REGIONDET: A BENCHMARK FOR REGION DETECTION BEYOND OBJECT INSTANCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Object detection is a fundamental task in computer vision and has achieved remarkable progress on standard benchmarks by localizing discrete and well-bounded object instances. However, many visual targets in real-world scenarios are not individual objects, but regions defined by visual states, scene context, object relations, and human activities, such as construction areas, damaged road regions, queues, group conversations, and vendor regions. Existing detection benchmarks are mainly built around object instances, providing limited support for systematically evaluating such region targets. To address this gap, we introduce **Region Detection**, a benchmark setting that extends conventional object detection beyond object instances, and construct **RegionDet**, a benchmark for region target localization. RegionDet contains eight region categories, including Construction, Crossing, Damage, Queuing, Talking, Vendor, Waiting, and Walking, with COCO-style bounding-box annotations and evaluation protocols. We evaluate representative closed-set and zero-shot/open-vocabulary detectors on RegionDet. Results show that closed-set detectors can partially learn region-level patterns under supervision, while zero-shot/open-vocabulary detectors remain limited even with region-oriented textual prompts, suggesting that current vision-language detectors still struggle to generalize to region targets. Further analyses highlight key challenges in Region Detection, including weak boundary cues, strong context dependency, and insufficient relation-level region understanding.

1 INTRODUCTION

Object detection is a fundamental task in computer vision, aiming to localize and recognize object instances in images. With the development of modern detection frameworks, remarkable progress has been achieved on standard benchmarks such as MS COCO Lin et al. (2014) and LVIS Gupta et al. (2019). However, most detection paradigms are built upon an object-centric assumption, where targets are typically discrete and countable entities with relatively clear physical boundaries, such as people, vehicles, animals, and everyday objects. Although this setting has greatly advanced generic object detection, many real-world localization needs go beyond discrete object instances.



(a) Object Detection



(b) Region Detection

Figure 1: Comparison between conventional object detection and Region Detection. (a) Object detection focuses on discrete and well-bounded object instances. (b) Region Detection localizes region targets whose spatial extents are defined by visual states, object relations, scene functions, and contextual cues beyond single-object boundaries.

In many real-world applications, visual systems need to localize not only individual object instances, but also regions associated with states, functions, activities, or events. For example, targets in urban streets, public spaces, and surveillance scenarios may include construction areas, damaged road regions, queues, group conversations, and vendor regions. Unlike standalone objects, these targets are defined by visual states, spatial relations, scene functions, human activities, and context, with spatial extents that may not align with clear object contours. For instance, a construc-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tion area is not determined by the boundary of any single worker, vehicle, or machine, but by the joint presence of workers, equipment, materials, road context, and ongoing activities. Similarly, a group conversation is not a person box, but a region determined by the proximity, pose orientation, and interaction among multiple participants. Fig. 1 illustrates the difference between conventional object-instance targets and region-level targets: the former correspond to well-bounded objects, whereas the latter are defined by states, relations, functions, activities, and context.

These region targets differ from conventional object instances in three key aspects: **(1) Weak boundary cues.** Their spatial extents are often not directly indicated by clear object contours, making localization difficult using standard boundary cues; **(2) Composite structure.** A region target may comprise multiple people, objects, and contextual elements rather than a single physical entity; **(3) Strong context dependency.** Many region targets depend on scene context and object relations and therefore cannot be reliably recognized from local appearance alone. This raises a key question: ***Can existing object detectors reliably localize and recognize targets whose identity and spatial extent are determined by visual states, scene context, and object interactions rather than the contour of a single object?***

To answer this question, we introduce and systematically study **Region Detection**, a box-level detection setting for localizing region targets beyond conventional object instances. Region Detection extends the target space from discrete and countable entities to regions defined by behavioral states, group relations, scene functions, and local abnormalities. It preserves the standard localization-and-classification formulation of object detection, enabling direct evaluation with existing detectors while better reflecting region localization needs in traffic monitoring, urban management, public safety, and scene understanding.

Based on this setting, we construct **RegionDet**, a benchmark dataset for Region Detection. RegionDet contains eight categories: Construction, Crossing, Damage, Queuing, Talking, Vendor, Waiting, and Walking. These categories span four representative types of region-forming cues: local abnormalities, human behaviors, group relations, and scene functions. To support consistent annotation and reproducible evaluation, we establish category-specific guidelines covering region extent, participant membership, and split/merge decisions, and annotate RegionDet using COCO-compatible bounding boxes and category labels. We further provide standard data splits and evaluation protocols, allowing RegionDet to be directly used with existing object detection frameworks. RegionDet thus provides a basis for systematically examining the limits of current detectors on region targets beyond conventional object instances.

Furthermore, we evaluate representative closed-set and zero-shot/open-vocabulary detectors across two-stage, one-stage, point/center-based, diffusion-based, query-based Transformer, and vision-language detection approaches. Experimental results show that RegionDet poses substantial challenges to existing detectors. Closed-set detectors can partially learn region-level patterns under supervised training, but their performance varies notably across models and categories. In contrast, zero-shot/open-vocabulary detectors remain limited even with region-oriented textual prompts, suggesting that current vision-language detectors still struggle to generalize from object-centric recognition to region targets defined by states, relations, and context.

2 Related Work

Object detection benchmarks. Large-scale benchmarks have been central to object detection. PASCAL VOC Everingham et al. (2010) established the standard box-level formulation, and MS COCO Lin et al. (2014) advanced the field with diverse scenes, scales, and instance annotations. Subsequent benchmarks and suites, such as Open Images Kuznetsova et al. (2020), Objects365 Shao et al. (2019), LVIS Gupta et al. (2019), ODinW Li et al. (2022a), and MSOSB Zhang et al. (2024), expanded category vocabulary, data scale, long-tailed distribution, and domain coverage. Despite these advances, their annotated targets remain largely object-centric: discrete, countable, and well-bounded instances such as persons, vehicles, animals, and everyday objects. In contrast, RegionDet shifts the target unit from standalone object instances to context- and relation-defined regions, such as construction areas, damaged regions, queues, group conversations, and vendor regions.

Region-level visual understanding. Beyond object detection, several tasks study visual regions with dense or higher-level semantics. Semantic segmentation datasets, such as PASCAL Con-

text Mottaghi et al. (2014), ADE20K Zhou et al. (2017), and Cityscapes Cordts et al. (2016), assign pixel-level semantic labels for dense scene parsing. Stuff and panoptic segmentation benchmarks, including COCO-Stuff Caesar et al. (2018) and COCO-Panoptic Kirillov et al. (2019), further cover amorphous regions such as road, sky, grass, and sidewalk. However, these tasks focus on pixel-level labeling of continuous object or stuff categories, whereas RegionDet adopts box-level annotations to localize semantic, functional, state, and activity regions. Many RegionDet targets, such as *Queuing*, *Talking*, and *Vendor*, may involve multiple people, objects, and contextual elements, and do not necessarily correspond to a single continuous mask region. RegionDet is also related to action and interaction understanding. Datasets such as AVA Gu et al. (2018), V-COCO Gupta & Malik (2015), and HICO-DET Chao et al. (2018) study human actions or human-object interactions, but usually focus on actor boxes, individual actions, or human-object-relation triplets. In contrast, RegionDet localizes the spatial extent of an activity, functional, or interaction region, which may include multiple participants, objects, and surrounding context. Thus, RegionDet provides a complementary benchmark for region-level localization beyond object instances.

Open-vocabulary and vision-language detection. Recent open-vocabulary and vision-language detectors aim to recognize novel categories specified by text descriptions. Representative methods, such as GLIP Li et al. (2022b), Grounding DINO Liu et al. (2024), OWL-ViT Minderer et al. (2022), YOLO-World Cheng et al. (2024), YOLOE Wang et al. (2025), and LLMDet Fu et al. (2025), leverage large-scale image-text pretraining or grounding data to align visual regions with language queries. These methods have shown promising generalization ability on conventional object detection and phrase grounding tasks. However, they are primarily trained and evaluated on object-like targets, and their text queries often correspond to nouns or noun phrases associated with individual object instances. It remains unclear whether such models can generalize to region targets defined by states, activities, group relations, and scene context. RegionDet provides a benchmark for evaluating this capability. Our experiments show that current zero-shot/open-vocabulary detectors remain limited on RegionDet even with region-oriented textual prompts, suggesting that vision-language detection models still face challenges in localizing region targets beyond object instances.

3 Region Detection and the RegionDet Benchmark

3.1 Problem Setting

We define **Region Detection** as a box-level detection setting for localizing visual regions whose meanings are determined by states, functions, activities, or events beyond conventional object instances. Given an input image I, the goal is to predict a set of region targets $\hat{\mathcal{Y}} = \{(\hat{b}_i, \hat{c}_i, \hat{s}_i)\}_{i=1}^{N}$, where $\hat{b}_i = (\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i)$ denotes the predicted bounding box, $\hat{c}_i \in \mathcal{C}$ is the predicted category, $\hat{s}_i$ is the confidence score, and N is the number of predictions. The label space $\mathcal{C}$ consists of region-level concepts rather than conventional object categories. For each image, the ground-truth annotations are represented as $\mathcal{Y} = \{(b_j, c_j)\}_{j=1}^{M}$, where b_j and c_j denote the bounding box and category of the j-th annotated region.



Figure 2: Representative examples of RegionDet. Region targets are defined by visual states, object relations, scene functions, and contextual cues rather than the physical contour of a single object.

Region Detection retains the standard box-level localization-and-classification formulation, making it compatible with existing detection frameworks and evaluation metrics. Its target unit, however, is not restricted to the physical contour of a single object. Instead, the identity and spatial extent of a region may be jointly determined by visual states,

Table 1: Comparison of RegionDet with related benchmark families.

Benchmark family	Annotation	Primary prediction unit	Main defining signal
COCO/LVIS	Box/Mask	Object instance	Physical object identity and contour
ADE20K/COCO-Stuff	Mask	Semantic pixel region	Object or stuff semantics
AVA/HICO-DET	Box/Triplet	Actor or human–object pair	Action or interaction label
Referring segmentation	Mask	Referred object or region	Linguistic expression
RegionDet	Box	Semantic region	State, relation, function, activity, and context

Table 2: Category definitions of RegionDet.

Category	Definition
Construction	Spatially coherent regions of construction activity involving workers, equipment, materials, and road context.
Crossing	Regions formed by pedestrians crossing roads or traffic-related spaces, as determined by pedestrian behavior and road context.
Damage	Local damaged or degraded regions, such as road damage and surface breakage.
Queuing	Regions formed by multiple people in a queue, as determined by their spatial arrangement and waiting behavior.
Talking	Regions of conversation or social interaction, defined by the proximity, orientation, and interaction among participants.
Vendor	Functional regions of street vending or trading involving vendors, goods, customers, and local context.
Waiting	Regions where people are waiting or remaining in place, inferred from posture, spatial arrangement, and scene context.
Walking	Regions formed by pedestrians walking together, as determined by motion states and spatial relations.

object interactions, group relations, scene functions, and surrounding context. Table 1 compares RegionDet with related benchmarks, while Fig. 2 shows representative examples.

3.2 Region Categories

RegionDet contains eight categories: *Construction*, *Crossing*, *Damage*, *Queuing*, *Talking*, *Vendor*, *Waiting*, and *Walking*. They cover four representative types of region-forming cues: local abnormalities (*Damage*), human behaviors (*Crossing*, *Waiting*, and *Walking*), group relations (*Queuing* and *Talking*), and scene functions (*Construction* and *Vendor*).

Unlike conventional object categories, the identity and spatial extent of these categories may be determined by multiple entities and contextual cues. For example, a *Talking* region is defined by the proximity, orientation, and interaction among participants, whereas a *Construction* region is jointly indicated by workers, equipment, materials, road context, and ongoing activities. Table 2 summarizes the category definitions, and Sec. 3.3 describes the operational annotation rules.

3.3 Data Collection and Annotation

Data collection. We collected candidate images from public web sources covering urban streets, public spaces, traffic scenes, and surveillance-like environments. Ten annotators participated in image collection, filtering, and initial annotation. Images were selected to cover diverse viewpoints, scales, densities, and contextual layouts. Low-quality, duplicated, irrelevant, or severely ambiguous images were removed.

Annotation unit and extent. Each region was assigned one of the eight predefined categories and annotated using the tightest axis-aligned bounding box enclosing the visible elements necessary to establish its region-level meaning. Relevant participants, objects, and contextual elements were included when they jointly defined the region, whereas unrelated background objects and bystanders were excluded. For group-related categories, participant membership was determined using spatial proximity, pose orientation, motion consistency, and observable interaction cues.

Split and merge rules. Spatially separated groups or activity centers were annotated as distinct regions. Nearby people or objects were merged into one region only when they jointly formed a coherent activity, interaction, or functional area. For example, a *Talking* box includes the interacting participants rather than an individual person, while a *Construction* box covers the locally coherent activity area formed by relevant workers, equipment, materials, and context. Disconnected damaged areas were annotated separately unless they formed a visually continuous damaged region. Cases without sufficient evidence for determining category identity or region membership were excluded.

Annotation completeness. Although candidate images were collected category-wise, each retained image was reviewed against the full eight-category ontology. All valid regions satisfying the annotation guidelines were annotated, allowing multiple categories to coexist within an image. The image counts in Table 3 refer to the category used during image collection, whereas region counts are computed from the final annotations.

Table 3: Category-wise statistics of RegionDet. Image counts are grouped by collection category.

Category	Images	Regions	Train Img.	Train Reg.	Test Img.	Test Reg.
Construction	1999	2031	1000	1016	999	1015
Crossing	401	415	201	208	200	207
Damage	1999	4739	1000	2437	999	2302
Queuing	409	426	204	212	205	214
Talking	406	574	203	257	203	317
Vendor	1995	3301	997	1673	998	1628
Waiting	400	571	200	320	200	251
Walking	401	531	200	248	201	283
Total	8010	12588	4005	6371	4005	6217

Quality control. To improve annotation consistency, all initial annotations underwent multi-annotator cross-checking by five additional annotators. The verification was conducted in two rounds. The first round independently examined category correctness, participant membership, split/merge decisions, and compliance with the region-extent guidelines. The second round focused on annotations flagged as ambiguous or inconsistent, removing invalid regions and refining inaccurate boxes. Disagreements were jointly re-examined with reference to the category-specific guidelines and resolved through consensus. Only annotations that passed both rounds of verification were retained in the final dataset.

3.4 Dataset Statistics

RegionDet contains 8,010 images and 12,588 annotated regions. It is divided into a training set with 4,005 images and 6,371 regions and a test set with 4,005 images and 6,217 regions. Table 3 reports the category-wise distribution.

We collected more samples for *Construction*, *Damage*, and *Vendor* because these categories exhibit greater appearance and scene variation. The interaction- and behavior-related categories are included to evaluate relation- and state-aware region understanding. Each category is divided approximately evenly between the training and test sets.

3.5 Evaluation Protocol

We use standard COCO-style detection metrics on the RegionDet test set. The primary metric is average precision (AP), averaged over all categories and IoU thresholds from 0.50 to 0.95 in increments of 0.05. We also report AP_{50}, AP_{75}, and per-category AP to analyze localization behavior under different overlap requirements.

For closed-set evaluation, models are trained using only the RegionDet training set and evaluated under the common eight-category label space. The test annotations are used only for final evaluation and are not accessed during model or checkpoint selection.

For zero-shot/open-vocabulary evaluation, models are evaluated without training or fine-tuning on RegionDet. We consider three text-query settings: bare category names, category definitions from Table 2, and concise region-oriented descriptions specifying relevant entities, relations, and context. The same queries are used across models within each setting, while input formatting follows the corresponding official implementations. All predictions are evaluated using the same COCO-style protocol. Because the closed-set and zero-shot settings use different supervision, their results characterize complementary evaluation regimes rather than a controlled architectural comparison.

4 Empirical Results and Analysis

4.1 Experimental Setup

Evaluation settings. We evaluate existing models under three settings. First, in closed-set detection, models are trained on the RegionDet training set and evaluated on the test set using the same eight-category label space. Second, in zero-shot/open-vocabulary detection, pretrained vision-language detectors are evaluated without training or fine-tuning on RegionDet. Third, we conduct

Table 4: Results of Closed-set Detection Methods on RegionDet (%).

Type	Methods	Backbone/Encoder	Construction	Crossing	Damage	Queuing	Talking	Vendor	Waiting	Walking	AP	AP_{50}	AP_{75}	Params. (M)
Dense one-stage	ATSS	ResNet-50	28.5	18.5	6.0	13.9	19.2	19.4	9.0	17.7	16.5	35.7	13.0	32.1
	YOLOv13-L	YOLOv13	31.8	54.0	6.61	30.7	33.7	24.8	15.6	40.2	29.7	48.5	29.8	27.6
	YOLO26-L	YOLO26	33.3	53.2	8.2	28.3	33.7	25.1	18.0	44.6	30.5	49.2	31.4	24.8
Two-stage	Faster R-CNN	ResNet-50	25.5	13.1	5.7	16.5	17.7	18.2	8.8	20.0	15.7	38.1	10.3	41.4
Point/center anchor-free	RepPoints	ResNet-50	28.8	17.6	6.8	15.9	19.4	19.9	9.6	17.5	16.9	36.7	13.3	36.8
	CenterNet++	ResNet-50	34.6	35.8	6.8	22.5	21.8	17.8	13.4	26.6	22.4	40.8	22.3	40.8
Diffusion based	DiffusionDet	ResNet-50	29.7	40.8	6.1	23.0	21.5	18.6	17.4	30.6	23.5	44.5	22.3	110.6
	DiffuDETR	ResNet-50	32.1	37.5	6.5	23.2	20.1	18.6	16.8	27.5	22.8	39.9	22.9	59.1
Query-based Transformer	DEIMv2-L	DINOv3-S16	45.5	67.7	15.5	31.4	37.0	35.0	25.3	38.8	37.0	57.1	38.7	32.0
	RT-DETRv4-L	DINOv3-B16	38.6	62.7	11.3	32.2	35.5	29.0	22.5	32.8	33.1	52.7	34.7	31.0
	RF-DETR-L	DINOv2-S16	37.5	60.0	13.3	32.8	43.7	34.5	24.5	37.8	35.5	53.6	37.4	33.6
	DINO	ResNet-50	29.7	36.6	5.7	17.8	15.7	15.0	10.7	23.2	19.3	34.0	19.1	47.6
	Co-DETR	ResNet-50	35.4	46.8	7.8	25.5	25.0	22.3	17.3	31.9	26.5	46.2	26.2	64.5
	MI-DETR	ResNet-50	33.6	47.0	6.9	23.9	22.2	20.1	15.2	31.9	25.1	44.4	25.0	77.3

Table 5: Zero-shot/open-vocabulary results on RegionDet with different prompts (%).

Prompt Type	Methods	Backbone/Encoder	Construction	Crossing	Damage	Queuing	Talking	Vendor	Waiting	Walking	AP	AP_{50}	AP_{75}	Params. (M)
Class name	MMGroundingDINO	Swin-T	3.4	0.2	1.2	0.0	0.3	0.6	0.3	0.2	0.8	1.9	0.6	173
		Swin-B	1.1	0.1	0.6	0.0	0.3	0.4	0.2	0.2	0.3	1.1	0.2	233
		Swin-L	1.4	0.3	0.1	0.0	0.5	3.6	0.3	0.3	0.8	2.1	0.6	343
	YOLO-World	YOLOv8-S	2.3	0.1	0.0	0.0	0.4	0.1	0.1	0.0	0.4	1.1	0.2	76
		YOLOv8-M	3.6	0.5	0.0	0.1	0.5	0.1	0.8	0.1	0.7	1.9	0.4	92
		YOLOv8-L	3.5	0.0	0.0	0.0	1.0	0.1	0.2	0.0	0.6	1.8	0.3	110
	YOLOE	YOLOv8-S	4.2	0.1	0.3	0.7	0.2	0.3	0.3	0.2	0.7	2.0	0.4	78
		YOLOv8-M	4.6	0.4	0.3	0.5	0.2	0.4	0.2	0.3	0.9	2.5	0.5	95
		YOLOv8-L	5.1	0.4	0.4	0.5	0.3	0.5	0.4	0.1	1.0	2.8	0.5	115
		YOLOv11-S	4.0	0.4	0.2	0.2	0.3	0.3	0.2	0.3	0.7	2.2	0.4	76
		YOLOv11-M	4.9	0.2	0.2	0.5	0.1	0.3	0.4	0.2	0.8	2.4	0.4	91
		YOLOv11-L	4.5	0.1	0.9	0.6	0.1	0.4	0.6	0.0	0.9	2.5	0.5	96
	LLMDet	Swin-T	0.7	0.1	0.8	0.0	0.4	0.2	0.1	0.1	0.3	0.9	0.2	744
		Swin-B	1.0	0.2	0.8	0.0	0.3	0.7	0.1	0.1	0.4	1.1	0.2	804
		Swin-L	0.8	0.2	0.4	0.0	0.4	0.2	0.1	0.1	0.3	0.9	0.2	879
Short phrase	MMGroundingDINO	Swin-T	1.8	0.1	0.3	2.1	1.4	0.1	0.1	0.1	0.7	1.5	0.6	173
		Swin-B	5.5	0.2	0.3	2.3	1.2	2.1	0.1	0.2	1.5	3.4	1.2	233
		Swin-L	3.4	0.1	0.3	1.4	2.7	0.2	0.1	0.1	1.0	2.1	0.9	343
	YOLO-World	YOLOv8-S	3.7	0.1	0.0	0.0	0.2	0.4	0.9	0.2	0.7	1.9	0.4	76
		YOLOv8-M	6.9	1.1	0.0	2.6	0.3	0.6	2.0	0.1	1.7	4.2	1.2	92
		YOLOv8-L	4.7	0.0	0.0	0.1	0.7	1.0	2.3	0.1	1.1	2.7	0.7	110
	YOLOE	YOLOv8-S	5.8	0.2	0.7	0.0	0.4	0.3	0.1	0.1	0.9	2.5	0.6	78
		YOLOv8-M	5.7	0.3	0.4	0.0	0.1	0.3	0.1	0.1	0.9	2.6	0.5	95
		YOLOv8-L	5.9	0.2	0.3	0.4	0.2	0.5	0.2	0.1	1.0	2.8	0.5	115
		YOLOv11-S	5.3	0.2	0.2	0.0	0.2	0.2	0.1	0.0	0.8	2.3	0.4	76
		YOLOv11-M	6.1	0.1	0.2	0.0	0.2	0.4	0.1	0.1	0.9	2.7	0.5	91
		YOLOv11-L	5.9	0.1	0.9	0.1	0.1	0.5	0.4	0.0	1.0	2.8	0.6	96
	LLMDet	Swin-T	1.5	0.1	0.4	3.1	0.4	0.2	0.1	0.2	0.7	1.6	0.6	744
		Swin-B	1.0	0.2	0.3	3.6	0.8	0.2	0.1	0.1	0.8	1.6	0.7	804
		Swin-L	2.0	0.3	0.2	3.6	1.4	0.2	0.2	0.2	1.0	2.0	0.9	879
Full definition	MMGroundingDINO	Swin-T	1.2	1.0	0.5	2.2	1.2	0.8	0.1	0.2	0.9	1.8	0.8	173
		Swin-B	3.0	0.5	0.6	3.6	1.4	3.2	0.1	0.1	1.6	3.5	1.3	233
		Swin-L	6.6	1.1	0.5	4.4	1.3	1.0	0.1	0.1	1.9	4.0	1.6	343
	YOLO-World	YOLOv8-S	1.4	1.9	0.4	0.1	0.3	0.7	0.4	0.2	0.7	1.6	0.4	76
		YOLOv8-M	1.5	3.1	0.2	0.2	0.5	3.3	1.0	0.1	1.2	2.7	0.9	92
		YOLOv8-L	2.7	1.2	0.1	0.3	0.7	2.4	1.4	0.1	1.1	2.6	0.8	110
	YOLOE	YOLOv8-S	0.4	6.0	0.2	0.0	0.2	1.5	0.5	0.1	1.1	2.0	1.1	78
		YOLOv8-M	0.6	3.5	0.1	0.0	0.2	1.2	0.1	0.1	0.7	1.4	0.6	95
		YOLOv8-L	0.7	4.9	0.2	0.0	0.4	1.8	0.3	0.1	1.1	2.1	0.9	115
		YOLOv11-S	0.6	1.7	0.1	0.0	0.1	1.1	0.1	0.1	0.5	1.1	0.4	76
		YOLOv11-M	0.4	3.7	0.1	0.0	0.2	1.5	0.2	0.1	0.8	1.5	0.7	91
		YOLOv11-L	0.6	1.7	0.2	0.0	0.1	2.3	0.1	0.1	0.6	1.5	0.5	96
	LLMDet	Swin-T	3.3	0.4	0.6	5.6	2.2	1.0	1.1	0.3	1.8	3.5	1.6	744
		Swin-B	6.2	0.2	0.7	2.1	1.7	2.2	0.1	0.2	1.7	3.7	1.3	804
		Swin-L	7.0	4.4	0.6	2.1	1.4	1.2	0.1	0.3	2.1	4.2	1.8	879

a category-conditioned localization experiment with general-purpose VLMs to examine whether broader visual-semantic reasoning benefits region localization when the target category is provided.

Prompt settings. For zero-shot/open-vocabulary detectors, we consider three text-query settings: (1) *class name*, containing only the category name; (2) *short phrase*, briefly describing the target as a region; and (3) *full definition*, specifying the relevant entities, relations, activities, and contextual cues. The semantic content of each prompt is kept consistent across models, while model-specific input formatting follows the corresponding official implementation. These settings vary the level of textual detail.

Evaluation metrics. We report AP, AP_{50}, AP_{75}, and per-category AP following the standard COCO evaluation protocol. Overall and per-category results are computed from the raw evaluator outputs and rounded independently. Model parameters are reported in millions when available.

Implementation details. We use the official implementations and pretrained checkpoints of all evaluated methods. Closed-set models are trained using their recommended optimization settings, while zero-shot/open-vocabulary models use their official inference pipelines. Both settings use the same RegionDet test split and eight-category label space, with predictions evaluated using the same COCO-style detection metrics. Because the evaluated models differ in backbone, pretraining, and optimization recipe, the benchmark provides a broad empirical comparison rather than a controlled architectural ablation.

Evaluated methods. The closed-set benchmark covers dense one-stage, two-stage, point/center-based, diffusion-based, and query-based Transformer detectors, as listed in Table 4. The zero-shot/open-vocabulary benchmark includes MMGroundingDINO, YOLO-World, YOLOE, and LLMDet. We additionally evaluate Qwen2.5-VL-7B and InternVL3.5-8B-HF in the category-conditioned VLM setting.

4.2 Closed-set Detection Results

Table 4 presents the closed-set results. Existing detectors can learn region-level patterns under supervision, although substantial differences remain across models and categories. DEIMv2-L achieves the highest overall AP of 37.0, AP50 of 57.1, and AP75 of 38.7. RF-DETR-L achieves the second-highest overall AP of 35.5, with an AP50 of 53.6 and an AP75 of 37.4.

Model-level differences. The strongest results are obtained by recent query-based Transformer detectors equipped with pretrained visual encoders. However, this advantage cannot be attributed solely to the detection paradigm because their backbones and training recipes differ. Among models using ResNet-50, for example, Co-DETR achieves 26.5 AP, compared with 23.5 for DiffusionDet, 22.4 for CenterNet++, and 19.3 for DINO. These results suggest that architecture, visual representation, and optimization jointly affect region-level detection performance.

Category-level difficulty. Performance varies considerably across region categories. *Crossing* is comparatively easier for strong supervised detectors, with DEIMv2-L and RF-DETR-L reaching 67.7 and 60.0 AP, respectively. In contrast, *Damage* remains the most difficult category, with the best result reaching only 15.5 AP. *Waiting* is also challenging, with a best AP of 25.3. These results indicate that weak visual boundaries, diverse appearances, and state-dependent semantics remain difficult even under full supervision.

Localization sensitivity. The gap between AP_{50} and AP_{75} indicates that approximate localization is easier than determining precise region extents. For example, DiffusionDet obtains 44.5 AP_{50} but 22.3 AP_{75}, while DEIMv2-L obtains 57.1 AP_{50} and 38.7 AP_{75}. Thus, detectors can identify relevant areas but remain sensitive to the contextual extent of target regions.

4.3 Zero-shot/Open-vocabulary Detection

Table 5 reports the complete zero-shot/open-vocabulary results under three prompt settings, including class names, short phrases, and full category definitions.

Region-oriented prompts improve performance but do not close the gap. Averaged over the 15 model variants, AP increases from 0.6 with class names to 1.0 with short phrases and 1.2 with full definitions. The best result improves from 1.0 AP under the class-name prompt to 2.1 AP under the full-definition prompt. This confirms that category names are insufficient descriptions of region targets. Nevertheless, even the best prompt-based result remains far below the best supervised result of 37.0 AP, highlighting the limited effectiveness of the tested region-oriented prompts.

Prompt effects are model dependent. Full definitions substantially benefit models with stronger language-conditioning mechanisms. For example, LLMDet with Swin-L improves from 0.3 AP using the class name to 2.1 AP using the full definition, while MMGroundingDINO with Swin-L improves from 0.8 to 1.9 AP. In contrast, YOLOE shows no consistent improvement from longer descriptions. Thus, additional linguistic context is useful only when the model can effectively associate relational and contextual descriptions with spatial regions.

Different prompts favor different categories. Full definitions yield noticeable gains for several semantically structured categories, reaching 7.0 AP on *Construction*, 6.0 AP on *Crossing*, and 5.6 AP on *Queuing* for the best corresponding models. However, *Damage* and *Walking* remain extremely difficult, with their best results across all prompt settings reaching only 1.2 and 0.3 AP, respectively. These results reveal that descriptive prompts can improve semantic disambiguation but remain insufficient for learning the spatial extent of weak-boundary or activity-defined regions.

Table 6: Category-conditioned localization results of general-purpose VLMs on RegionDet (%).

Method	Construction	Crossing	Damage	Queuing	Talking	Vendor	Waiting	Walking	AP	AP_{50}	AP_{75}
Qwen2.5-VL-7B	10.7	10.7	0.8	24.4	0.6	0.2	1.9	0.9	6.3	13.8	4.6
InternVL3.5-8B-HF	1.2	0.6	0.3	2.3	0.0	0.0	0.3	0.0	0.6	2.2	0.2

4.4 Category-conditioned VLM Localization

We evaluate whether general-purpose VLMs provide complementary region-level reasoning capabilities. For each image-category pair, the target category and its description are provided to the model, which is asked to return all corresponding bounding boxes. Generated coordinates are parsed using a deterministic procedure and mapped to the image resolution. No RegionDet training examples or parameter updates are used. Because the target category is given in advance, this experiment is an oracle-category localization diagnostic and is not directly comparable to multi-class detection.

Table 6 shows that general-purpose VLMs exhibit stronger but highly model-dependent localization capabilities. Qwen2.5-VL-7B reaches 6.27 AP and 13.85 AP_{50}, outperforming the evaluated open-vocabulary detectors under their multi-class zero-shot setting. However, this comparison should be interpreted cautiously because the VLM receives the target category in advance.

Qwen2.5-VL performs particularly well on *Queuing*, reaching 24.44 AP, and obtains approximately 10.7 AP on both *Construction* and *Crossing*. This suggests that its visual-semantic reasoning can help identify regions with distinctive group arrangements or scene structures. However, its performance remains below 2 AP on *Damage*, *Talking*, *Vendor*, *Waiting*, and *Walking*, indicating persistent difficulty in converting semantic understanding into complete and precise region extents.

InternVL3.5 obtains only 0.60 AP despite receiving the same category-level information. The large difference between the two VLMs shows that general visual-semantic capability does not automatically provide reliable coordinate-level grounding. Region localization remains strongly dependent on the model's spatial representation, coordinate generation, and ability to aggregate multiple visual elements into a coherent region.

4.5 Qualitative Analysis

Closed-set detectors. Fig. 3 shows that supervised detectors can identify relevant region areas, consistent with the quantitative results. For *Construction* and *Vendor*, predictions often focus on salient components such as workers, equipment, goods, or pedestrians while missing parts of the functional region. For *Damage*, boxes are frequently shifted or incomplete. For *Queuing*, *Talking*, and *Waiting*, detectors sometimes localize individual people or subgroups rather than the complete relation-defined region. These examples illustrate the difficulty of aggregating multiple visual elements and determining context-dependent spatial extents.



Figure 3: Qualitative results of representative closed-set detectors. Although supervised detectors identify relevant areas, their predictions are often incomplete or spatially imprecise.

Zero-shot/open-vocabulary detectors. Fig. 4 presents representative qualitative results of LLMDet. Although the model identifies relevant visual elements such as people, workers, goods, and local structures, its predictions often focus on individual objects or local components rather than complete region targets. This behavior is particularly evident for activity-, relation-, and function-defined



Figure 4: Representative zero-shot results of LLMDet. Predictions frequently correspond to individual objects or local components rather than complete region targets.

categories whose spatial extents depend on multiple entities and contextual cues. These observations are consistent with the quantitative results and indicate an object-oriented localization tendency in current vision-language detectors.

4.6 Key Findings

First, region-level patterns are learnable under supervision, but precise localization remains difficult. Recent closed-set detectors achieve moderate performance, while the persistent AP_{50}–AP_{75} gaps and low *Damage* performance reveal sensitivity to context-dependent region extents.

Second, performance cannot be explained by detection paradigm alone. The strongest models combine effective architectures, pretrained visual representations, and optimized training recipes. RegionDet should therefore be interpreted as a broad benchmark rather than a controlled comparison of architectural families.

Third, region-oriented prompts help but remain insufficient. Short phrases and full definitions improve average zero-shot performance, yet the best result reaches only 2.1 AP. The remaining gap indicates limitations beyond prompt selection.

Fourth, general-purpose VLMs provide complementary but unstable localization capability. Qwen2.5-VL performs well on selected structured categories, particularly *Queuing*, whereas InternVL3.5 remains weak. This suggests that semantic reasoning can help, but reliable region detection also requires accurate spatial grounding and contextual aggregation.

5 Conclusion and Discussion

In this paper, we introduce **Region Detection**, a box-level detection setting that extends conventional object detection beyond discrete and well-bounded object instances. We construct **RegionDet**, a benchmark containing eight region categories, COCO-compatible bounding-box annotations, and standard evaluation protocols. RegionDet provides a testbed for localizing visual regions defined by states, relations, scene functions, activities, and contextual cues. Our experiments show that region-level patterns are learnable under supervision, but existing detectors remain limited by weak boundary cues, context-dependent spatial extents, and relation-defined semantics. Zero-shot/open-vocabulary detectors perform poorly even with region-oriented prompts, indicating that their limitations cannot be explained solely by prompt selection. Category-conditioned experiments with general-purpose VLMs show improved localization on selected structured categories, but the results remain highly model dependent and substantially below supervised detection performance. Together, these findings highlight the need for stronger contextual reasoning, relational modeling, and spatial grounding.

Limitations and Future Directions. RegionDet currently represents each target using a single bounding box, which may not fully capture the uncertainty and multiple plausible extents of context-defined regions. Its eight categories also primarily cover urban streets and public-space scenarios. Future work may extend RegionDet to broader domains and investigate multiple acceptable annotations, soft or uncertain boundaries, segmentation-based representations, and models designed explicitly for context- and relation-aware region localization. We hope RegionDet facilitates progress from object-instance localization toward richer region-level visual understanding.

AI USE STATEMENT

We used generative AI tools only to assist with language polishing, grammar correction, and improving the readability and organization of the manuscript. We did not use generative AI tools for task formulation, dataset construction, annotation, experimental design, model implementation, data processing, result analysis, or drawing scientific conclusions. All AI-assisted text was carefully reviewed, revised, and approved by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work introduces RegionDet, a benchmark for Region Detection. The dataset is constructed from images collected by the authors and is intended solely for academic research. RegionDet focuses on region-level visual categories such as construction, road damage, queuing, talking, vendor activity, waiting, walking, and crossing. We do not collect private user data, conduct human-subject experiments, or intentionally annotate personally identifiable information. Although some images may contain people in public scenes, the annotations focus only on region-level visual semantics rather than individual identities or personal attributes. The dataset will be released for research purposes with appropriate documentation and usage restrictions. Researchers are expected to use the dataset responsibly and avoid applications that may infringe privacy, enable surveillance abuse, or cause harm to individuals or communities.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the Region Detection task, the RegionDet dataset, category definitions, annotation protocol, data statistics, model training settings, evaluation metrics, and baseline configurations in the paper. RegionDet follows a COCO-style bounding-box annotation format and a standard detection evaluation protocol to facilitate reuse in existing detection frameworks. To support reproducibility and future research, we will make the processed annotation files, dataset splits, and evaluation scripts publicly available.

REFERENCES

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pp. 381–389. IEEE, 2018.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16901–16911, 2024.

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

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338, 2010.

Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14987–14997, 2025.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6047–6056, 2018.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5356–5364, 2019.

Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9404–9413, 2019.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35: 9287–9301, 2022a.

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*, pp. 10965–10975, 2022b.

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*, pp. 740–755, 2014.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. Simple open-vocabulary object detection with vision transformers. arxiv 2022. *arXiv preprint arXiv:2205.06230*, 2, 2022.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 891–898, 2014.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8430–8439, 2019.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anything. *arXiv preprint arXiv:2503.07465*, 2025.

Yupeng Zhang, Shuqi Zheng, Ruize Han, Yuzhong Feng, Junhui Hou, Linqi Song, Wei Feng, and Liang Wan. Rethinking the one-shot object detection: Cross-domain object search. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9573–9581, 2024.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.