar robustness like end-to-end finetuning using **2.41 M parameters**, **96**× lesser number of parameters.

Robustness advancements rarely extend to VLMs-based OV-ODs, primarily because of the complexity (*e.g.* 3-5 transformer backbones, GLEE trained on 64 GPUs across 18 datasets, *etc.*), limiting progress to smaller CNN models such



a) 2/3 correct b) 2/4 correct c) 4/16 correct d) 0 / 18 correct

Fig. 13: Empirical Dataset Bias: Models detect small # objects despite heavy occlusion (a,b); fail for large # objects (c, d). FIBER-B on pixelated OdinW-13.

as Faster R-CNN or simpler tasks. Insights from our Robust Onion can enable cost-efficient extensions that improve robustness under real-world noise.

5.2 Validating insights on Previous Works & Dataset Bias

Tab. 3 shows how Robust Onion analysis aligns well with several prior findings, while providing a unified explanation for some of them. We also show dataset bias for ODinW-13 (pixelated) in Fig. 13. FIBER-B can detect 3-4 objects in pixelated images when the number of objects is small (a & b) despite heavy occlusion Fig. 9 (mid and right). However, when the number of objects are large, the model clubs all of them under one large box.

6 Conclusion

Robust Onion provides a detailed analysis of robustness against visual distortions in OV-ODs, systematically *'peeling'* each component, to assess their individual impact. Analyzing detectors under realistic, yet underexplored, visual distortions (mirroring real-world noise feature collapse) reveals: 1) Vision backbone dominates robustness, outweighing architectural variations or pre-training choices. 2) Shallow layers are most noise-sensitive. 3) Cross-exchange of features within the vision backbone improves robustness. 4) Prompt/caption expressiveness has minimal effect on robustness during both inference and fine-tuning. To further validate our analysis, we test cross-exchanging backbone features to show improvement on real-world autonomous driving datasets. We additionally verify and explain some of the robustness-related observations as reported in previous works. Robust Onion provides a clear, actionable roadmap for advancing the robustness of next-generation OV-OD.

Acknowledgements

The authors thank Steven Dick (UCF High-Performance Computing) and Shyam Marjit (Centre for Data Science, Indian Institute of Science) for their help and contributions to this project.

References

1. Abdelhamed, A., Afifi, M., Go, A.: What do you see? enhancing zero-shot image classification with multimodal large language models (2025), https://arxiv.org/abs/2405.15668
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Bao, W., Deng, R., He, J.: Mint: A simple test-time adaptation of vision-language models against common corruptions. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=yJpBVE4vfo
4. Baranwal, A., Mueez, A., Voelker, J., Bhatia, G., Vyas, S.: Synspill: Improved industrial spill detection with synthetic data. In: 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). p. 1425–1434. IEEE (Oct 2025). https://doi.org/10.1109/iccvw69036.2025.00152, http://dx.doi.org/10.1109/ICCVW69036.2025.00152
5. Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A.: Understanding robustness of transformers for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10231–10241 (October 2021)
6. Bianchi, L., Carrara, F., Messina, N., Gennaro, C., Falchi, F.: The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22520–22529 (2024)
7. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv preprint arXiv:2001.10773 (2020)
8. Chai, J.C.L., Ng, T.S., Low, C.Y., Park, J., Teoh, A.B.J.: Recognizability embedding enhancement for very low-resolution face recognition and quality estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9957–9967 (2023)
9. Chen, J., Guo, H., Yi, K., Li, B., Elhoseiny, M.: Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18030–18040 (2022)
10. Chen, W.T., Vong, Y.J., Kuo, S.Y., Ma, S., Wang, J.: Robustsam: Segment anything robustly on degraded images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4081–4091 (June 2024)
11. Cheng, K., Song, W., Fan, J., Ma, Z., Sun, Q., Xu, F., Yan, C., Chen, N., Zhang, J., Chen, J.: Caparena: Benchmarking and analyzing detailed image captioning in the llm era (2025), https://arxiv.org/abs/2503.12329
12. Cheng, Z., Zhu, X., Gong, S.: Low-resolution face recognition. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 605–621. Springer (2019)
13. Chhipa, P.C., De, K., Chippa, M.S., Saini, R., Liwicki, M.: Open-vocabulary object detectors: Robustness challenges under distribution shifts. In: European Conference on Computer Vision. pp. 62–79. Springer (2024)

14. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
15. Davila, D., Du, D., Lewis, B., Funk, C., Van Pelt, J., Collins, R., Corona, K., Brown, M., McCloskey, S., Hoogs, A., et al.: Mevid: Multi-view extended videos with identities for video person re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1634–1643 (2023)
16. Deng, J., Zhang, H., Ding, K., Hu, J., Zhang, X., Wang, Y.: Zero-shot generalizable incremental learning for vision-language object detection (2024), https://arxiv.org/abs/2403.01680
17. Dou, Z.Y., Kamath, A., Gan, Z., Zhang, P., Wang, J., Li, L., Liu, Z., Liu, C., LeCun, Y., Peng, N., Gao, J., Wang, L.: Coarse-to-fine vision-language pre-training with fusion in the backbone. In: NeurIPS (2022)
18. Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., Tian, Q.: Uavdt dataset (2018), https://sites.google.com/view/grli-uavdt/%E9%A6%96%E9%A1%B5
19. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
20. Gupta, A., Dollar, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
21. Gupta, H., Kotlyar, O., Andreasson, H., Lilienthal, A.J.: Robust Object Detection in Challenging Weather Conditions . In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7508–7517. IEEE Computer Society, Los Alamitos, CA, USA (Jan 2024). https://doi.org/10.1109/WACV57701.2024.00735, https://doi.ieeecomputersociety.org/10.1109/WACV57701.2024.00735
22. Gupta, H., Kotlyar, O., Andreasson, H., Lilienthal, A.J.: Robust object detection in challenging weather conditions. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7508–7517 (2024). https://doi.org/10.1109/WACV57701.2024.00735
23. He, W., Deng, Y., Tang, S., Chen, Q., Xie, Q., Wang, Y., Bai, L., Zhu, F., Zhao, R., Ouyang, W., et al.: Instruct-reid: A multi-purpose person re-identification task with instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17521–17531 (2024)
24. Huynh, N.D., Bouadjenek, M.R., Aryal, S., Razzak, I., Hacid, H.: Visual question answering: from early developments to recent advances – a survey (2025), https://arxiv.org/abs/2501.03939
25. Kazemzadeh, S., Ordonez, V., andre Matten, M., Berg, T.L.: Referitgame: Referring to objects in photographs of natural scenes. In: Conference on Empirical Methods in Natural Language Processing (2014), https://api.semanticscholar.org/CorpusID:6308361
26. Kenk, M.A., Hassaballah, M.: Dawn: vehicle detection in adverse weather nature dataset. arXiv preprint arXiv:2008.05402 (2020)
27. Li, C., Chen, X., Zhao, K., Zhu, J., Chen, J.: Zero-shot quantization for object detection (2025), https://openreview.net/forum?id=XNr6sexQGj
28. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)

29. Li, P., Prieto, L., Mery, D., Flynn, P.J.: On low-resolution face recognition in the wild: Comparisons and new techniques. IEEE Transactions on Information Forensics and Security **14**(8), 2000–2012 (2019). https://doi.org/10.1109/TIFS.2018.2890812
30. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015)
31. Ling, X., Lu, Y., Xu, W., Deng, W., Zhang, Y., Cui, X., Shi, H., Wen, D.: Dive into the resolution augmentations and metrics in low resolution face recognition: A plain yet effective new baseline. arXiv preprint arXiv:2302.05621 (2023)
32. Liu, J., Wang, Z., Ma, L., Fang, C., Bai, T., Zhang, X., Liu, J., Chen, Z.: Benchmarking object detection robustness against real-world corruptions. Int. J. Comput. Vision **132**(10), 4398–4416 (May 2024). https://doi.org/10.1007/s11263-024-02096-6, https://doi.org/10.1007/s11263-024-02096-6
33. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016)
34. Mao, X., Chen, Y., Zhu, Y., Chen, D., Su, H., Zhang, R., Xue, H.: Coco-o: A benchmark for object detectors under natural distribution shifts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6339–6350 (2023)
35. Mao, X., Chen, Y., Zhu, Y., Chen, D., Su, H., Zhang, R., Xue, H.: Coco-o: A benchmark for object detectors under natural distribution shifts (2023), https://arxiv.org/abs/2307.12730
36. Mao, Z., Chimitt, N., Chan, S.H.: Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2021)
37. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
38. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N.: Simple open-vocabulary object detection with vision transformers (2022), https://arxiv.org/abs/2205.06230
39. Minderer, M., Gritsenko, A.A., Houlsby, N.: Scaling open-vocabulary object detection. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=mQPNcBWjGc
40. Mishra, M., Stallone, M., Zhang, G., Shen, Y., Prasad, A., Soria, A.M., Merler, M., Selvam, P., Surendran, S., Singh, S., et al.: Granite code models: A family of open foundation models for code intelligence. arXiv preprint arXiv:2405.04324 (2024)
41. Pathak, P., Marjit, S., Vyas, S., Rawat, Y.S.: LR0.FM: Low-Res Benchmark and Improving robustness for Zero-Shot Classification in Foundation Models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=AsFxRSLtqR
42. Pathak, P., Rawat, Y.S.: Coarse attribute prediction with task agnostic distillation for real world clothes changing reid. In: 36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025. BMVA Press (2025)
43. Pathak, P., Rawat, Y.S.: Colors see colors ignore: Clothes changing reid with color disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16797–16807 (October 2025)

44. Patil, J.S., Pawase, R.S., Dandawate, Y.H.: Classification of low resolution astronomical images using convolutional neural networks. In: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). pp. 1168–1172. IEEE (2017)
45. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 2641–2649 (2015). https://doi.org/10.1109/ICCV.2015.303
46. Qin, Q., Chang, K., Huang, M., Li, G.: Denet: Detection-driven enhancement network for object detection under adverse weather conditions. In: Proceedings of the Asian Conference on Computer Vision. pp. 2813–2829 (2022)
47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
48. Saha, O., Horn, G.V., Maji, S.: Improved zero-shot classification by adapting vlms with text descriptions (2024), https://arxiv.org/abs/2401.02460
49. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**(9), 973–992 (Sep 2018), https://doi.org/10.1007/s11263-018-1072-8
50. Schiappa, M.C., Azad, S., Vs, S., Ge, Y., Miksik, O., Rawat, Y.S., Vineet, V.: Robustness analysis on foundational segmentation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1786–1796 (2024)
51. Shen, H., Zhao, T., Zhu, M., Yin, J.: Groundvlp: Harnessing zero-shot visual grounding from vision-language pre-training and open-vocabulary object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4766–4775 (2024)
52. Shermeyer, J., Van Etten, A.: The effects of super-resolution on object detection performance in satellite imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
53. Tian, X., Gu, J., Li, B., Liu, Y., Hu, C., Wang, Y., Zhan, K., Jia, P., Lang, X., Zhao, H.: Drivevlm: The convergence of autonomous driving and large vision-language models. ArXiv **abs/2402.12289** (2024), https://api.semanticscholar.org/CorpusID:267750682
54. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems **34**, 200–212 (2021)
55. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
56. Wu, J., Jiang, Y., Liu, Q., Yuan, Z., Bai, X., Bai, S.: General object foundation model for images and videos at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3783–3795 (June 2024)
57. Xue, X., Wei, G., Chen, H., Zhang, H., Lin, F., Shen, C., Zhu, X.X.: Reovlm: Transforming vlm to meet regression challenges in earth observation. arXiv preprint arXiv:2412.16583 (2024)
58. Yamada, Y., Otani, M.: Does robustness on imagenet transfer to downstream tasks? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9215–9224 (June 2022)
59. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: A face detection benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

60. Yao, L., Pi, R., Han, J., Liang, X., Xu, H., Zhang, W., Li, Z., Xu, D.: Detclipv3: Towards versatile generative open-vocabulary object detection. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5610–5619 (2024), https://api.semanticscholar.org/CorpusID:269148793
61. Yoo, J., Lee, D., Chung, I., Kim, D., Kwak, N.: What how and when should object detectors update in continually changing test domains? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23354–23363 (2024)
62. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
63. Yu, H., Yi, S., Niu, K., Zhuo, M., Li, B.: Umit: Unifying medical imaging tasks via vision-language models. arXiv preprint arXiv:2503.15892 (2025)
64. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 69–85. Springer (2016)
65. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14393–14402 (2021)
66. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey (2024), https://arxiv.org/abs/2304.00685
67. Zhang, Z., Gong, H., Feng, Y., Chu, Z., Liu, H.: Enhancing object detection in adverse weather conditions through entropy and guided multimodal fusion. In: Proceedings of the Asian Conference on Computer Vision. pp. 2922–2938 (2024)
68. Zhao, S., Schulter, S., Zhao, L., Zhang, Z., Suh, Y., Chandraker, M., Metaxas, D.N., et al.: Taming self-training for open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13938–13947 (2024)
69. Zhao, X., Chen, Y., Xu, S., Li, X., Wang, X., Li, Y., Huang, H.: An open and comprehensive pipeline for unified object grounding and detection. arXiv preprint arXiv:2401.02361 (2024)
70. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16793–16803 (2022)
71. Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., Alvarez, J.M.: Understanding the robustness in vision transformers. In: International conference on machine learning. pp. 27378–27394. PMLR (2022)
72. Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., Ling, H.: Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 7380–7399 (2021)