Roboflow100-VL: A Multi-Domain Object Detection Benchmark for Vision-Language Models

Peter Robicheaux[1,*], **Matvei Popov**[1,*], **Anish Madan**[2], **Isaac Robinson**[1], **Joseph Nelson**[1], **Deva Ramanan**[2], **Neehar Peri**[2]
[1]Roboflow, [2]Carnegie Mellon University

rf100-vl.org

Abstract

Vision-language models (VLMs) trained on internet-scale data achieve remarkable zero-shot detection performance on common objects like `car`, `truck`, and `pedestrian`. However, state-of-the-art models still struggle to generalize to out-of-distribution classes, tasks and imaging modalities not typically found in their pre-training. Rather than simply re-training VLMs on more visual data, we argue that one should align VLMs to new concepts with annotation instructions containing a few visual examples *and* rich textual descriptions. To this end, we introduce Roboflow100-VL, a large-scale collection of 100 multi-modal object detection datasets with diverse concepts not commonly found in VLM pre-training. We evaluate state-of-the-art models on our benchmark in zero-shot, few-shot, semi-supervised, and fully-supervised settings, allowing for comparison across data regimes. Notably, we find that VLMs like GroundingDINO and Qwen2.5-VL achieve less than 2% zero-shot accuracy on challenging medical imaging datasets within Roboflow100-VL, demonstrating the need for few-shot concept alignment. Lastly, we discuss our recent CVPR 2025 Foundational FSOD competition and share insights from the community. Notably, the winning team significantly outperforms our baseline by 17 mAP! Our code and dataset are available on GitHub and Roboflow.

1 Introduction

Vision-language models (VLMs) trained on web-scale datasets achieve remarkable zero-shot performance on many popular academic benchmarks [156, 87, 132]. However, the performance of such foundation models varies greatly when evaluated in-the-wild, particularly on out-of-distribution classes, tasks (e.g. material property estimation, defect detection, and contextual action recognition) and imaging modalities (e.g. X-rays, thermal spectrum data, and aerial imagery). In this paper, we introduce Roboflow100-VL (RF100-VL), a large-scale multi-domain dataset to benchmark state-of-the-art VLMs on hundreds of diverse concepts not typically found in internet pre-training.

Status Quo. Foundation models are often trained on large-scale datasets curated from diverse sources around the web. However, despite their scale and diversity, these pre-training datasets still follow a long-tail distribution [121], causing foundation models to generalize poorly to rare concepts [106]. A common approach for improving the performance of VLMs is to scale up training data and model size [24]. However, we argue that some data will always remain out-of-distribution, whether due to being sequestered from the internet or being created after the model's training cutoff [144], motivating the need to learn new concepts from a few examples.

Evaluating Out-of-Distribution Generalization. Existing benchmarks primarily assess spatial understanding through visual question answering (VQA) and common sense reasoning [87, 157,

[*]Equal Contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

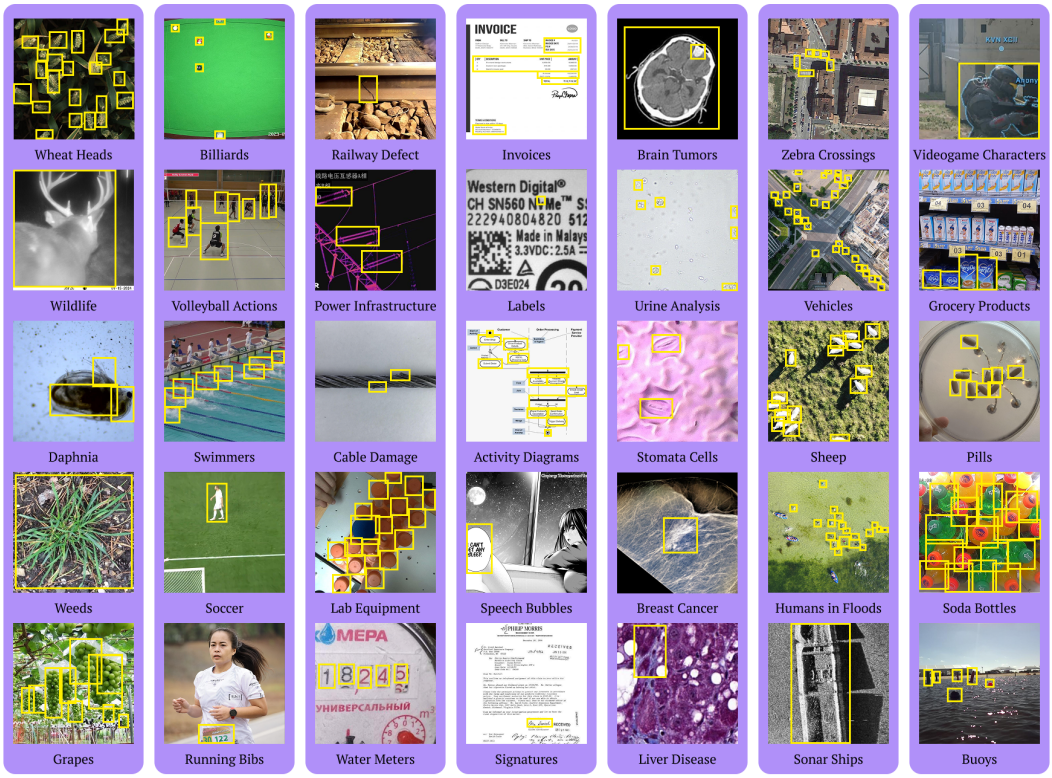
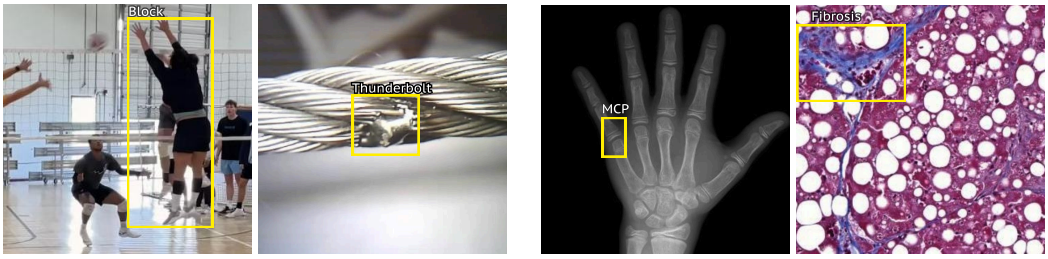


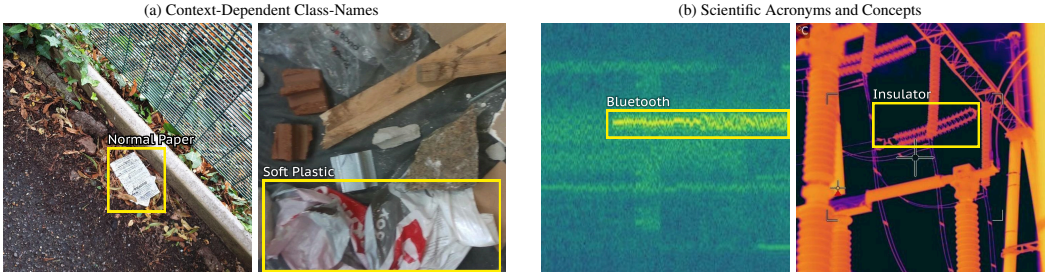
Figure 1: **Roboflow100-VL Dataset.** We identify a set of 100 challenging datasets from Roboflow Universe that contain concepts not typically found in internet-scale pre-training. To simplify analysis, we cluster these 100 datasets using per-dataset CLIP [113] embeddings into seven categories. We visualize examples from each of these categories above. Furthermore, we also generate multi-modal instructions for each dataset with a few visual examples and rich textual descriptions per class to facilitate few-shot concept alignment.

132]. However, we argue that evaluating model performance on compositional reasoning benchmarks alone does not effectively measure generalization to out-of-distribution tasks. Moreover, current spatial understanding and grounding benchmarks (e.g. RefCOCO [155] and OdinW [82]) typically evaluate performance on classes commonly found in internet pre-training. We demonstrate that such benchmarks artificially inflate model performance and are not representative of many real-world applications (cf. Table 1). To address this limitation, we introduce RF100-VL, a large-scale detection benchmark comprised of 100 multi-modal datasets from diverse domains (cf. Fig. 1). Importantly, we carefully curate RF100-VL such that it cannot be solved by simply prompting state-of-the-art models with class names. Specifically, we include datasets where classes are labeled using scientific names (e.g. liver `fibrosis` and `steatosis`), acronyms (e.g. `DIP` and `MCP`), context-dependent names (e.g. detecting a `block` vs. `set` in the context of volleyball), material properties (e.g. `paper` vs. `soft plastic`), and diverse imaging modalities (cf. Fig. 2). We posit that models must leverage multi-modal contextual information (presented in the form of multi-modal annotator instructions) to effectively align to target concepts in RF100-VL.

Multi-Modal Annotator Instructions. Annotating large-scale datasets is an iterative process that often requires extensive discussions between data curators and annotators to clarify class definitions and ensure label consistency. These (often multi-modal) labeling instructions provide rich contextual information not provided by class names alone. We argue that aligning foundation models to target concepts can be principally addressed through the lens of few-shot learning by presenting vision-language models with visual examples and rich textual descriptions per class (cf. Fig. 3). Importantly, this approach mirrors how we align human annotators to concepts of interest with few-shot multi-modal examples [34, 91].



(a) Context-Dependent Class-Names (b) Scientific Acronyms and Concepts



(c) Catch-all Categories (d) Multi-Modal Imaging Modalities

Figure 2: **Hard Examples in Roboflow100-VL.** Our dataset is particularly challenging because it is difficult to detect objects in RF100-VL using class-names alone. Specifically, we select datasets where classes are labeled using scientific names, acronyms, context-dependent names, material properties. We posit that models must leverage multi-modal contextual annotations to address such hard examples.

Contributions. We present three major contributions. First, we introduce RF100-VL, a large-scale, multi-domain benchmark designed to evaluate vision-language models (VLMs) on challenging real-world use cases. We evaluate state-of-the-art models on our benchmark in zero-shot, few-shot, semi-supervised, and fully-supervised settings, allowing for comparison across data regimes. Our extensive experiments highlight the difficulty of adapting VLMs to out-of-distribution tasks and reveal the limitations of current state-of-the-art methods. Lastly, we highlight the results of our recent CVPR 2025 challenge hosted in conjunction with the Workshop on Visual Perception via Learning in An Open World.

2 Related Works

Vision Language Models are trained using large-scale, weakly supervised image-text pairs sourced from the web. Although many VLMs primarily focus on classification [113] or image understanding, recent methods address spatial understanding with open-vocabulary detectors. Early approaches adapted VLMs for object detection by classifying specific image regions [62, 63] or integrating detection components into frozen [78] or fine-tuned [98, 97, 52] encoders. In contrast, RegionCLIP [160] employs a multi-stage training strategy that involves generating pseudo-labels from captioning data, performing region-text contrastive pre-training, and fine-tuning on detection tasks. GLIP [83] treats detection as a phrase grounding problem by using a single text query for the entire image. Detic [161] improves long-tail detection performance by utilizing image-level supervision from ImageNet [117]. Notably, recent VLMs achieve remarkable zero-shot performance and are widely used as "black box" models in diverse downstream applications [90, 108, 75, 103, 130]. More recently, multi-modal large language models (MLLMs) such as Qwen2.5-VL [28] and Gemini 2.5 Pro [47] frame spatial understanding as a text generation task. Interestingly, such generalist MLLMs perform worse at object detection than task-specific models like GroundingDINO [86]

Fine-Tuning Vision-Language Models is crucial for adapting foundation models to downstream tasks [68, 158, 59]. Traditional fine-tuning methods, such as linear probing [37, 67] and full fine-tuning [146, 151] can be computationally expensive. Instead, parameter-efficient approaches like CLIP-Adapter [59] and Tip-Adapter [159] optimize lightweight MLPs while keeping encoders frozen. Although prior few-shot learners commonly used meta-learning [154], more recent approaches show that transfer learning generalizes better to novel categories [145]. In particular, Pan et. al. [105] demonstrates that transfer learning can be effectively used to fine-tune foundation models using

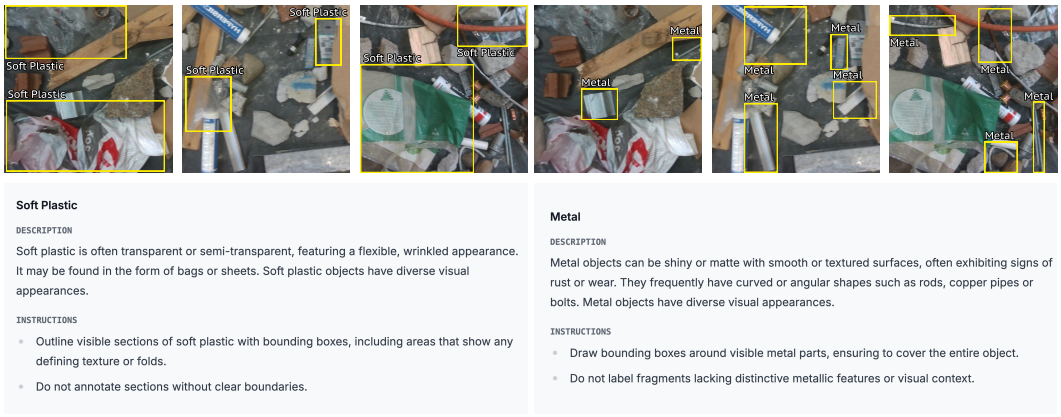


Figure 3: **Multi-Modal Few-Shot Examples.** We present an example of the few-shot visual examples and rich text descriptions used for in-context prompting and fine-tuning. Notably, image examples used for each class may overlap and are only guaranteed to have exhaustive annotations for one class. Such multi-modal examples help clarify ambiguous concepts like `soft plastic` and `metal`.

a few multi-modal examples. More recently, in-context learning [152] demonstrates promising results for test-time few-shot adaptation without gradient-based fine-tuning. We explore such test-time fine-tuning strategies in the context of MLLMs [47, 28] to learn from multi-modal annotator instructions.

Benchmarking Vision-Language Models is of significant interest to the community. State-of-the-art VLMs are typically evaluated using benchmarks such as MMStar [35], MMMU [157], MME [84], ScienceQA [89], MMBench [87], MM-Vet [156], Seed-Bench [81], and MMVP [133]. These benchmarks evaluate a broad set of vision-language tasks, including fine-grained perception, reasoning, common sense knowledge, and problem solving in various domains. However, existing evaluations primarily focus on multi-modal understanding in the context of visual question answering (VQA). In contrast, RF100-VL evaluates VLM detection accuracy given a few visual examples and rich textual descriptions. Prior VLM grounding benchmarks like RefCOCO [155] often focus on referential grounding of common object categories. Recent efforts like ODinW [82] consider more challenging scenarios by sourcing real-world data from Roboflow [38]. However, we find that state-of-the-art methods achieve high zero-shot accuracy on RefCOCO and OdinW [28], suggesting that these datasets may not be well suited for evaluating foundational few-shot object detection [91].

3 Roboflow100-VL Benchmark

As shown in Fig. 1, RF100-VL consists of diverse datasets not typically found in internet-scale pre-training. We highlight our data curation procedure (Section 3.1) and present several baselines to evaluate state-of-the-art models in the zero-shot and few-shot settings (Section 3.2). We also evaluate models under the semi-supervised and fully-supervised settings in Appendix F.

3.1 Creating Roboflow100-VL

We source our datasets from Roboflow Universe, a community-driven platform that hosts diverse open-source datasets created to solve real-world computer vision tasks. With more than $500,000$ public datasets spanning medical imaging, agriculture, robotics, and manufacturing, we focus on selecting high-quality datasets not commonly found in internet-scale pre-training (e.g. COCO [85], Objects365 [124], GoldG [73], CC4M [125]) to better assess VLM generalization to rare concepts. When selecting candidates for RF100-VL, we prioritized datasets where images contained multiple objects, ensuring more realistic evaluation beyond classification. In addition, we sought out datasets with semantically ambiguous names (e.g. "button" can refer to both clothing and electronics) to encourage algorithms to leverage multi-modal annotator instructions rather than simply relying on class names. We manually validate the labeling quality of each dataset to ensure exhaustive

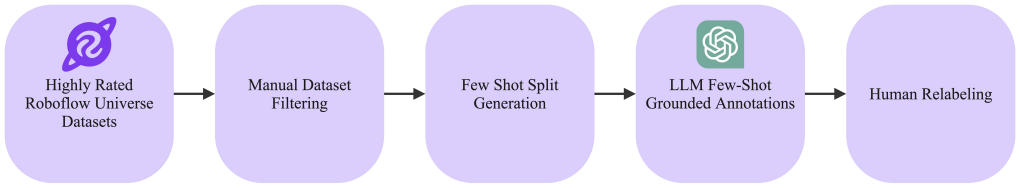


Figure 4: **Dataset Curation.** We begin by sorting all object detection datasets on Roboflow Universe by stars as a proxy for quality and usefulness to the community. Next, we manually filter out all datasets with common classes, datasets where images only have a single focal object, or datasets with watermarks. We generate 10-shot splits following the protocol defined by Wang et.al. [145], where we find a subset of images with 10 total instances per class. We use these 10-shot splits to generate visually grounded "annotator instructions", and manually update these instructions to add any salient details missed by GPT-4o. Finally, human labelers verify that all images within a dataset follow consistent annotation policies (e.g. bounding-box fit, semantic legibility of class names, and completeness of annotation instructions).

annotations. In cases without exhaustive annotations, we manually re-annotate the dataset to the best of our ability (cf. Fig 4). In total, we spent 1693 hours labeling, reviewing, and preparing the dataset.

Multi-Modal Annotation Generation. Annotator instructions offer precise class definitions and visual examples that help clarify annotation policies (e.g. by highlighting typical cases, corner cases, and negative examples) and improve labeling accuracy. Despite providing significant value during the labeling process, few datasets publicly release these annotator instructions. Recognizing the importance of these instructions in aligning humans with target concepts of interest, we generate multi-modal annotator instructions for all 100 datasets within RF100-VL (cf. Fig 3).

We prompt GPT-4o [24] to generate an initial set of annotator instructions, providing in-context examples based on the nuImages annotator guidelines. Our prompt includes a structured output template, along with dataset metadata, class names, and few-shot visual examples per class. In practice, we find that GPT-4o often overlooks the few-shot images and instead relies heavily on class names to generate class descriptions. Notably, GPT-4o struggles when class names are uninformative and sometimes produces overly vague instructions that, while correct, lack useful detail. To address this, we manually verify all generated annotator instructions to mitigate hallucinations and incorporate additional informative visual details missed by the model. We include our annotation generation prompt in Appendix P.

Dataset Statistics. Figure 5 (left) presents an overview of the different types of datasets within RF100-VL, detailing the number of classes, images and annotations per category. RF100-VL contains a total of 564 classes and 164,149 images, with over 1.3 million annotations. The "Other" category has the highest number of classes (142), followed by "Industrial" (122) and "Flora & Fauna" (70). Despite having fewer classes, the "Flora & Fauna" category has the highest number of images (46,718) and annotations (441,677), indicating a higher density of annotations per image. Figure 5 (right) provides a visual representation of class distribution, reinforcing the dominance of the "Other", "Industrial", and "Flora & Fauna" categories. In contrast, "Sports" has the fewest classes (36) and the least representation in RF100-VL. Despite consisting of 100 datasets, RF100-VL has about half the number of images as COCO [85], making this an approachable benchmark for the academic community.

3.2 State-of-the-Art Baselines

We train and evaluate all models on each dataset within RF100-VL independently. Importantly, we do not tune any parameters or modify zero-shot prompts per-dataset. For all models, we compute metrics using pycocotools with maxDets set to 500 instead of the usual 100 because there are many images with more than 100 objects. We discuss our evaluation protocol further in Appendix B.

Zero-Shot Baselines prompt models with class names or expressive descriptions [96] to detect target concepts. However, the effectiveness of zero-shot prompting depends on the pre-training data: If the target class name is semantically meaningful and aligns well with the model's foundational

Dataset Type	# Classes	# Images	# Anno.
Aerial	29	11,627	186,789
Document	88	21,418	127,129
Flora & Fauna	70	46,718	441,677
Industrial	122	29,758	205,627
Medical	77	16,369	125,433
Sports	36	8,443	58,508
Other	142	29,816	210,328
All	564	164,149	1,355,491

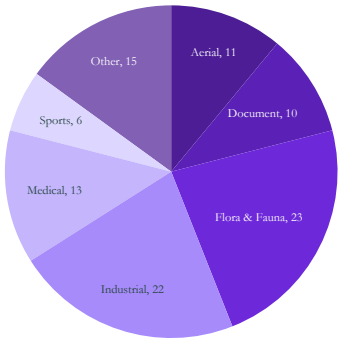


Figure 5: **Dataset Statistics.** The table on the left provides details on the number of classes, images, and annotations across different dataset types within RF100-VL. The figure on the right illustrates the distribution of dataset types by count. Notably, despite containing 100 datasets, RF100-VL is 50% the size of COCO [85] (by number of images) and can feasibly be benchmarked on academic-level compute.

pre-training, performance is strong; otherwise, the model fails catastrophically. We benchmark the zero-shot performance of Detic [161], OWLv2 [97], GroundingDINO [86], MQ-GLIP [152], QwenV2.5-VL [28] and Gemini 2.5 Pro [47].

Few-Shot Baselines. We evaluate three types of few-shot baselines: visual prompting, multi-modal prompting, and federated fine-tuning. Visual prompting uses images of target concepts that are difficult to describe through text as prompts to help models learn novel concepts in-context. For example, while "hard plastic" is a broad and ambiguous category that is hard to define through text, providing image examples improves concept alignment. Typically, visual prompts are tokenized and fed as inputs to a frozen VLM. Here, we apply MQ-GLIP [152] with image prompting. Multi-modal prompting combines language and visual prompts to leverage multi-modal features. Intuitively, using both text and images yields better alignment than using either modality alone. In the case of "soft plastic", ambiguous concepts can be clarified with textual descriptions (e.g., "thin plastic film" and "plastic bag") alongside visual examples. Both visual and language prompts are tokenized and separately fed into a frozen VLM. We evaluate MQ-GLIP [152], and Gemini 2.5 Pro [47] by prompting models with class names, few-shot images, and annotator instructions. Lastly, federated fine-tuning modifies the standard cross-entropy classification to only treat exhaustively annotated classes as true negatives for each image. We follow the implementation from Madan et. al. [91] when fine-tuning Detic [161]. We slightly modify the federated loss when fine-tuning YOLO [71, 74] to avoid using Madan et. al's frequency prior, opting to instead determine hard negatives using per-image annotations.

4 Experiments

We conduct extensive experiments to evaluate the performance of state-of-the-art models on RF100-VL. We present our zero-shot and few-shot results below. See Appendix A for additional implementation details and Appendix F for semi-supervised and fully supervised results.

4.1 Metrics

Each dataset within RF100-VL is independently evaluated using AP. We report the average accuracy per super-category to simplify analysis. RF100-VL includes datasets that are out-of-distribution from typical internet-scale pre-training data, making it particularly challenging (even for VLMs). To construct the few-shot split, we follow the K-shot dataset creation process established by [145]. See Appendix E for further discussion on few-shot split selection. Importantly, all methods across data regimes are evaluated on the same fully annotated test set. In Table 1, we highlight that prior methods report different results on COCO and OdinW than our reproduced results. YOLOv8 [71] and YOLOv11 [74] achieve slightly different performance on COCO because the original results are reported using Ultralytics, whereas our results are computed using pycocotools. Importantly,

this discrepancy in tooling yields a larger disparity on RF100-VL, discussed further in Appendix B. Further, we find that Qwen2.5-VL evaluates on ODinW using a referential grounding protocol (reported, see GitHub issue) instead of a traditional object detection protocol (ours). Specifically, referential grounding prompts a model with only the true positive classes in each test image, while object detectors prompt a model with *all* classes. The former dramatically reduces the number of false positives. We evaluate Gemini 2.5 Pro using both protocols for completeness.

4.2 Empirical Analysis of Results

State-of-the-Art Zero-Shot and Few-Shot Models Struggle on Roboflow100-VL. RF100-VL is a much harder dataset than prior open-vocabulary object detection benchmarks. Specifically, GroundingDINO achieves 49.2 mAP on ODinW-13, but only reaches 15.7 mAP on RF100-VL. Similar trends can be seen with Qwen2.5-VL and Gemini 2.5 Pro (cf. Table 1). Notably, both RF100-VL and ODinW-13 are sourced from Roboflow Universe, but our dataset is carefully curated to evaluate performance on target concepts not typically found in internet-scale pre-training.

Open-Vocabulary Object Detectors Outperform MLLMs. We find that open-vocabulary object detectors like Detic, GroundingDINO, OWLv2, and MQ-GLIP consistently outperform MLLMs like Qwen 2.5 VL, Gemini 2.5 Pro, despite these MLLMs pre-training on orders of magnitude more data. We posit that this poor performance can be attributed to MLLMs not reporting per-box confidence scores or ensuring that predictions don't overlap (e.g. non-maximal suppression). This highlights the advantage of task-specific architectures over generalist models.

Multi-Modal Annotator Instructions Provide Limited Benefit. Somewhat surprisingly, state-of-the-art MLLMs struggle to benefit from multi-modal annotator instructions. In fact, prompting with instructions provides inconsistent benefit compared to prompting with class names (e.g. Qwen2.5VL improves but Gemini 2.5 Pro degrades considerably). Intuitively, we expect annotator instructions to improve object detection performance by resolving semantic ambiguity in class names and providing rich contextual information. However, we posit that this performance decline can be attributed to the fact that MLLMs are instruction-tuned for open vocabulary detection with rigid prompt structures, making it difficult to effectively leverage additional contextual information.

Large-Scale Pre-Training Improves Fine-Tuned Few-Shot Performance in Specialists. We find that fine-tuning GroundingDINO [83] achieves the best few-shot performance, significantly outperforming all YOLO variants by more than 10%. Notably, all gradient-based fine-tuning baselines outperform in-context visual prompting and multi-modal prompting methods, suggesting that in-context prompting provides limited benefit for rare classes not seen in pre-training. We posit that GroundingDINO's large-scale task-specific pre-training makes it easier to learn new concepts during fine-tuning.

Table 1: **Comparison to Other Benchmarks.** We find that state-of-the-art MLLMs achieve considerably lower performance on RF100-VL compared to OdinW-13, highlighting the difficulty of our proposed dataset. Further, models that performed better on COCO did not consistently perform better on the RF100-VL, indicating that the newer YOLO models might be overfitting to COCO. Lastly, we highlight a discrepancy between reported and reproduced numbers on both COCO and OdinW. Discrepancies in COCO evaluation can be attributed to differences in evaluation toolkits, while discrepancies in ODinW evaluation can be attributed to prior work evaluating models using referential grounding evaluation protocols, while we use standard object detection evaluation protocols. We discuss this further in section 4.1. Following prior work, we use single-class prompts for MLLMs in this table (cf. Appendix A).

Method	COCO Val		OdinW-13		Roboflow100-VL
	Reported	Ours	Reported	Ours	Ours
Zero-Shot					
Qwen 2.5-VL (72B) [28] (Class Names Only)	-	-	43.1	30.9	7.8
Gemini 2.5 Pro [47] (Class Names Only)	-	-	41.9	33.7	8.0
Fully-Supervised					
YOLOv8n [71]	37.3	37.4	-	-	54.9
YOLOv11n [74]	39.5	39.4	-	-	55.3
YOLOv8s [71]	44.9	45.0	-	-	56.2
YOLOv11s [74]	47.0	46.9	-	-	56.2
YOLOv8m [71]	50.2	50.3	-	-	56.4
YOLOv11m [74]	51.5	51.5	-	-	56.5

Table 2: **Roboflow100-VL Benchmark.** We evaluate the zero-shot, few-shot, semi-supervised, and fully-supervised performance of state-of-the-art methods on the RF100-VL benchmark. We find that RF100-VL is particularly challenging for zero-shot and few-shot approaches, with most methods struggling to achieves 10% mAP averaged over all 100 datasets. Notably, we find that GroundingDINO achieves the best zero-shot and few-shot accuracy. We use a double horizontal bar to separate specialist models from generalist MLLMs. Note that we use multi-class prompts for MLLMs in this table (cf. Appendix A).

Method	Aerial	Document	Flora & Fauna	Industrial	Medical	Sports	Other	All
Zero-Shot								
Detic [161]	12.2	4.5	17.9	6.0	0.8	7.6	11.2	9.5
GroundingDINO [86]	21.5	9.2	27.9	10.3	2.1	13.3	17.5	15.7
OWLv2 [97] (Class Names Only)	19.8	12.2	23.3	7.8	2.1	12.5	14.3	13.6
MQ-GLIP-Text [152](Class-Names Only)	12.1	10.0	23.2	7.8	1.4	9.3	14.2	12.2
Qwen 2.5 VL (72B) [28] (Class Names Only)	4.6	3.9	10.4	4.1	1.6	6.0	5.6	5.6
Qwen 2.5 VL (72B) [28] (Instructions Only)	5.4	5.0	14.8	5.6	1.7	7.6	7.6	7.6
Gemini 2.5 Pro [47] (Class Names Only)	8.7	11.8	18.3	8.6	5.3	6.5	15.4	11.6
Gemini 2.5 Pro [47] (Instructions Only)	1.8	6.2	7.9	3.5	0.6	2.1	5.9	4.5
Few-Shot (10 shots)								
Detic [161] w/ Federated Loss [91]	19.5	19.6	28.4	25.9	8.5	26.6	25.7	22.8
MQ-GLIP-Image [152] (Images Only)	4.4	3.2	13.3	3.9	1.4	7.4	6.9	6.4
MQ-GLIP [152] (Class Names + Images)	12.1	9.5	23.1	7.8	1.4	9.3	14.3	12.2
GroundingDINO [86]	32.4	30.6	41.3	37.8	18.3	33.2	32.0	33.6
YOLOv8n [71]	12.8	22.8	20.9	28.1	13.7	13.6	19.9	20.2
YOLOv8n [71] w/ Federated Loss [91]	13.5	25.4	22.0	25.9	14.6	14.6	21.3	21.7
YOLOv8s [71]	15.9	22.8	22.1	24.7	13.9	18.0	21.7	20.7
YOLOv8s [71] w/ Federated Loss [91]	17.4	24.9	25.4	26.5	16.8	18.5	23.3	23.6
YOLOv8m [71]	14.3	24.0	19.7	24.9	13.1	19.7	22.9	20.3
YOLOv8m [71] w/ Federated Loss [91]	16.9	23.3	20.8	26.6	16.0	21.4	22.6	22.6
Qwen 2.5 VL (72B)[28] (Instructions + Images)	5.7	6.6	14.8	5.8	1.7	7.3	6.8	7.6
Gemini 2.5 Pro [47] (Images)	6.2	9.4	17.5	9.5	2.7	5.0	9.7	9.8
Gemini 2.5 Pro [47] (Instructions + Images)	5.3	8.8	15.0	8.8	2.1	4.9	9.5	8.8

Do COCO Detectors Generalize Beyond COCO? Real-time object detectors are often optimized for COCO, assuming better performance on COCO translates to real-world improvements. However, real-world datasets (such as those in RF100-VL) are often much smaller and more diverse than COCO, challenging this assumption. Specifically, although RF100-VL has half as many images as COCO, it has more than seven times as many classes (cf. Fig. 5). Interestingly, we find that models that achieved higher performance on COCO did not necessarily improve real-world performance on RF100-VL. For example, YOLOv11 outperforms YOLOv8 on COCO but performs similarly to YOLOv8 across all three tested sizes (nano, small, medium) on RF100-VL. This suggests that newer YOLO models may be overfitting to COCO, as gains on that dataset don't transfer to real-world datasets. Lastly, we find that increasing model size leads to smaller performance improvements on RF100-VL compared to COCO. The performance difference between the smallest and largest models within a model family is at most 1.9 mAP, suggesting that simply increasing model capacity may not lead to significant performance gains on RF100-VL.

4.3 CVPR 2025 Foundational FSOD Challenge

We hosted a challenge at CVPR 2025 to encourage broad community involvement in addressing the problem of aligning foundation models to target concepts with few-shot visual examples and rich textual descriptions. Importantly, we use a subset of 20 datasets from RF100-VL for this challenge to lower the barrier to entry. Our competition received submissions from 25 teams (some submissions are private) at the close of our competition on June 8th, 2025 AOE. Notably, ten teams beat our best baseline. We present the current top three teams in Table 3. We summarize the contributions of the top three teams in Appendix M and include a link to full technical reports and code here.

4.4 Limitations and Future Work

Reliance on Crowdsourced Annotations. All our datasets are sourced from Roboflow Universe, a community platform where anyone can upload dataset annotations. Although this allows us to source diverse datasets, it introduces uncertainty regarding overall annotation quality. While we manually inspect and re-annotate all datasets to ensure quality to the best of our ability, verifying annotations in specialized domains like medical imaging remains a significant challenge.

Table 3: **CVPR 2025 Foundational FSOD Challenge with Roboflow20-VL.** This year's challenge winner beat our best few-shot baseline by 17 AP! For more details about top performing methods, see Appendix M.

Method	Aerial	Document	Flora & Fauna	Industrial	Medical	Sports	Other	All
Zero-Shot								
Detic [161]	4.1	1.4	22.2	6.3	0.1	1.0	9.7	8.4
GroundingDINO [86]	28.5	5.1	33.7	12.8	0.4	5.1	16.9	16.8
OWLv2 [97] (Class Names Only)	35.5	4.9	24.4	12.0	0.1	3.2	12.7	14.2
MQ-GLIP-Text [152](Class-Names Only)	30.1	2.5	32.8	5.5	0.5	6.4	10.8	14.0
Qwen 2.5 VL (72B) [28] (Class Names Only)	3.8	3.5	10.2	2.8	0.1	9.6	3.9	5.1
Qwen 2.5 VL (72B) [28] (Instructions Only)	4.9	7.8	13.4	5.1	0.4	11.5	5.8	7.4
Gemini 2.5 Pro [47] (Class Names Only)	3.5	9.7	21.5	13.3	0.4	8.9	12.2	11.5
Gemini 2.5 Pro [47] (Instructions Only)	1.0	8.0	9.7	7.9	0.1	4.1	4.8	5.7
Few-Shot (10 shots)								
Detic w/ Federated Loss [91]	11.6	14.3	30.8	24.7	8.9	17.4	21.0	20.3
GroundingDINO [86]	39.9	34.5	45.7	37.8	23.3	26.3	24.7	33.4
MQ-GLIP-Image [152] (Images Only)	1.8	1.1	17.6	1.8	0.1	6.6	6.8	6.7
MQ-GLIP [152] (Class Names + Images)	29.8	2.5	32.7	5.6	0.5	6.5	10.9	14.0
Qwen 2.5 VL (72B) [28] (Instructions + Images)	5.1	9.3	15.2	2.9	0.2	8.5	5.7	7.2
Gemini 2.5 Pro [47] (Images Only)	7.7	14.2	24.3	4.0	0.2	12.8	9.7	9.7
Gemini 2.5 Pro [47] (Instructions + Images)	8.4	12.4	12.4	19.3	0.2	8.6	4.9	8.6
Challenge Submissions								
BEATON	52.4	46.9	56.3	62.0	42.9	42.0	45.8	50.4
FDUROILab	52.3	49.9	56.9	61.6	42.1	41.9	42.4	49.8
NJUST-KMG	49.5	43.8	57.4	59.1	42.1	42.9	43.1	49.0

Generated Annotator Instructions May Not Reflect Real Instructions. Our annotator instructions are automatically generated by GPT-4o and are manually verified for correctness. However, they may not fully reflect the nuances of real-world instructions typically developed alongside dataset collection. We encourage the community to release real annotator instructions generated through iterative discussions between annotators and stakeholders. Furthermore, although our annotator instructions provide high-level class descriptions, they often do not directly incorporate image evidence to identify typical cases, edge cases, and negative examples. Future work should explore how to create better automatic annotator instructions.

Generalist and Specialist Models have Complementary Strengths. Although specialist models like GroundingDINO [86] outperform generalist models like Qwen2.5-VL [28], MLLMs can more easily process few-shot visual examples and rich textual descriptions. Future work should combine the versatility of MLLMs with the precision of specialist models.

5 Conclusion

In this paper, we introduce Roboflow100-VL, a large-scale benchmark to evaluate state-of-the-art VLMs on concepts not typically found in internet-scale pre-training. RF100-VL is curated to evaluate detection performance on out-of-distribution tasks (e.g. material property estimation, defect detection, and contextual action recognition) and imaging modalities (e.g. X-rays, thermal spectrum data, and aerial imagery) using a few visual examples and rich textual descriptions. We find that state-of-the-art models struggle on this challenging benchmark, demonstrating the limitations of existing methods, highlighting opportunities to develop better algorithms that effectively use multi-modal annotator instructions. We hope that RF100-VL will be a rigorous test-bench for future VLMs and MLLMs.

6 Acknowledgments

This work was supported in part by compute provided by NVIDIA, and the NSF GRFP (Grant No. DGE2140739).

References

[1] Roboflow 100. *activity diagrams Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/activity-diagrams-qdobr.

[2] Roboflow 100. *aerial pool Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/aerial-pool.

[3] Roboflow 100. *bees Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/bees-jt5in.

[4] Roboflow 100. *cable damage Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/cable-damage.

[5] Roboflow 100. *circuit voltages Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/circuit-voltages.

[6] Roboflow 100. *flir camera objects Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/flir-camera-objects.

[7] Roboflow 100. *grass weeds Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/grass-weeds.

[8] Roboflow 100. *halo infinite angel videogame Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/roboflow-100/halo-infinite-angel-videogame.

[9] Roboflow 100. *paper parts Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/paper-parts.

[10] Roboflow 100. *peixos fish Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/peixos-fish.

[11] Roboflow 100. *signatures Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/signatures-xc8up.

[12] Roboflow 100. *soda bottles Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/soda-bottles.

[13] Roboflow 100. *stomata cells Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/stomata-cells.

[14] Roboflow 100. *thermal cheetah Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/thermal-cheetah-my4dp.

[15] Roboflow 100. *trail camera Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/trail-camera.

[16] Roboflow 100. *truck movement Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/roboflow-100/truck-movement.

[17] Roboflow 100. *underwater objects Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/underwater-objects-5v7p8.

[18] Roboflow 100. *wine labels Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roboflow-100/wine-labels.

[19] Roboflow 100-VL. *dentalai-i4clz-fsuo Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/rf100-vl/dentalai-i4clz-fsuo-ung2d.

[20] Roboflow 100-VL. *recode-waste-czvmg-yxsw Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/rf100-vl/recode-waste-czvmg-yxsw-fj9b9.

[21] Roboflow 100-VL. *speech-bubbles-detection-r22zt-ou0u6-jols Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/rf100-vl/speech-bubbles-detection-r22zt-ou0u6-jols.

[22] WorBots 4145. *2024 FRC Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/worbots-4145/2024-frc.

[23] abdulla. *grapes-5 Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/abdulla-ooj7n/grapes-5.

[24] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[25] apron activities. *aircraft turnaround dataset Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/apron-activities/aircraft-turnaround-dataset.

[26] Airport. *13 Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/airport-j6ozu/13-lkc01.

[27] AutoClashRoyale. *ClashRoyaleCharDetector Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/autoclashroyale/clashroyalechardetector.

[28] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. "Qwen2. 5-vl technical report". In: *arXiv preprint arXiv:2502.13923* (2025).

[29] BEPROJ. *Varroa mites detection (test-set) Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/beproj/varroa-mites-detection--test-set.

[30] Exploratorium + BioBus. *Exploratorium Daphnia Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/exploratorium-biobus/exploratorium-daphnia.

[31] Onno Bos. *actions Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/onno-bos-hdinj/actions-zzid2.

[32] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. "nuscenes: A multimodal dataset for autonomous driving". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[33] CAU. *water meter Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/cau-r0usm/water-meter-jbktv.

[34] Nadine Chang, Francesco Ferroni, Michael J Tarr, Martial Hebert, and Deva Ramanan. "Thinking Like an Annotator: Generation of Dataset Labeling Instructions". In: *arXiv preprint arXiv:2306.14035* (2023).

[35] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. "Are we on the right way for evaluating large vision-language models?" In: *arXiv preprint arXiv:2403.20330* (2024).

[36] Qiang Chen, Xiangbo Su, Xinyu Zhang, Jian Wang, Jiahui Chen, Yunpeng Shen, Chuchu Han, Ziliang Chen, Weixiang Xu, Fanrong Li, et al. "LW-DETR: a transformer replacement to yolo for real-time detection". In: *arXiv preprint arXiv:2406.03459* (2024).

[37] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[38] Floriana Ciaglia, Francesco Saverio Zuppichini, Paul Guerrie, Mark McQuade, and Jacob Solawetz. "Roboflow 100: A rich, multi-domain object detection benchmark". In: *arXiv preprint arXiv:2211.13523* (2022).

[39] Intelligent Digital Communications. *ISM Band Packet Detection Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/intelligent-digital-communications/ism-band-packet-detection.

[40] corrosion. *screw detect classification Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/corrosion-a1pkl/screw_detect_classification.

[41] CountingPills. *CountingPills Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/countingpills-rbjwo/countingpills.

[42] cse499. *Defect Detection Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/cse499/defect-detection-yjplx.

[43] G.M. Sangiovanni D. Ventura. *Sea cucumbers new tiles Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/olo-dan/sea-cucumbers-new-tiles.

[44] D4MS. *marine-sharks Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/d4ms/marine-sharks.

[45] DataCatalogue. *Macro segmentation Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/datacatalogue/macro-segmentation.

[46] My datasets. *Crystal Clean: Brain Tumors MRI Dataset Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/my-datasets-k2sei/crystal-clean-brain-tumors-mri-dataset.

[47] Google DeepMind. *Introducing Gemini 2.0: our new AI model for the agentic era*. Dec. 2024. URL: https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

[48] Detect Detect. *New-Defects in Wood Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/detect-detect/new-defects-in-wood.

[49] Worker Detect. *Conveyor T-shirts Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/worker-detect/conveyor-t-shirts.

[50] Pavement Distresses Detection. *Asphalt Distress Detection Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/pavement-distresses-detection/asphalt_distress_detection.

[51] Dorna. *tube Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/dorna/tube-4rv8o.

[52] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. "Learning to prompt for open-vocabulary object detection with vision-language model". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 14084–14093.

[53] Edesak. *NIH-Xray Dataset*. Open Source Dataset. 2021. URL: https://universe.roboflow.com/edesak/nih-xray.

[54] Elmir. *Pig detection Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/elmir/pig-detection-kaimq.

[55] NCRAAI Mehran University of Engineering, Technology Jamshoro, and University of Malaga Spain. *Wheel Defect Detection Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/ncraai-mehran-university-of-engineering-and-technology-jamshoro-and-university-of-malaga-spain/wheel-defect-detection-e53jb.

[56] Fieldlytics. *Everday New Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/fieldlytics-sn6h2/everday_new.

[57] Pukyong Univ master Fisheries of science. *Jellyfish Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/pukyong-univ-master-fisheries-of-science/jellyfish-pqc6u.

[58] Forreport. *Canal Stenosis Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/forreport/canal_stenosis.

[59] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. "Clip-adapter: Better vision-language models with feature adapters". In: *International Journal of Computer Vision* 132.2 (2024), pp. 581–595.

[60] garbagedetection. *Floating Waste Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/garbagedetection-gqqya/floating-waste.

[61] GDIT. *Aerial Airport Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/gdit/aerial-airport.

[62] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. "Open-vocabulary object detection via vision and language knowledge distillation". In: *arXiv preprint arXiv:2104.13921* (2021).

[63] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. "Open-vocabulary object detection via vision and language knowledge distillation". In: *arXiv preprint arXiv:2104.13921* (2021).

[64] wei guo. *GWHD2021 Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/wei-guo/gwhd2021.

[65] Agrim Gupta, Piotr Dollar, and Ross Girshick. "LVIS: A dataset for large vocabulary instance segmentation". In: *CVPR*. 2019.

[66] Chung H. *all-elements Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/chung-h/all-elements.

[67] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.

[68] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "Lora: Low-rank adaptation of large language models". In: *arXiv preprint arXiv:2106.09685* (2021).

[69] ImageEnhancement. *SSS OD Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/imageenhancement/sss_od.

[70] Inpaklijn. *fruitjes Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/inpaklijn/fruitjes.

[71] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. *Ultralytics YOLO*. Version 8.0.0. Jan. 2023. URL: https://github.com/ultralytics/ultralytics.

[72] joywolves. *ball Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/joywolves/ball-qgqhv.

[73] Gaoussou Youssouf Kebe, Padraig Higgins, Patrick Jenkins, Kasra Darvish, Rishabh Sachdeva, Ryan Barron, John Winder, Donald Engel, Edward Raff, Francis Ferraro, and Cynthia Matuszek. "A Spoken Language Dataset of Descriptions for Speech-Based Grounded Language Learning". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 2021.

[74] Rahima Khanam and Muhammad Hussain. "Yolov11: An overview of the key architectural enhancements". In: *arXiv preprint arXiv:2410.17725* (2024).

[75] Mehar Khurana, Neehar Peri, Deva Ramanan, and James Hays. "Shelf-Supervised Multi-Modal Pre-Training for 3D Object Detection". In: *arXiv preprint arXiv:2406.10115* (2024).

[76] kolly. *COD MW Warzone Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/kolly-ku5ew/cod-mw-warzone.

[77] Akshay Krishna. *INBreast Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/akshay-krishna-gie3b/inbreast-zzlbj.

[78] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. "F-vlm: Open-vocabulary object detection upon frozen vision and language models". In: *arXiv preprint arXiv:2209.15639* (2022).

[79] DataCluster Labs. *Car Logo Detection Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/datacluster-labs-agryi/car-logo-detection-cxyfl.

[80] Salo Levy. *The Dreidel Project Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/salo-levy-nlqrn/the-dreidel-project.

[81] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. "Seed-bench: Benchmarking multimodal llms with generative comprehension". In: *arXiv preprint arXiv:2307.16125* (2023).

[82] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. "ELEVATER: A Benchmark and Toolkit for Evaluating Language-Augmented Visual Models". In: *Neural Information Processing Systems* (2022).

[83] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. "Grounded language-image pre-training". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10965–10975.

[84] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. "A Survey of Multimodel Large Language Models". In: *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*. 2024, pp. 405–409.

[85] 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: *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer. 2014, pp. 740–755.

[86] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection". In: *arXiv preprint arXiv:2303.05499* (2023).

[87] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. "Mmbench: Is your multi-modal model an all-around player?" In: *European conference on computer vision*. Springer. 2024, pp. 216–233.

[88] Liver. *Liver diseases Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/liver-t5yvf/liver-diseases.

[89] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. "Learn to explain: Multimodal reasoning via thought chains for science question answering". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 2507–2521.

[90] Yechi Ma, Neehar Peri, Shuoquan Wei, Wei Hua, Deva Ramanan, Yanan Li, and Shu Kong. "Long-Tailed 3D Detection via 2D Late Fusion". In: *arXiv preprint arXiv:2312.10986* (2023).

[91] Anish Madan, Neehar Peri, Shu Kong, and Deva Ramanan. "Revisiting few-shot object detection with vision-language models". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 19547–19560.

[92] MBCET MajorProject. *Human Detection in Floods Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/mbcet-majorproject/human-detection-in-floods-a6aun.

[93] University Science Malaysia. *Label Printing Defect Version 2 Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/university-science-malaysia/label-printing-defect-version-2.

[94] MALEK. *Tomatoes 2 Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/malek-hnaln/tomatoes-2.

[95] AI For Mankind. *Wildfire Smoke Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/brad-dwyer/wildfire-smoke.

[96] Sachit Menon and Carl Vondrick. "Visual Classification via Description from Large Language Models". In: *The Eleventh International Conference on Learning Representations (ICLR)*. 2023.

[97] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. "Scaling Open-Vocabulary Object Detection". In: *arXiv preprint arXiv:2306.09683* (2023).

[98] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. "Simple open-vocabulary object detection". In: *European Conference on Computer Vision*. Springer. 2022, pp. 728–755.

[99] Roman Nguyen. *Mahjong Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/roman-nguyen/mahjong-vtacs.

[100] Than Quan Nguyen. *SMD Components Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/than-quan-nguyen/smd-components-dnljh.

[101] Zhaosi Nicholas. *pill Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/zhaosi-nicholas-z2xrh/pill-j8vgy.

[102] OrionProducts. *OrionProducts Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/orionproducts/orionproducts.

[103] Aljosa Osep, Tim Meinhardt, Francesco Ferroni, Neehar Peri, Deva Ramanan, and Laura Leal-Taixe. "Better Call SAL: Towards Learning to Segment Anything in Lidar". In: *ECCV*. 2024.

[104] Vale Outdoors. *Into the Vale Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/vale-outdoors/into-the-vale.

[105] Hongpeng Pan, Shifeng Yi, Shouwei Yang, Lei Qi, Bing Hu, Yi Xu, and Yang Yang. "The Solution for CVPR2024 Foundational Few-Shot Object Detection Challenge". In: *arXiv preprint arXiv:2406.12225* (2024).

[106] Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. "The Neglected Tails in Vision-Language Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 12988–12997.

[107] Penguins471. *penguin-finder-seg Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/penguins471/penguin-finder-seg.

[108] Neehar Peri, Achal Dave, Deva Ramanan, and Shu Kong. "Towards Long-Tailed 3D Detection". In: 2023.

[109] Onboarding Project. *Buoy Onboarding Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/onboarding-project/buoy-onboarding.

[110] project-ujdhs. *UAV small object detection Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/project-ujdhs/uav-small-object-detection.

[111] RF Projects. *X-Ray ID Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/rf-projects/x-ray-id.

[112] Robin Public. *Electric-Pylon-Detection-in-RSI Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/robin-public/electric-pylon-detection-in-rsi.

[113] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

[114] Research. *APOCE (Aerial Photographs for Object Detection of Construction Equipment) Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/research-s0une/apoce-aerial-photographs-for-object-detection-of-construction-equipment.

[115] Riis. *Aerial Sheep Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/riis/aerial-sheep.

[116] Roboflow. *Aquarium Combined Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/brad-dwyer/aquarium-combined.

[117] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115 (2015), pp. 211–252.

[118] RySEAI. *Lacrosse-Object-Detection Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/ryseai/lacrosse-object-detection.

[119] Inkyu Sa. *deepFruits mango Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/inkyu-sa-e0c78/deepfruits-mango.

[120] Saxion2. *Roboflow-Trained-Dataset Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/saxion2/roboflow-trained-dataset.

[121] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. "Laion-5b: An open large-scale dataset for training next generation image-text models". In: *Advances in neural information processing systems* 35 (2022), pp. 25278–25294.

[122] teeth segmentation. *ufba-425 Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/teeth-segmentation/ufba-425.

[123] sgcortes. *ZebrasatAsturias Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/sgcortes/zebrasatasturias.

[124] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. "Objects365: A Large-Scale, High-Quality Dataset for Object Detection". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 8429–8438. DOI: 10.1109/ICCV.2019.00852.

[125] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 2556–2565.

[126] SoftTeacherrail. *L10 UL 50 2 Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/softteacherrail-xxepi/l10_ul_50_2.

[127] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. "A simple semi-supervised learning framework for object detection". In: *arXiv preprint arXiv:2005.04757* (2020).

[128] spine1000. *spine frx normal vindr Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/spine1000-fqdvf/spine_frx_normal_vindr.

[129] Augmented Startups. *Football-Player-Detection Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/augmented-startups/football-player-detection-kucab.

[130] Ayca Takmaz, Cristiano Saltori, Neehar Peri, Tim Meinhardt, Riccardo de Lutio, Laura Leal-Taixe, and Aljosa Osep. "Towards Learning to Complete Anything in Lidar". In: *International Conference on Machine Learning (ICML)*. 2025.

[131] PSG College of Technology. *Urine analysis1 Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/psg-college-of-technology/urine-analysis1.

[132] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. "Winoground: Probing vision and language models for visio-linguistic compositionality". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5238–5248.

[133] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. "Eyes wide shut? exploring the visual shortcomings of multimodal llms". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 9568–9578.

[134] Phongsakhon Tongcham. *BIB Detection Dataset*. Open Source Dataset. 2025. URL: https://universe.roboflow.com/phongsakhon-tongcham/bib_detection.

[135] Mohamed Traore. *TACO: Trash Annotations in Context Dataset Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/mohamed-traore-2ekkp/taco-trash-annotations-in-context.

[136] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features". In: *arXiv preprint arXiv:2502.14786* (2025).

[137] TSF. *dataconvert Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/tsf/dataconvert.

[138] UNi. *Needle Base Tip Min Max Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/uni-tb9sq/needle-base-tip-min-max.

[139] Leo Unit. *xray Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/leo-unit/xray-2vqog.

[140] UNITUS. *WB-prova Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/unitus-vj7wf/wb-prova.

[141] chungnam university. *GRCCS Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/chungnam-university-lprwc/-grccs.

[142] Shanghai Jiao Tong University. *Infraredimageofpowerequipment Dataset*. Open Source Dataset. 2024. URL: https://universe.roboflow.com/shanghai-jiao-tong-university-xwhvl/infraredimageofpowerequipment.

[143] varun. *Invoice Processing Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/varun-yvyoy/invoice-processing-nl2cz.

[144] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. "A comprehensive survey of continual learning: Theory, method and application". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[145] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. "Frustratingly Simple Few-Shot Object Detection". In: *International Conference on Machine Learning (ICML)*. 2020.

[146] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. "Growing a brain: Fine-tuning by increasing model capacity". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2471–2480.

[147] weedsdetection. *weeds4 Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/weedsdetection-nqzsa/weeds4-evltl.

[148] Tack hwa Wong. *DeepPCB Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/tack-hwa-wong-zak5u/deeppcb-4dhir.

[149] New Workspace. *train Dataset*. Open Source Dataset. 2022. URL: https://universe.roboflow.com/new-workspace-d4ab7/train-i4unu.

[150] New WS. *org harvest Dataset*. Open Source Dataset. 2023. URL: https://universe.roboflow.com/new-ws/org_harvest.

[151] Wenhao Wu, Zhun Sun, and Wanli Ouyang. "Revisiting classifier: Transferring vision-language models for video recognition". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 37. 3. 2023, pp. 2847–2855.

[152] Yifan Xu, Mengdan Zhang, Chaoyou Fu, Peixian Chen, Xiaoshan Yang, Ke Li, and Changsheng Xu. "Multi-modal queried object detection in the wild". In: *Advances in Neural Information Processing Systems* 36 (2024).

[153] Weihao Xuan, Qingcheng Zeng, Heli Qi, Junjue Wang, and Naoto Yokoya. "Seeing is Believing, but How Much? A Comprehensive Analysis of Verbalized Calibration in Vision-Language Models". In: *arXiv preprint arXiv:2505.20236* (2025).

[154] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. "Meta r-cnn: Towards general solver for instance-level low-shot learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 9577–9586.

[155] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. "Modeling context in referring expressions". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer. 2016, pp. 69–85.

[156] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. "Mm-vet: Evaluating large multimodal models for integrated capabilities". In: *arXiv preprint arXiv:2308.02490* (2023).

[157] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 9556–9567.

[158] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 3836–3847.

[159] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. "Tip-adapter: Training-free clip-adapter for better vision-language modeling". In: *arXiv preprint arXiv:2111.03930* (2021).

[160] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. "Regionclip: Region-based language-image pretraining". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16793–16803.

[161] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. "Detecting twenty-thousand classes using image-level supervision". In: *European Conference on Computer Vision*. Springer. 2022, pp. 350–368.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We justify our main claims with experimental evidence.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include a discussion of limitations in Section 4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We do not provide any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include an implementation details section in the supplement.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We host all of our data on Roboflow and code on Github

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We include a section on implementation details in the supplement.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We include an experiment analyzing the variance of different model types in the supplement.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include a section on implementation details in the supplement.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work addresses fine-tuning foundation models, which can have significant societal impact as they may amplify biases in the data.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The data and models we evaluate do not have a high risk for misuse. Roboflow validates data uploaded to its platform.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets in Roboflow are permissible for commerical use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our GitHub repository provides details on how to access and use our dataset.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not use any data from human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not perform any human-subject research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We describe our proposed method in detail in Section 3.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Implementation Details

We present additional implementation details to reproduce our baseline experiments below. Our code is available on GitHub.

Detic. We use Detic [161] with a SWIN-L backbone for all zero-shot experiments. Additionally, we use the model checkpoint trained on LVIS, COCO and ImageNet-21K. We use class names provided as text prompts for Detic's CLIP classifier.

GroundingDINO. We use GroundingDINO [86] with pretrained weights from mmdetection (MM-GroundingDINO-L*). We prompt the model with all the class names combined into a single prompt. We fine-tune GroundingDINO on each few-shot dataset for 1000 iterations with a batch size 4 and learning rate of 3e-4. We resize all images to (640, 1333) and don't use any additional data augmentations.

MQ-GLIP. MQ-Det [152] proposes a learnable module that enables multi-modal prompting. We choose GLIP with a SWIN-L backbone as the underlying detection model for our experiments. We use the model checkpoint trained on Objects365, FourODs, GoldG, and Cap24M. Laslty, we use class names as the text prompts and few-shot visual examples as visual prompts.

OWLv2. We use OWLv2 [97] as implemented in HuggingFace. We prompt the model with each class name independently.

Qwen-2.5VL. We conduct all experiments using the "qwen2.5-vl-72b-instruct" model via API. We prompt the model based on guidelines from Qwen's official documentation. We also improve the base prompt through small-scale validation on multiple datasets and select the best prompt:

System Prompt

"You are a helpful assistant capable of object detection."

Multi-Class Detection Prompt

"Locate all of the following objects: {class names} in the image and output the coordinates in JSON format."

Single-Class Detection Prompt

"Locate every {class name} in the image and output the coordinates in JSON format."

Gemini 2.5 Pro. We conduct all experiments using the Gemini API with the "gemini-2.5-pro-preview-03-25" model. We prompt the model based on guidelines from Gemini's official documentation, but also improve the base prompt through small-scale validation on multiple datasets and select the best prompt:

System Prompt

"Return bounding boxes as a JSON array with labels. Never return masks or code fencing."

Multi-Class Detection Prompt

"Detect the 2d bounding boxes of the following objects: {class names}"

Single-Class Detection Prompt

"Detect all 2d bounding boxes of {class name}."

Prompting with Rich Textual Descriptions To evaluate Qwen and Gemini with dataset-specific annotator instructions, we appended the following prompt after our main prompt:

"Use the following annotator instructions to improve detection accuracy: {annotator instructions}"

We include the rich textual description for all classes when using the multi-class detection prompt. In contrast, we only append the relevant class description (extracted using GPT-4o) when using the single-class detection prompt.

Prompting with Few-Shot Visual Examples We provide one image at a time to Qwen and Gemini to mimic their turn-based pre-training. We use all few-shot images when prompting Gemini. However, we only use three images when prompting Qwen due to API limitations.

We prompt Gemini with native resolution images, but limit Qwen's few-shot visual examples to a minimum of 4*28*28 pixels and a maximum of 12800*28*28 pixels due to API limitations. To manage costs, we limit Gemini to only output 8192 tokens per request. We do not set any token limits for Qwen. Lastly, we implement a robust parser to handle minor JSON formatting errors. In some cases with many few-shot image examples, the API fails to return a valid response for requests of excessive size. In such cases, we simply assign a score of 0 AP for those images. Due to Gemini and Qwen not always predicting a confidence score for their bounding boxes, we set it to 1.0 by default.

YOLOv8 and YOLOv11. We train our YOLOv8 [71] and YOLOv11 [74] family of models using the Ultralytics package with default parameters. For all models, we follow the established protocol in Ciaglia et. al. [38] and train for 100 epochs with a batch size of 16. However, we evaluate all YOLO models using pycocotools instead of Ultralytics (cf. Appendix B)

B Additional Evaluation Details

We find that metrics reported with pycocotools (500 maxDets) differs significantly from those reported by Ultralytics on RF100-VL (cf. Table 4). Notably, all YOLO models report metrics using Ultralytics' implementation of mAP by default. Our preliminary investigation, supported by similar observations on Github, suggest that this disparity can be largely attributed to differences in the integration method of the precision-recall curve. Ultralytics uses a trapezoidal sum, which inflates model performance by as much 2.7% compared to pycocotools. We choose to report results for YOLO models using pycocotools in the main paper to standardize our results with our other baselines.

Table 4: **Impact of Evaluation Toolkit on RF100-VL Performance.** We find that the Ultralytics mAP calculation significantly over-estimates mAP compared with pycocotools. For fair comparison with other baselines, we choose to report metrics using pycocotools.

Method	pycocotools mAP (Ours)	Ultralytics mAP
YOLOv8n [71]	55.4	57.2
YOLOv11n [74]	56.1	57.8
YOLOv8s [71]	56.5	59.0
YOLOv11s [74]	57.0	59.4
YOLOv8m [71]	56.9	59.6
YOLOv11m [74]	57.0	59.6

C Ablation on Prompting MLLMs

We evaluate Gemini 2.5 Pro and Qwen 2.5-VL performance on RF100-VL using two prompting strategies: single-class prompting and multi-class prompting. The single-class prompting strategy separately performs a forward pass for each class and merges the results per image. The multi-class prompting strategy performs a single forward pass for all classes. Both Gemini 2.5 Pro and Qwen

2.5-VL recommend the single-class prompting strategy. Importantly, we do not perform non-maximal suppression for either strategy as both MLLMs do not report confidence scores per box.

Interestingly, we observe that Qwen2.5VL performs better with single-class prompts, while Gemini performs better with multi-class prompts. We posit that this can be attributed to Qwen's extensive referential object detection pre-training, which typically requires detecting a single class. In contrast, Gemini achieves better performance with multi-class prompting, which is more aligned with traditional object detection setups. We argue that multi-class prompting should be the default for assessing a MLLM's object detection capabilities since this more closely mirrors standard object detection protocols.

Table 5: **Analysis of Prompting Strategy.** MLLMs typically evaluate detection performance with single-class prompts. We find that Qwen2.5VL achieves better performance with single-class prompts, while Gemini 2.5 Pro achieves better performance with multi-class prompts. We advocate for multi-class prompting since this more closely matches object detection evaluation.

Method	Single-Class Prompt	Multi-Class Prompt
Gemini 2.5 Pro	8.0	11.6
Qwen 2.5-VL (72B)	7.8	5.6

D Comparing Different Model Sizes

In Figure 6, we evaluate the performance of the Gemini model family over time (e.g. Gemini Flash 2.0 was released before Gemini Flash 2.5). Although Gemini has not been explicitly fine-tuned on RF100-VL, we see a significant increase in performance. This suggests that Gemini is making real progress towards zero-shot open-vocabulary object detection in the wild. Unsurprisingly, base MLLMs outperform faster distilled models (e.g. Gemini 2.5 Pro achieves 35% better performance than Gemini Flash 2.5), but distilled models provide considerably better performance per dollar. Importantly, all models are prompted with multi-class prompts (cf. Appendix C).

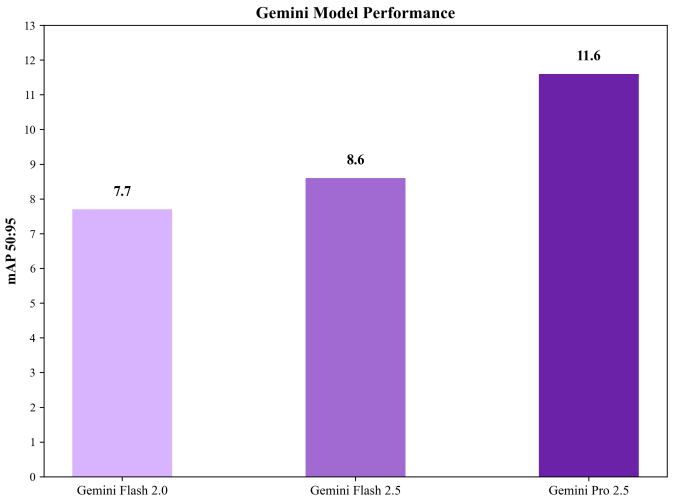


Figure 6: **Gemini Improves on RF100-VL over Time**. Despite not explicitly fine-tuning on RF100-VL, we find that newer Gemini models consistently improve over older models on our benchmark. This suggests that Gemini is making real progress towards improving zero-shot open-vocabulary detection in-the-wild.

E Ablation on Few-Shot Split Selection

Prior work typically selects few-shot training examples at random. However, Madan et. al. [91] demonstrates that the specific few-shot examples used for fine-tuning greatly affects target class performance. Specifically, Madan et. al. selects the most informative K-shot examples for each class in nuImages [32] by evaluating Detic w/ Federated Fine-Tuning's class-wise performance on a held-out validation set. For instance, in a 5-shot task with three random splits, we may select our five-shot car examples from split 1, our five-shot bicycles from split 3, and our five-shot debris from

split 2 based on which split has the highest per-class accuracy. As shown in Table 6, this "best split [91]" approach consistently outperforms random selection.

Despite the effectiveness of this approach, it has two primary limitations. First, it uses the validation performance of a specific model to inform few-shot selection. This inherently biases the few-shot images towards a particular model. Next, Madan et. al.'s proposed algorithm is computationally expensive since it requires fine-tuning a model on many candidate few-shot splits. This approach is computationally infeasible with RF100-VL's 100 datasets.

To address these two issues, we propose a learning-free approach that leverages the key insight from Madan et. al's analysis: the "best" examples are typically large and unoccluded. Concretely, we generate K random candidate few-shot splits for each class and pick the split that has the largest average bounding box area. Similar to Madan et. al., in a 5-shot task with three random splits, we may select our five-shot car examples from split 1, our five-shot bicycles from split 3, and our five-shot debris from split 2. We evaluate our proposed sampling strategy on nuImages and find that this approach performs better than random, but underperforms Madan et. al.'s approach. Future work should consider more effective strategies for selecting the "best" few-shot examples for concept alignment.

Table 6: **"Best" Split Construction**. We evaluate the quality of few-shot example selection using a (1) random baseline, (2) Madan et al.'s "best split" approach, which chooses per-class few-shot examples based on Detic w/ Federated Fine-Tuning's validation accuracy, and (3) our proposed learning-free method that selects splits with the largest average bounding box area. While Madan et al.'s method performs the best, it is biased towards Detic and is computationally expensive. Our approach offers a tractable alternative that improves over the random baseline.

Approach	Average Precision (AP)			
	`All`	`Many`	`Medium`	`Few`
Detic (Zero-Shot) [161]	14.40	25.83	16.59	2.32
Detic w/ Federated Fine-Tuning *(5-shots, Random Split)*	16.58	27.12	19.71	4.13
Detic w/ Federated Fine-Tuning *(5-shots, Best Split [91])*	**18.30**	**28.66**	**21.81**	**5.56**
Detic w/ Federated Fine-Tuning *(5-shots, Best Split, Ours)*	16.94	28.41	20.32	3.45
Detic w/ Federated Fine-Tuning *(10-shots, Random Split)*	17.24	28.07	20.71	4.18
Detic w/ Federated Fine-Tuning *(10-shots, Best Split [91])*	**18.24**	**28.63**	22.00	**5.19**
Detic w/ Federated Fine-Tuning *(10-shots, Best Split, Ours)*	17.48	26.36	**22.42**	4.32

F Semi-Supervised and Fully Supervised Results

We present results from semi-supervised and fully-supervised baselines in Table 7. Importantly, these models are evaluated on the same data splits as our zero-shot and few-shot baselines. To construct the semi-supervised split, we randomly sample 10% of the training set.

Semi-Supervised Baselines. We evaluate variants of YOLO [71, 74] and YOLO with STAC [127] trained on 10% of each dataset in RF100-VL. STAC generates high-confidence pseudo-labels for localized objects in unlabeled images and updates the model by enforcing consistency through strong augmentations. We follow the training protocol defined by Sohn et. al. [127]. First, we train a teacher model on the labeled subset of the data. Then, we use the teacher model to pseudo-label the remaining unlabeled subset of the data. We keep all detections above a confidence C, where the confidence tuned to maximize the F1 score of the teacher model on a validation set. Finally, we combine the subset of data with true ground truth labels and the subset with pseudo-labels to form a training set for a student model of the same architecture. We train this student model until convergence with heavy augmentations. We use the same hyperparameters as our supervised YOLOv8 and YOLOv11 implementation. Because YOLO models already train with significant augmentation, we don't add any new augmentations for the student training.

Fully-Supervised Baselines. We benchmark YOLOv8 [71], YOLOv11 [74], and LW-DETR [36] on all datasets within RF100-VL. YOLOv8, developed by Ultralytics, builds on the YOLOv5 architecture with improvements in model scaling and architectural refinements. YOLOv11 adds more architecture improvements, and is primarily validated on COCO. LW-DETR is a lightweight detection transformer that outperforms YOLO models for real-time object detection, and is SOTA on the original Roboflow100 [38] dataset, the predecessor to RF100-VL. Its architecture consists of a ViT encoder, a projector, and a shallow DETR decoder. This baseline serves as an upper bound on

performance, though in rare cases, few-shot foundation models may surpass it when the target dataset only has a few examples.

Semi-Supervised Learners are Data Efficient. We find that leveraging simple semi-supervised learning algorithms like STAC [127] significantly improves model performance when learning with limited labels. In half (7 out of 14) of combinations of model size and data domain, semi-supervised learners improved mAP at least as much as stepping up a model size. For example, YOLOv8s (small) trained on 10% labeled data (and 90% STAC psuedo-labels) achieves better performance overall than YOLOv8m (medium) trained on just 10% labeled data.

Table 7: **Roboflow100-VL Semi-Supervised and Fully-Supervised Benchmark.** We find that semi-supervised learners are able to reach nearly 80% of the performance of fully supervised models using 10% labeled data.

Method	Aerial	Document	Flora & Fauna	Industrial	Medical	Sports	Other	All
Semi-Supervised (10% Labels)								
YOLOv8n [71]	32.8	35.1	41.5	50.7	30.2	29.4	37.6	39.3
YOLOv8n [71] w/ STAC [127]	35.3	38.5	43.6	51.5	31.5	32.1	40.1	41.2
YOLOv8s [71]	37.8	40.8	42.6	52.6	32.8	36.8	41.3	42.3
YOLOv8s [71] w/ STAC [127]	38.2	42.7	43.4	52.4	34.0	38.4	42.3	43.1
YOLOv8m [71]	37.8	40.5	41.3	52.9	33.4	41.4	42.1	42.5
YOLOv8m [71] w/ STAC [127]	37.5	42.5	43.2	52.7	34.2	40.2	43.5	43.3
Fully-Supervised								
YOLOv8n [71]	50.4	56.4	53.9	64.3	50.0	49.2	54.6	55.4
YOLOv11n [74]	51.2	58.4	54.8	64.6	50.3	49.2	55.5	56.1
YOLOv8s [71]	51.6	58.9	54.9	64.6	50.2	51.2	56.7	56.5
YOLOv11s [74]	53.1	58.8	55.5	64.7	50.3	52.0	57.4	57.0
LW-DETRs [36]	54.5	57.7	54.2	66.8	51.7	54.7	56.3	57.4
YOLOv8m [71]	52.5	59.9	55.1	64.7	49.5	52.6	57.5	56.9
YOLOv11m [74]	53.0	60.5	54.8	65.1	49.9	52.4	57.3	57.0
LW-DETRm [36]	57.1	60.1	56.7	68.2	52.8	57.5	61.0	59.8

G Analysis of Accuracy vs. Parameter Count

In Figure 7, we observe a counter-intuitive trend: larger models perform worse in our evaluations. This is likely due to the mismatch between general-purpose MMLMs and specialized object detectors. Despite being the largest model pre-trained on the most data, Qwen2.5-VL (72B) underperforms GroundingDINO in the zero-shot setting and is also considerably slower. Interstingly, we find that GroundingDINO fine-tuned on few-shot examples surpasses all YOLO models fine-tuned on few-shot examples, indicating that large pre-trained backbones enable more efficient fine-tuning in specialist models.

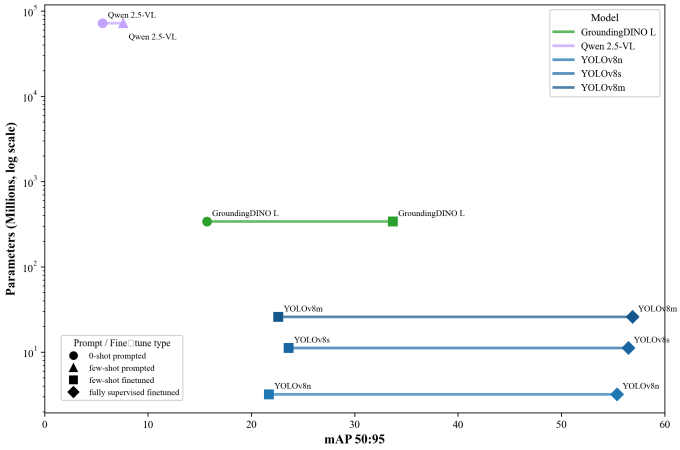


Figure 7: **Accuracy vs. Parameter Count.** Somewhat counterintuitively, we find that the model with the most parameters (Qwen2.5-VL 72B) performs worse than significantly smaller models pre-trained on less data (GroundingDINO) in the zero-shot setting. This suggests that generalist MLLMs are parameter inefficient for specialized tasks.

H Correlation Between Model Type and Per-Dataset Performance

Figure 8 presents four scatterplots comparing mAP 50:95 across different model pairs on RF100-VL, with each axis representing one model's mAP and each point labeled by a dataset index (sorted

alphabetically). These plots help identify whether certain datasets are universally easy, medium, or hard across models.

We compare Gemini vs. GroundingDINO, Qwen vs. GroundingDINO, Gemini vs. Qwen, and GroundingDINO vs. YOLO. Gemini and Qwen, as well as GroundingDINO and YOLO, show stronger linear correlations in their per-dataset scores, suggesting alignment in perceived difficulty. In contrast, comparisons between generalists (Gemini and Qwen) and specialists (GroundingDINO and YOLO) show weaker correlation. This suggests that large-scale MLLMs, likely trained on similar web data, align more closely with each other, while specialist models like GroundingDINO and YOLO show stronger consistency. These results imply that dataset difficulty levels (easy, medium, hard) may not generalize across model classes, but may be better defined within model types.

Additionally, among the top 15 datasets where Gemini outperforms Qwen and GroundingDINO, seven overlap. This suggests that Gemini may excel on datasets similar to those found in its pretraining, but struggles to generalize to novel domains.

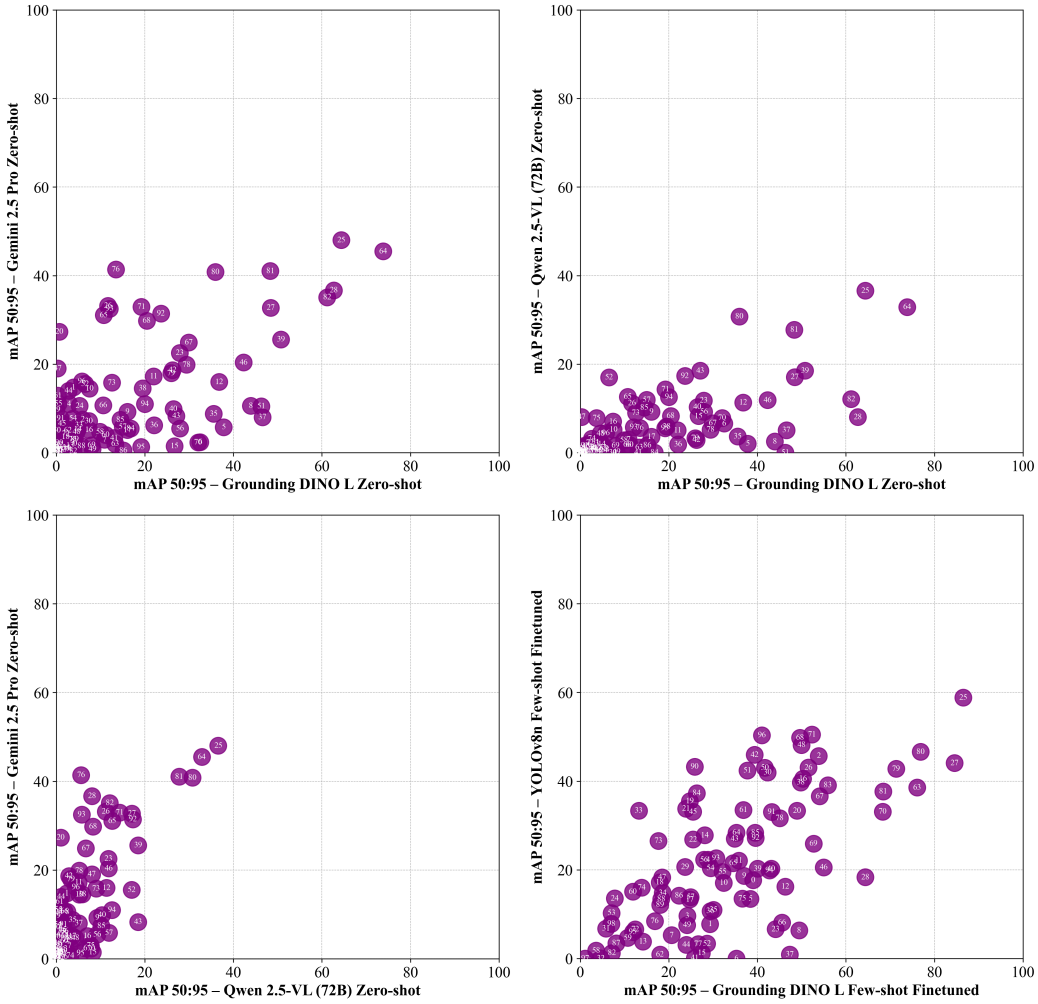


Figure 8: **Correlation Between Models Type and Performance.** We see stronger linear trends between Gemini and Qwen, and between GroundingDINO and YOLO, indicating aligned perceptions of dataset difficulty within model groups.

I Performance Variance for Few-Shot Models

Tables 8 and 9 measure the variance of YOLOv8 on RF100-VL. We use this model as a proxy for understanding few-shot learning variance and the statistical significance of our results. We train

YOLOv8n and YOLOv8s [71] with federated loss [91] ten times on each dataset, using ten different random seeds to determine model initialization and augmentation selection. We report the mean and standard deviation in two ways. In Table 8, we take the average mAP across all datasets in a given category (e.g. Industrial, Sports, All, etc.), and report the mean mAP and standard deviation across ten different runs. In Table 9, we measure the mean and standard deviation for each dataset across 10 different runs, and then report the average mean and standard deviation over each category. This will result in a higher standard deviation. Table 9 conveys the variance of a single dataset in RF100-VL and motivates averaging mAP across multiple datasets as a more stable metric.

Table 8: **Roboflow100-VL Overall Variance.** We evaluate the mean mAP and standard deviation of ten runs of YOLOv8 [71] with federated loss [91] over different subsets of Roboflow100-VL. These results can be used as a proxy to calculate whether a new entry to Table 2 is statistically significant. Unsurprisingly, averaging over 100 datasets yields a less noisy estimate of model performance

Method	Aerial	Document	Flora & Fauna	Industrial	Medical	Sports	Other	All
YOLOv8n [71]	13.5 ± .778	24.9 ± 1.19	20.8 ± .246	30.1 ± .409	15.9 ± .771	14.8 ± 1.270	21.8 ± .891	21.6 ± .230
YOLOv8s [71]	17.3 ± .791	26.4 ± .902	23.4 ± .407	30.0 ± .680	17.8 ± .565	19.2 ± .536	25.0 ± .252	23.7 ± .216

Table 9: **Roboflow100-VL Dataset Variance.** We evaluate the mean mAP and standard deviation over 10 runs of YOLOv8 [71] with federated loss [91] for each of the 100 datasets in Roboflow100-VL. These results helps quantify how much a model should improve on a single dataset to be statistically significant. This approach for quantifying statistical significance shows a much higher variance.

Method	Aerial	Document	Flora & Fauna	Industrial	Medical	Sports	Other	All
YOLOv8n [71]	13.5 ± 2.29	24.9 ± 2.65	20.8 ± 2.80	30.1 ± 2.48	15.9 ± 2.21	14.8 ± 2.07	21.8 ± 2.55	21.6 ± 2.50
YOLOv8s [71]	17.3 ± 2.25	26.4 ± 2.86	23.4 ± 3.24	30.0 ± 2.88	17.8 ± 2.43	19.2 ± 1.92	25.0 ± 2.44	23.7 ± 2.71

J Impact of Instruction Quality

We evaluate few-shot detection performance on RF20-VL using annotator instructions generated by GPT4o, Qwen 2.5-VL, Gemini 2.5 Pro, GPT4o with a human-in-the loop (our original instructions), and human written instructions in Table 10. We evaluate the impact of instruction source on Qwen 2.5 VL and Gemini 2.5 Pro. Notably, we do not find a clear correlation between instruction source and downstream model performance. We find that Qwen 2.5 VL achieves better performance with annotator instructions from all sources compared to class names only, while Gemini 2.5 Pro performs worse with annotator instructions from all sources compared to class names only. Somewhat surprisingly, we find that instructions from GPT 4o with a human-in-the-loop performs the best on both Qwen 2.5VL and Gemini 2.5 Pro, beating human written instructions. Although prompting with annotator instructions yields inconsistent benefits, future work should explore novel ways of incorporating such rich contextual information.

Table 10: **Impact of Instruction Origin.** We find that there is no strong correlation between instruction origin and MLLM detection accuracy.

Instruction Source	Qwen 2.5VL	Gemini 2.5 Pro
Class Names Only	5.1	11.5
GPT4o Instructions	6.4	5.3
Qwen 2.5 VL Instructions	6.4	4.7
Gemini 2.5 Pro Instructions	7.2	5.2
GPT-4o Instructions with Edits (Main Paper)	7.4	5.7
Human Written Instructions	6.6	4.4

On average, the class names only prompts had 31.75 words, the GPT4o instructions with a human-in-the-loop had 502 words, the shortened GPT4o instructions with a human-in-the-loop had 170.55 words, and the human instructions had 482.95 words. Somewhat surprisingly, we find that the length of the prompt does not correlate well with model performance (cf. Table 11). This suggests that the models are not fine-tuned to leverage such instructions, regardless of context length. Future work should consider more adaptive prompt designs or fine-tuning strategies.

Table 11: **Impact of Instruction Length.** We find that there is no strong correlation between instruction length and MLLM detection accuracy.

Instruction Source	Qwen 2.5VL	Gemini 2.5 Pro
Class Names Only	5.1	11.5
GPT-4o Instructions with Edits (Main Paper)	7.4	5.7
Shortened GPT-4o Instructions with Edits	5.9	5.4
Human Instructions	6.6	4.4

K Impact of Detector-Style Post-Processing with MLLMs

Unlike specialist detectors, MLLMs directly predict bounding boxes without confidence scores, and do not leverage common detector post-processing techniques like NMS. We investigate the impact of such post-processing steps on MLLM detection accuracy with RF20-VL in Table 12. First, we estimate the confidence score of each predicted bounding box with SigLIPv2 [136]. We compute the cosine similarity of the predicted class name text embedding and bounding box image crop embedding. Although one can prompt an MLLM to predict its own confidence scores in theory, recent work [153] demonstrates that MLLMs struggle to verbalize confidence estimates in practice. We expect that the challenging out-of-distribution classes in RF20-VL make directly verbalizing confidence estimates even more difficult. Next, we run NMS per-class. Importantly, these post-processing steps can be applied to both zero-shot and few-shot prompted MLLMs. We find that adding confidence scores from SigLIP significantly improves performance for both Qwen 2.5VL and Gemini 2.5 Pro. Further, NMS seems to have a negligible impact, suggesting the LLMs implicitly learn to avoid making duplicate predictions.

Table 12: **Impact of Detector-Style Post Processing.** We find that adding confidence scores from SigLIP significantly improves performance for both Qwen 2.5VL and Gemini 2.5 Pro, but NMS seems to have negligible impact.

Model	Qwen 2.5VL	Gemini 2.5 Pro
Instructions	7.4	5.7
+ SigLIPv2 Score	9.7	8.3
+ NMS	9.8	8.4
Instructions + Images	7.2	8.6
+ SigLIPv2 Score	8.8	10.6
+ NMS	8.9	10.6

L Analysis of Failure Cases

We can infer the causes of model failure by comparing the relative performance of standard object detectors (e.g. YOLOv8) with VLMs (e.g. Detic) for few-shot object detection. Importantly, unlike Detic, YOLOv8 is not pre-trained on large-scale datasets and is not promptable with class names. Therefore, when both models achieve low performance, we can attribute model failures to difficulties in feature extraction. In contrast, when YOLOv8 performs well, but Detic performance suffers, we can attribute model failures to semantic ambiguity.

Using this heuristic, we analyze YOLOv8m w/ Federated Loss and Detic w/ Federated Loss because they have similar overall performance on RF100-VL (cf. Table 2). Notably, we find that Detic achieves 19.6 AP on Documents, while YOLOv8m achieves 23.3 AP. This suggests that these datasets contain many semantically ambiguous classes. Similarly, Detic achieves 8.5 AP on Medical while YOLOv8m achieves 16.0 AP. In contrast, we posit that datasets where Detic outperforms YOLOv8 (like Flora & Fauna and Sports) are semantically unambiguous and more similar to Detic's pre-training.

M Summary of CVPR 2025 Competition Top Performers

We summarize the contributions of top teams below. We present full technical reports and code here.

BEATON uses Nebula-CV as the base detector, an unpublished model built on the DINO architecture with Swin-B as the visual backbone and BERT as the text encoder, enabling open-set detection through cross-modal fusion. The model is pre-trained in two stages: first on five million curated

web-scale images, then fine-tuned on one million high-quality grounding examples distilled from Qwen2.5-VL. To address the few-shot setting, they introduce strategies including optimized text prompts generated with Qwen2.5-VL, a carefully tuned combination of data augmentations (e.g., flip, crop, HSV augmentation, copy-paste), pseudo-labeling to supplement sparse annotations, and dataset-specific inference resolution selection. They also tested but found minimal benefit from federated fine-tuning and LLM-based post-processing.

FDUROILab proposes a structured fine-tuning strategy enhanced by aggressive data augmentation techniques such as CachedMosaic, YOLOXHSVRandomAug, CachedMixUp, and RandomCrop, which increase data diversity and model robustness. The team employs MM-GroundingDINO with a Swin-L backbone as the base detector and uses Qwen2.5-VL-32B for post-processing to refine classification results by correcting errors made by the primary detector. Training is conducted across 20 datasets, each undergoing 50 independent runs to ensure robust optimization. Their ablation study shows a stepwise improvement in performance, with significant gains from fine-tuning, additional augmentations, multiple training runs, and the MLLM-based post-processing.

NJUST-KMG integrates dynamic data augmentation, feature consistency regularization, a dynamic freezing mechanism, grid search optimization, and inference enhancements via Test-Time Augmentation (TTA) and Weighted Boxes Fusion (WBF). The augmentation pipeline dynamically adjusts the probabilities of CachedMosaic, MixUp, HSV jitter, and RandomCrop based on training progression, while the freezing strategy customizes parameter updates depending on dataset size and domain similarity. NJUST-KMG also uses a grid search process to tune hyperparameters and configurations for each dataset to maximize validation mAP. During inference, predictions are refined by combining outputs from the top models using WBF with confidence calibration.

N Dataset Comparison

We present a detailed comparison of RF100-VL with related datasets in Table 13. Notably, RF100-VL is the only dataset to support rich textual descriptions and evaluates models in the zero-shot, few-shot, semi-supervised, and fully-supervised data regimes.

Table 13: **Dataset Comparison**. We compare the characteristics of RF100-VL with Roboflow100, COCO, LVIS, Objects365, and OdinW. Notably, RF100-VL is the only dataset with rich textual descriptions and evaluates models across data regimes.

Dataset	# Datasets	# Images	# Annotations	# Classes	Rich Textual Descriptions	Multi-Spectral Imagery	Zero-Shot Evaluation	Few-Shot Evaluation	Semi-Supervised Evaluation	Fully-Supervised Evaluation	Image Source
Roboflow100-VL	100	164,149	1,355,491	564	✓	✓	✓	✓	✓	✓	Roboflow Universe
Roboflow100 [38]	100	224,714	1,319,307	805	✗	✓	✗	✗	✗	✓	Roboflow Universe
COCO [85]	1	328,000	2,500,000	91	✗	✗	✗	✗	✗	✓	Flickr
LVIS (v0.5) [65]	1	82,000	745,000	1230	✗	✗	✗	✗	✗	✓	COCO
Objects365 [124]	1	638,000	10,101,000	365	✗	✗	✗	✗	✗	✓	Flickr
OdinW [82]	35	152,384	1,073,455	314	✗	✗	✓	✓	✗	✓	Roboflow Universe

O RF100-VL Bounding Box Annotation Refinement

We hired external contractors to refine RF100-VL's bounding box annotations according to a set of guidelines, described below. In total, 30 annotators spent 2168 hours validating annotations, with the authors performing additional quality control. Our annotation guidelines provide instructions for improving the quality of existing annotations across datasets. Our primary goal was to ensure consistent annotation style, with every possible instance of each class labeled by a single, tightly fitting bounding box.

The annotation refinement process emphasizes the following corrections:

- *Merged Bounding Boxes*: Annotators must ensure each object has its own bounding box, redrawing boxes that encompass multiple instances of an object.
- *Incomplete Bounding Boxes*: Bounding boxes must fully contain the entire object. If an object extends beyond the box's boundaries, the box needs to be expanded to include the whole object.

- *Missing Annotations*: It is crucial to identify and label every instance of each class within a dataset. This is the most challenging and important aspect of refinement. Annotators are advised to check the background for missing annotations.
- *Incorrectly Labeled Objects*: Bounding boxes around objects that do not belong to the specified class must be removed.
- *Wrong Class Names*: Annotators need to correct instances where class names are misassigned to objects (e.g., doors labeled as windows, or generic numerical labels instead of descriptive class names like "enemy" or "head").
- *Duplicated Bounding Boxes*: If the same object has multiple bounding boxes, one of the duplicates must be removed.

These instructions focus on correcting annotations within each dataset to achieve consistency. When inconsistencies are found in a dataset's labeling scheme, annotators are instructed to make a determination for consistency and ensure all images follow that scheme.

P Annotation Generation Instructions

We present our prompt for generating multi-modal annotator with GPT-4o below.

````
Pay attention to the following example annotation instructions for nu-images,
an object detection dataset:

{nuImages Annotator Instructions}

That was an example of object detection annotation instructions.

Using the above instructions as rough inspiration, come up with annotation
instructions for a dataset.

The annotation instructions should be in markdown format, and follow the
following outline:

```markdown
# Overview
Table of contents

# Introduction
Introduction to the dataset. Introduce what task the dataset is trying to
solve. List all of the classes and provide a brief description of each class.

# Object Classes
## Class 1
### Description
Provide a description of the class, paying attention to visually distinctive
elements of the class.
### Instructions
Provide detailed instructions for how to annotate this class. Give specific
references to the class, and pay attention to the example labeled images that
will be provided. Provide specific descriptions of what not to label, if applicable.
...

## Class 2
...

## Class n
...
```
````

```
Please pay specific attention to the provided visual example images and ground
your response in those examples. Be brief and concise, but comprehensive.
Make sure ### Instructions in each class provides visual descriptions of what
exactly to annotate.

Visual descriptions should make specific reference to how the object looks in
each image. If the object is not something everyone knows, describe its distinctive
shape, color, texture, etc. Look at the example pictures when coming up with these
instructions.

Respond with only the markdown content, no other text (and no backticks). Do not
describe the color of the bounding box, just describe how to find the spatial extent
of the object in the image.

The final markdown file should not make specific reference to the provided example
images. Those are simply to help you come up with the instructions. An annotator
should be able to recreate the annotations in the example images using your
generated instructions.

If the classes are similar, make sure the instructions specify how to disambiguate
between them (visually, which specific visual features to look for).

The visual content of the image should be used to clarify the description of each
class. Feel free to generalize about what is present in the dataset from the example
images.

Here is general metadata about the dataset:
{Metadata}

Here are the class names:
{Class Names}

Here are the example images:
{Few-Shot Example Images}
```

Q Sample Annotation Instructions

We present sample annotator instructions below. We use dataset metadata, class names and few-shot visual examples and prompt GPT-4o [24] to generate annotator instructions (cf. Appendix P). We then manually verify that the instructions accurately describe the few-shot examples. These annotator instructions are from recode-waste-czvmg-fsod-yxsw.

```
# Overview
- [Introduction](#introduction)
- [Object Classes](#object-classes)
  - [Aggregate](#aggregate)
  - [Cardboard](#cardboard)
  - [Hard Plastic](#hard-plastic)
  - [Metal](#metal)
  - [Soft Plastic](#soft-plastic)
  - [Timber](#timber)

# Introduction
This dataset is designed for waste classification within different material
classes. The goal is to accurately identify and annotate different types of
waste materials for sorting and recycling purposes. The classes represented
are: Aggregate, Cardboard, Hard Plastic, Metal, Soft Plastic, and Timber.
```

```
# Object Classes

## Aggregate
### Description
Aggregate refers to small, granular materials, often irregular in shape with
rough surfaces. They generally appear as pieces of stone or concrete.

### Instructions
Annotate all visible portions of aggregate items. Ensure to include entire
objects even if occluded by other materials, estimating boundaries if necessary.
Exclude dust or very fine particles that do not form distinct objects.

## Cardboard
### Description
Cardboard objects are typically flat and have a layered texture. They may appear
as boxes or sheets.

### Instructions
Annotate only distinguishable pieces of cardboard, focusing on their flat surfaces
and any visible layering. Do not annotate cardboard that is part of another object
or soiled beyond recognition.

## Hard Plastic
### Description
Hard plastics are rigid and maintain their shape. They can be cylindrical, tubular,
or robust objects often found in industrial contexts.

### Instructions
Annotate the entire visible area of hard plastic objects, ensuring to capture their
solid structure. Avoid labeling small, indistinct pieces or any plastic that appears
flexible.

## Metal
### Description
Metal objects are robust, often shiny or reflective. They can appear as rods, sheets,
or other distinct shapes.

### Instructions
Label all distinct metal parts, taking care to capture their complete form. Avoid
labeling rust marks or indistinct metallic fragments lacking shape.

## Soft Plastic
### Description
Soft plastics are flexible and often transparent or translucent. They may appear in
the form of bags or wrappers.

### Instructions
Focus on full pieces of soft plastic material, ensuring to include areas with visible
creases or folds indicating flexibility. Do not label pieces smaller than a
recognizable package or those mixed with other materials.

## Timber
### Description
Timber objects are wooden, either rough or smooth, often elongated or rectangular.

### Instructions
Annotate the entire visible portion of timber, focusing on the grain or wood texture.
Do not label splinters or fragments that do not exhibit a clear wooden structure.
```

R Roboflow100-VL Datasets

We present a table with links to all datasets within Roboflow100-VL (fully-supervised and FSOD datasets) below.

Flora & Fauna	Link
aquarium-combined [116]	FSOD, Fully Supervised
bees [3]	FSOD, Fully Supervised
deepfruits [119]	FSOD, Fully Supervised
exploratorium-daphnia [30]	FSOD, Fully Supervised
grapes-5 [23]	FSOD, Fully Supervised
grass-weeds [7]	FSOD, Fully Supervised
gwhd2021 [64]	FSOD, Fully Supervised
into-the-vale [104]	FSOD, Fully Supervised
jellyfish [57]	FSOD, Fully Supervised
marine-sharks [44]	FSOD, Fully Supervised
orgharvest [150]	FSOD, Fully Supervised
peixos-fish [10]	FSOD, Fully Supervised
penguin-finder-seg [107]	FSOD, Fully Supervised
pig-detection [54]	FSOD, Fully Supervised
roboflow-trained-dataset [120]	FSOD, Fully Supervised
sea-cucumbers-new-tiles [43]	FSOD, Fully Supervised
thermal-cheetah [14]	FSOD, Fully Supervised
tomatoes-2 [94]	FSOD, Fully Supervised
trail-camera [15]	FSOD, Fully Supervised
underwater-objects [17]	FSOD, Fully Supervised
varroa-mites-detection–test-set [29]	FSOD, Fully Supervised
wb-prova [140]	FSOD, Fully Supervised
weeds4 [147]	FSOD, Fully Supervised

Industrial	Link
-grccs [141]	FSOD, Fully Supervised
13-lkc01 [26]	FSOD, Fully Supervised
2024-frc [22]	FSOD, Fully Supervised
aircraft-turnaround-dataset [25]	FSOD, Fully Supervised
asphaltdistressdetection [50]	FSOD, Fully Supervised
cable-damage [4]	FSOD, Fully Supervised
conveyor-t-shirts [49]	FSOD, Fully Supervised
dataconvert [137]	FSOD, Fully Supervised
deeppcb [148]	FSOD, Fully Supervised
defect-detection [42]	FSOD, Fully Supervised
fruitjes [70]	FSOD, Fully Supervised
infraredimageofpowerequipment [142]	FSOD, Fully Supervised
ism-band-packet-detection [39]	FSOD, Fully Supervised
l10ul502 [126]	FSOD, Fully Supervised
needle-base-tip-min-max [138]	FSOD, Fully Supervised
recode-waste [20]	FSOD, Fully Supervised
screwdetectclassification [40]	FSOD, Fully Supervised
smd-components [100]	FSOD, Fully Supervised
truck-movement [16]	FSOD, Fully Supervised
tube [51]	FSOD, Fully Supervised
water-meter [33]	FSOD, Fully Supervised
wheel-defect-detection [55]	FSOD, Fully Supervised

Document	Link
activity-diagrams [1]	FSOD, Fully Supervised
all-elements [66]	FSOD, Fully Supervised
circuit-voltages [5]	FSOD, Fully Supervised
invoice-processing [143]	FSOD, Fully Supervised
label-printing-defect-version-2 [93]	FSOD, Fully Supervised
macro-segmentation [45]	FSOD, Fully Supervised
paper-parts [9]	FSOD, Fully Supervised
signatures [11]	FSOD, Fully Supervised
speech-bubbles-detection [21]	FSOD, Fully Supervised
wine-labels [18]	FSOD, Fully Supervised

Medical	Link
canalstenosis [58]	FSOD, Fully Supervised
crystal-clean-brain-tumors-mri-dataset [46]	FSOD, Fully Supervised
dentalai [19]	FSOD, Fully Supervised
inbreast [77]	FSOD, Fully Supervised
liver-disease [88]	FSOD, Fully Supervised
nih-xray [53]	FSOD, Fully Supervised
spinefrxnormalvindr [128]	FSOD, Fully Supervised
stomata-cells [13]	FSOD, Fully Supervised
train [149]	FSOD, Fully Supervised
ufba-425 [122]	FSOD, Fully Supervised
urine-analysis1 [131]	FSOD, Fully Supervised
x-ray-id [111]	FSOD, Fully Supervised
xray [139]	FSOD, Fully Supervised

Aerial	Link
aerial-airport [61]	FSOD, Fully Supervised
aerial-cows [61]	FSOD, Fully Supervised
aerial-sheep [115]	FSOD, Fully Supervised
apoce-aerial-photographs-for-object-detection-of-construction-equipment [114]	FSOD, Fully Supervised
electric-pylon-detection-in-rsi [112]	FSOD, Fully Supervised
floating-waste [60]	FSOD, Fully Supervised
human-detection-in-floods [92]	FSOD, Fully Supervised
sssod [69]	FSOD, Fully Supervised
uavdet-small [110]	FSOD, Fully Supervised
wildfire-smoke [95]	FSOD, Fully Supervised
zebrasatasturias [123]	FSOD, Fully Supervised

Sports	Link
actions [31]	FSOD, Fully Supervised
aerial-pool [2]	FSOD, Fully Supervised
ball [72]	FSOD, Fully Supervised
bibdetection [134]	FSOD, Fully Supervised
football-player-detection [129]	FSOD, Fully Supervised
lacrosse-object-detection [118]	FSOD, Fully Supervised

Other	Link
buoy-onboarding [109]	FSOD, Fully Supervised
car-logo-detection [79]	FSOD, Fully Supervised
clashroyalechardetector [27]	FSOD, Fully Supervised
cod-mw-warzone [76]	FSOD, Fully Supervised
countingpills [41]	FSOD, Fully Supervised
everdaynew [56]	FSOD, Fully Supervised
flir-camera-objects [6]	FSOD, Fully Supervised
halo-infinite-angel-videogame [8]	FSOD, Fully Supervised
mahjong [99]	FSOD, Fully Supervised
new-defects-in-wood [48]	FSOD, Fully Supervised
orionproducts [102]	FSOD, Fully Supervised
pill [101]	FSOD, Fully Supervised
soda-bottles [12]	FSOD, Fully Supervised
taco-trash-annotations-in-context [135]	FSOD, Fully Supervised
the-dreidel-project [80]	FSOD, Fully Supervised