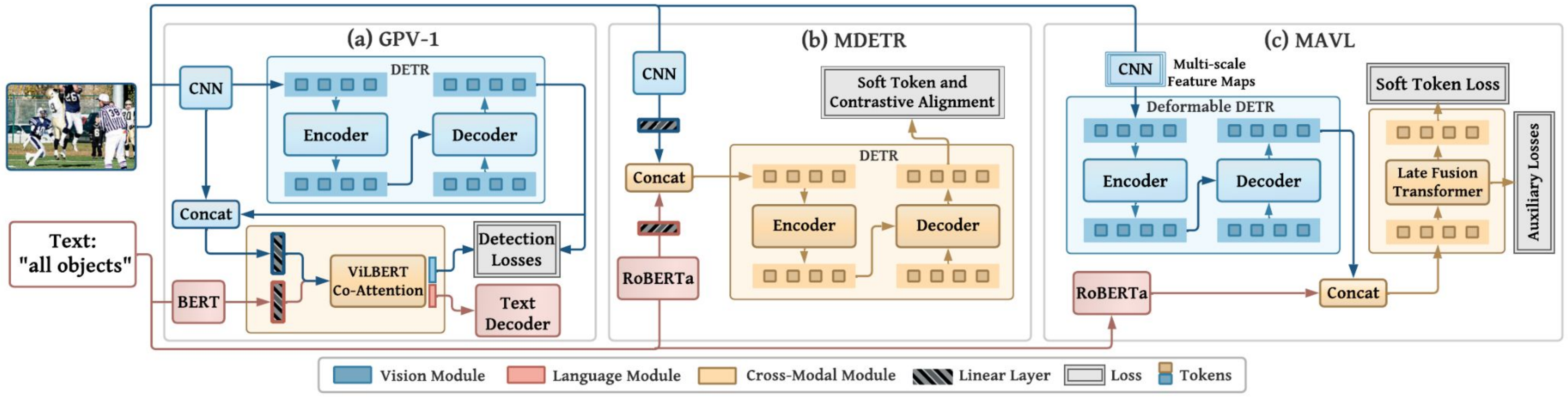
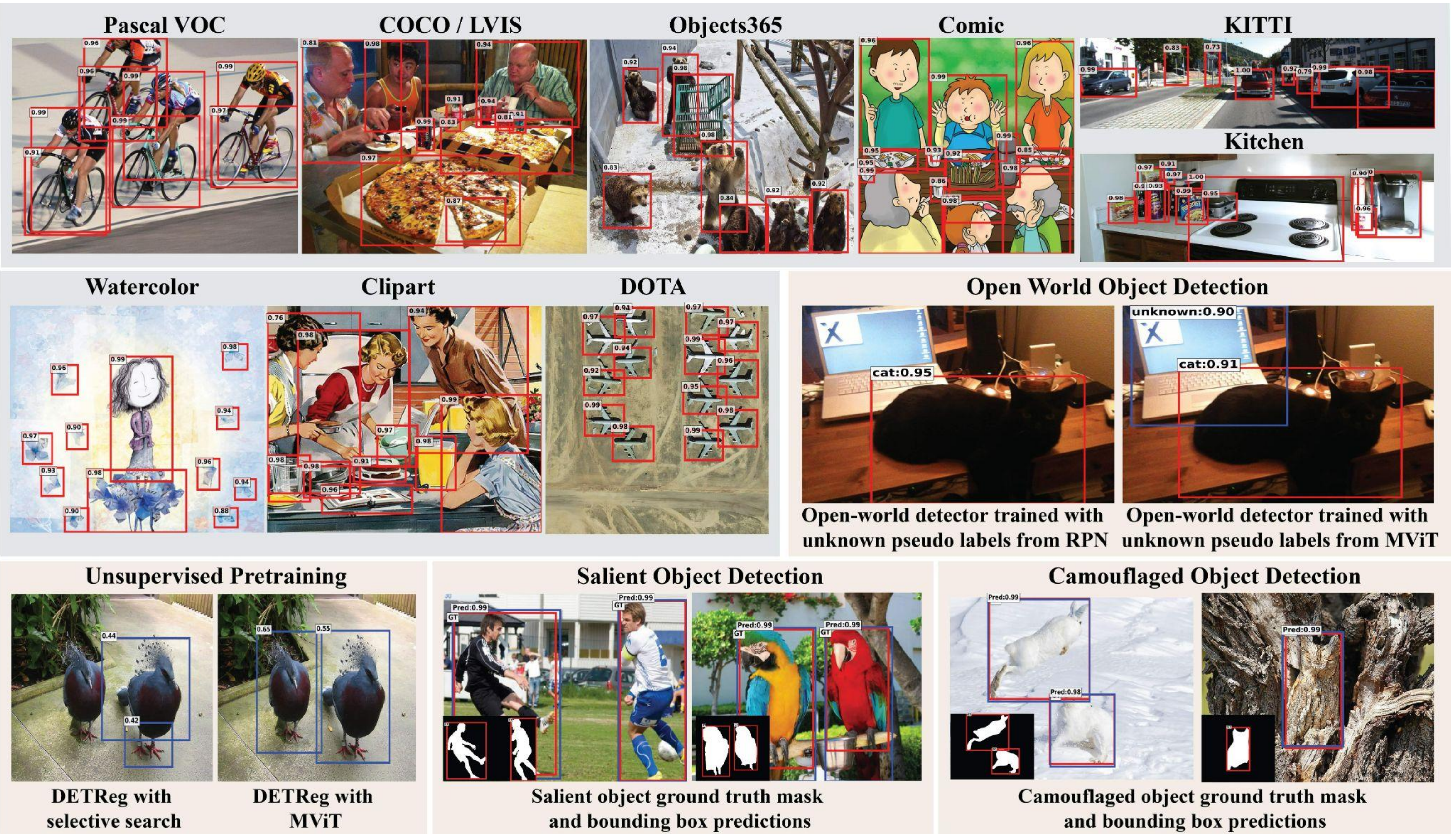
Class-agnostic Object Detection with Multi-modal Transformer

Highlights

- Multi-modal Vision Transformers (MViTs) excel at Class-agnostic OD (COD) across multiple domains.
- COD using human intuitive natural language text queries (e.g., “all objects”, “all entities”, etc.).
- Propose an efficient MViT model, Multiscale Attention ViT with Late fusion (MAVL), with state-of-the-art COD performance.
- Class-agnostic detectors (MViTs) can be applied to several downstream applications.
- In Open-world OD, unknown pseudo-labels generated using MViT improves novelty detection.
- In Salient and Camouflaged OD, task specific queries perform competitively against supervised models without any tuning.

Architecture overview of MViTs





Open-world detector trained with unknown pseudo labels from RPN

Open-world detector trained with unknown pseudo labels from MViT

DETReg with selective search

DETReg with MViT

Salient object ground truth mask and bounding box predictions

Camouflaged object ground truth mask and bounding box predictions

Results of Class-agnostic Object Detection

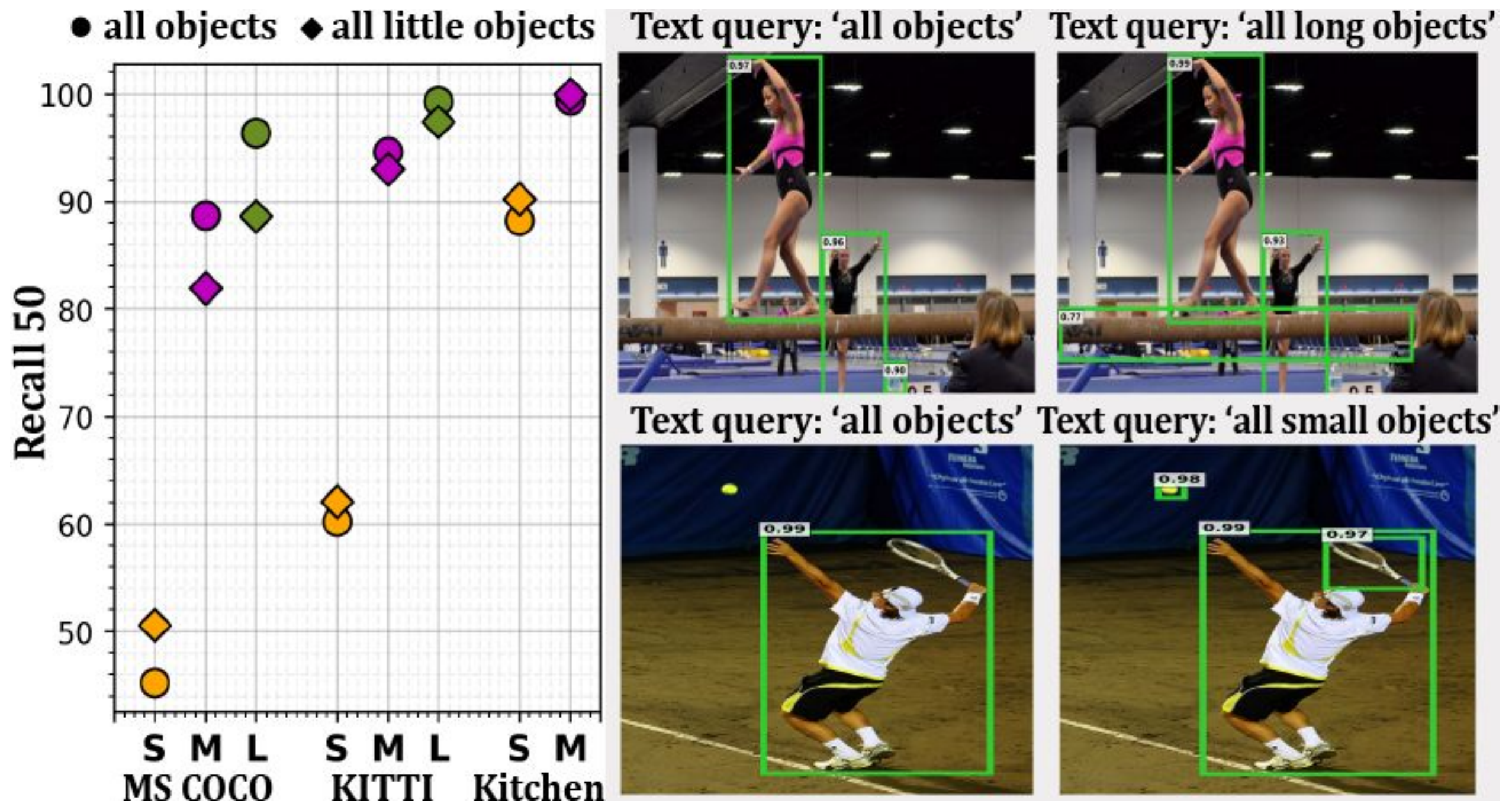
Dataset →	Pascal-VOC		COCO		KITTI		Objects365		LVIS	
Model ↓	AP50	R50	AP50	R50	AP50	R50	AP50	R50	AP50	R50
Edge Boxes	0.08	7.14	0.09	5.16	0.09	6.58	0.07	3.27	0.05	3.00
Selective Search	0.32	21.4	0.27	12.7	0.03	4.85	0.38	10.7	0.24	9.31
Deep Mask	5.92	40.4	2.16	19.2	1.33	15.5	1.31	14.5	0.51	8.17
Faster-RCNN	42.9	85.8	26.4	58.7	23.5	53.2	24.8	54.6	8.91	35.6
RetinaNet	43.2	86.6	24.6	59.1	30.4	57.6	24.3	54.8	8.57	35.7
Def-DETR	30.1	81.0	20.0	53.5	23.7	55.0	17.0	45.9	6.60	30.7
GPV-I	61.9	91.1	38.0	64.4	43.0	64.4	25.6	50.2	9.18	27.5
MDETR	66.0	90.1	40.7	62.2	46.7	**67.2**	30.4	54.0	10.7	32.8
MAVL (Ours)	**68.6**	**91.3**	**43.6**	**65.0**	**48.2**	63.5	**33.2**	**57.9**	**11.7**	**37.0**
	+25.4	+4.7	+19.0	+5.9	+17.8	+5.9	+8.4	+3.1	+2.8	+1.3

Class-agnostic OD results of MViTs in comparison with bottom-up approaches and uni-modal detectors trained to localize generic objects. In general, MViTs achieve state-of-the-art performance using intuitive text queries.

Dataset →	Kitchen		Clipart		Comic		Watercolor		DOTA†	
Model ↓	AP50	R50	AP50	R50	AP50	R50	AP50	R50	AP50	R50
RetinaNet	35.3	89.5	27.0	90.0	33.1	86.1	47.8	91.9	0.72	15.6
GPV-1	24.5	84.8	35.1	86.1	42.3	83.6	50.3	89.5	0.55	9.33
MDETR	38.4	**91.4**	44.9	90.7	55.8	**89.5**	63.6	94.3	1.94	21.8
MAVL (Ours)	**45.4**	91.0	**50.6**	**92.9**	**57.7**	89.2	**63.8**	**95.6**	**2.86**	**24.2**

Class-agnostic OD performance of MViTs in comparison with RetinaNet on several out-of-domain datasets.

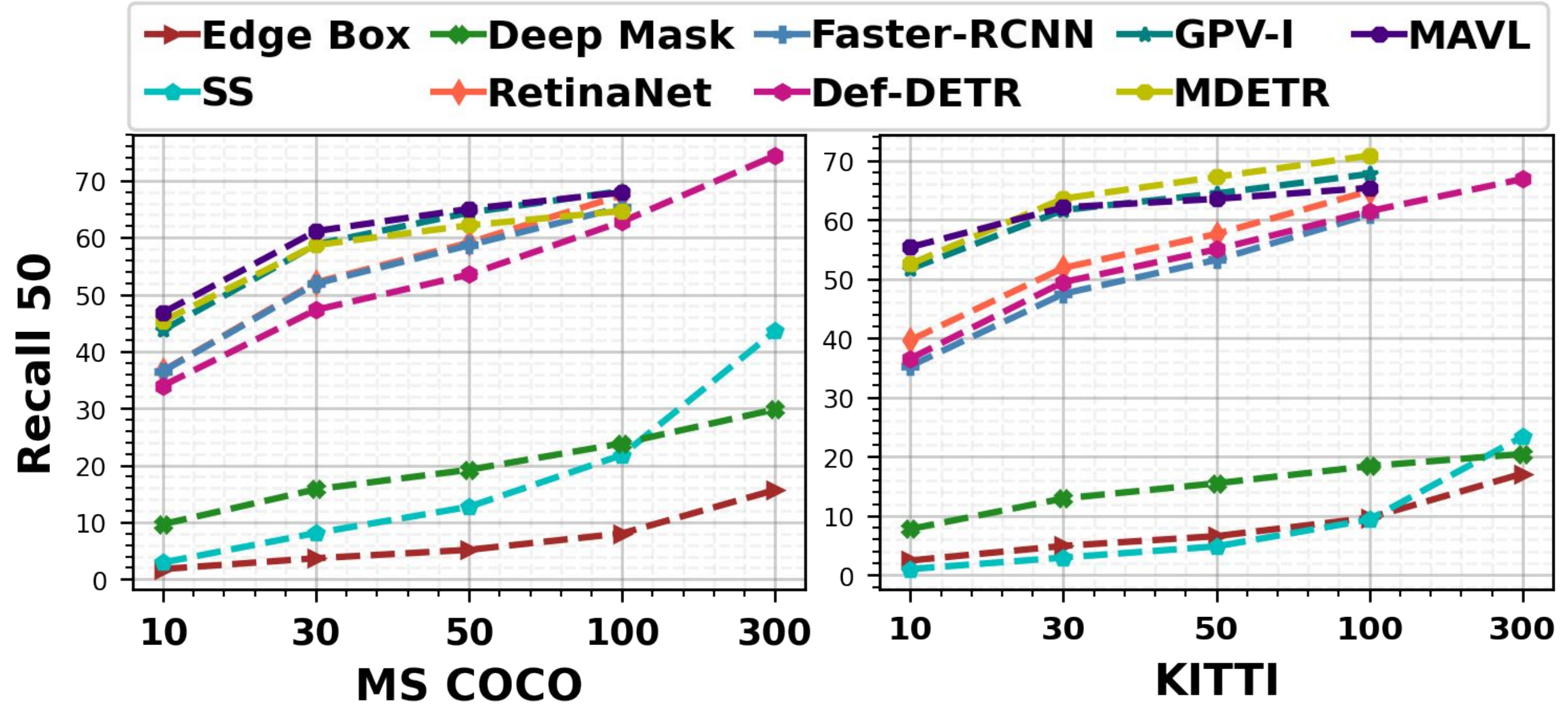
Some Use Cases of Using Different Intuitive Text Queries



Dataset →	Pascal-VOC		COCO		KITTI	
Text Query ↓	AP50	R50	AP50	R50	AP50	R50
all objects	51.3	85.5	33.3	58.4	40.2	64.0
all entities	65.2	88.4	34.6	54.6	41.9	59.5
all visible entities & objects	63.3	89.0	37.9	61.6	42.0	63.0
all obscure entities & objects	59.5	86.6	35.2	59.1	42.4	63.5
all small objects	40.0	83.9	28.9	58.9	40.4	65.2
combined detections (CD)	63.7	91.0	42.0	**65.0**	**48.2**	**63.5**
CD w/o ‘all small objects’	**68.6**	**91.3**	**43.6**	**65.0**	45.8	61.6

Combining MAVL detections from multiple human intuitive natural language queries captures varying aspects of objectness

Using Different Number of Proposals



Analysis on the Importance of Language Structure

Dataset→		Lang.	Pascal-VOC		COCO		KITTI	
Model ↓	Lexicon	Structure	AP50	R50	AP50	R50	AP50	R50
MDETR	✓	✓	63.9	88.0	38.1	58.5	42.5	60.9
MAVL	✓	✓	65.0	89.1	39.3	62.0	39.0	61.0
MDETR	×	✓	59.7	86.4	33.4	57.9	36.9	55.0
MAVL	×	✓	61.6	86.7	34.4	58.3	36.5	58.9
MAVL †	×	×	35.1	82.7	21.2	56.3	21.5	58.5

Effect of removing language branch from MViTs keeping the data loader structure intact. The performance is not affected largely as the language structure is still intact (boxes from caption are seen together). However, it degrades significantly when language structure is removed.

Applications

State-of-the-art Results on,

- Open-world Object Detection
- Pretraining for Class-aware Object Detection
- Salient Object Detection
- Camouflaged Object Detection
- Improving Two-stage Object Detection

جـــامـــعـــة مـحـــمـــد بـــن زايـــد
للـــذكـــاء الاصـــطـــنـــاعـــي
MOHAMED BIN ZAYED UNIVERSITY
OF ARTIFICIAL INTELLIGENCE

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022
October 23-27, 2022, Tel Aviv