Multi-modal Queried Object Detection in the Wild

Yifan Xu[1,3†*], **Mengdan Zhang**[2†], **Chaoyou Fu**[2],
Peixian Chen[2], **Xiaoshan Yang**[1,3], **Ke Li**[2], **Changsheng Xu**[1,3‡]
[1]MAIS, Institute of Automation, Chinese Academy of Sciences [2]Tencent Youtu Lab
[3]University of the Chinese Academy of Sciences
{yifan.xu, csxu}@nlpr.ia.ac.cn, davinazhang@tencent.com

Code: https://github.com/YifanXu74/MQ-Det

From language query to multi-modal query

Fish?

Fish?

Plane?

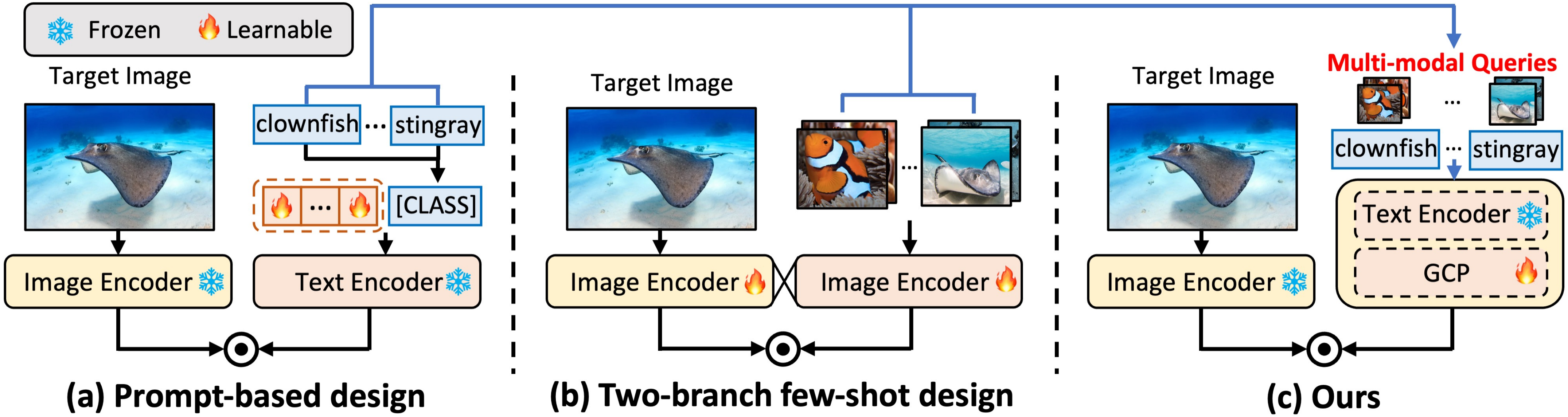
Plane?

Bat?

Bat?

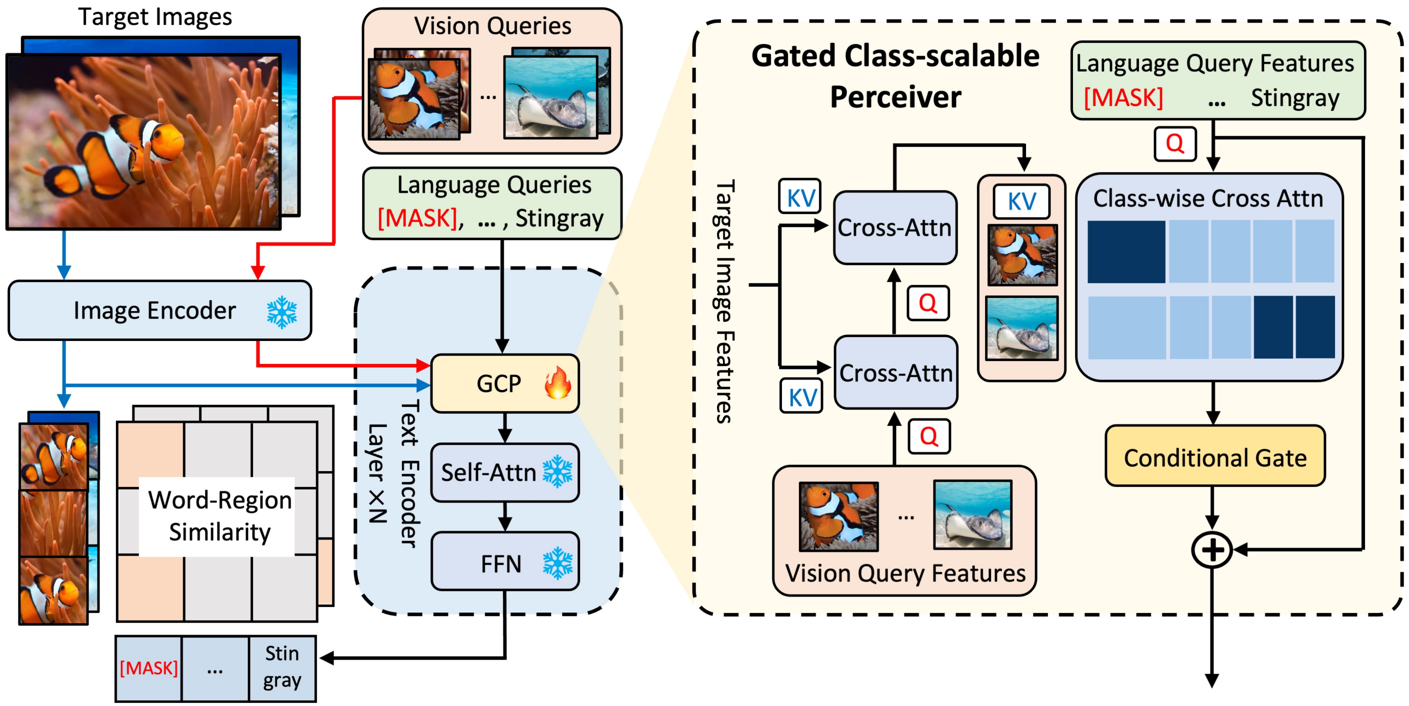
- **Multi-modal queried object detection**
  - One can detect customized objects through textual descriptions, visual exemplars, or both.

- **Language-queried object detector** (current open-world detectors):
  - ✓ Pros: high information density and **strong generalization capability**
  - ✗ Cons: **insufficient granularity** and ambiguous queries
- **Vision-queried object detector** (few-shot detectors):
  - ✓ Pros: **rich description granularity**
  - ✗ Cons: redundant information and **low generalization**
- **Multi-modal queried object detector (ours)**
  - ✓ **Open-set generalization**
  - ✓ **Rich description granularity**



Contributions

- **The first multi-modal queried open-world object detector**. We take the first step on multi-modal queried object detection.
- **Wide applicability.** We design a plug-and-play Gated Class-scalable Perceiver (GCP) structure and a vision conditioned masked language prediction strategy to enable multi-modal queries on most language-queried detectors.
- **High performance.** The proposed MQ-Det significantly boosts open-world detection in both finetuning-free and few-shot finetuning scenarios. For example, +7.8 AP over previous SOTA on finetuning-free LVIS.



- Gated Class-scalable Perceiver (GCP)
  - Language-vision fusion

$$\bar{\mathbf{v}}_i = \text{X-MHA}(\mathbf{v}_i, I), \quad \hat{v}_i = \text{X-MHA}(t_i, \bar{\mathbf{v}}_i),$$

$$\hat{t}_i = t_i + \sigma(gate(\hat{v}_i)) \cdot \hat{v}_i$$

  - To bridge class-wise visual cues and textual cues in each high-level stage of the text encoder of the detector.

- Vision conditioned masked language prediction

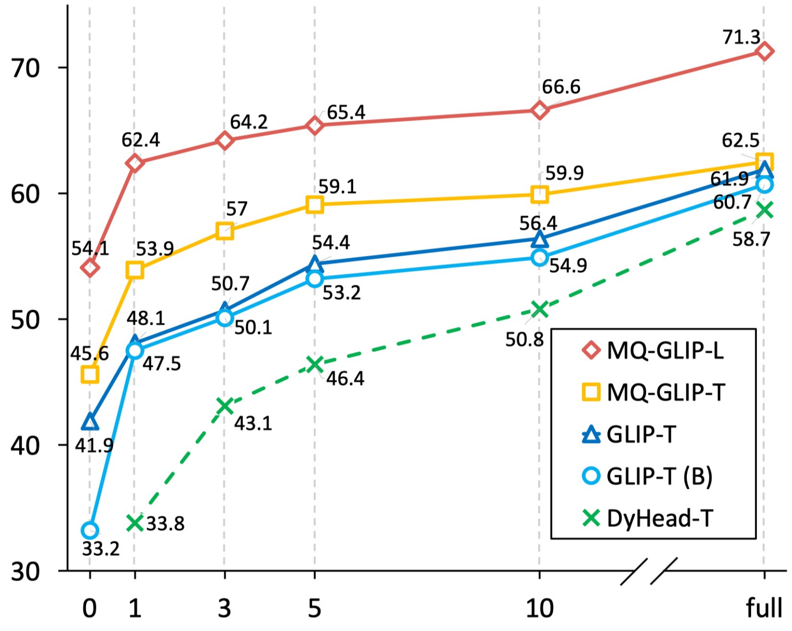
$$\mathcal{T} = \{t_1, t_2, \ldots, [\text{MASK}], \ldots, t_{|C|}\}$$

  - To ensure sufficient visual intervention in the modulating stage.

Finetuning-free LVIS

Model	Backbone	Pre-Train Data	Data Size	Training Time (V100 days)	#Vision Query	MiniVal (%) AP	MiniVal (%) AP_r	MiniVal (%) AP_c	MiniVal (%) AP_f	Val v1.0 (%) AP	Val v1.0 (%) AP_r	Val v1.0 (%) AP_c	Val v1.0 (%) AP_f
MDETR [20]*	RN101	GoldG,RefC	0.9M	400	0	24.2	20.9	24.9	24.3	22.5	7.4	22.7	25.0
Mask R-CNN [17]*	RN101	-	-	-	0	33.3	26.3	34.0	33.9	-	-	-	-
Supervised-RFS [13]*	RN50	-	-	-	0	-	-	-	-	25.4	12.3	24.3	32.4
GLIP-T (B) [25]	Swin-T	O365	0.66M	300	0	17.8	13.5	12.8	22.2	11.3	4.2	7.6	18.6
GLIP-T [25]	Swin-T	O365,GoldG,CC4M	5.5M	480	0	26.0	20.8	21.4	31.0	17.2	10.1	12.5	25.5
GLIPv2-T [48]	Swin-T	O365,GoldG,CC4M	5.5M	-	0	29.0	-	-	-	-	-	-	-
GroundingDINO-T [27]	Swin-T	O365,GoldG,Cap4M	5.5M	-	0	25.7	15.2	21.9	30.9	-	-	-	-
GLIP-L [25]	Swin-L	FourODs,GoldG,Cap24M	27.5M	600	0	37.3	28.2	34.3	41.5	26.9	17.1	23.3	35.4
GroundingDINO-L [27]	Swin-L	O365,OI,GoldG,Cap4M,COCO,RefC	15.8M	-	0	33.9	22.2	30.7	38.8	-	-	-	-
MQ-GLIP-T-Img	Swin-T	O365†	0.66M	10	5	17.6	12.0	14.5	21.2	12.4	8.9	9.2	18.3
MQ-GLIP-T-Txt	Swin-T	O365†	0.66M	10	0	26.0	20.8	21.4	31.0	17.2	10.1	12.5	25.5
MQ-GroundingDINO-T	Swin-T	O365†	0.66M	10	5	30.2	21.7	26.2	35.2	22.1	12.9	17.4	31.4
MQ-GLIP-T	Swin-T	O365†	0.66M	10	5	30.4	21.0	27.5	34.6	22.6	15.4	18.4	30.4
MQ-GLIP-L	Swin-L	O365†	0.66M	22	5	**43.4**	**34.5**	**41.2**	**46.9**	**34.7**	**26.9**	**32.0**	**41.3**

Compare with GLIP



Few-shot ODinW

Model	Language Query	Vision Query	Backbone	Pre-train Data	Data Size	ODinW-35 AP_{avg}	ODinW-13 AP_{avg}
Finetuning-free Setting							
MDETR [20]	✓	✗	ENB5 [38]	GoldG,RefC	0.9M	10.7	25.1
OWL-ViT [30]	✓	✓	ViT L/14(CLIP)	O365, VG	0.8M	18.8	40.9
GLIP-T [25]	✓	✗	Swin-T	O365,GoldG,Cap4M	5.5M	18.7	41.9
GLIP-L [25]	✓	✗	Swin-L	FourODs,GoldG,Cap24M	27.5M	22.6	51.0
OmDet [50]	✓	✗	ConvNeXt-B	COCO,O365,LVIS,PhraseCut	1.8M	16.0	43.6
GLIPv2-T [48]	✓	✗	Swin-T	O365,GoldG,Cap4M	5.5M	22.3	50.7
DetCLIP [42]	✓	✗	Swin-T	O365,GoldG,YFCC1M	2.4M	-	43.3
GroundingDINO-T [27]	✓	✗	Swin-T	O365,GoldG,Cap4M	5.5M	21.7	49.8
MQ-GroundingDINO-T	✓	✓	Swin-T	O365†	0.66M	22.5	50.9
MQ-GLIP-T	✓	✓	Swin-T	O365†	0.66M	20.8	45.6
MQ-GLIP-L	✓	✓	Swin-L	O365†	0.66M	**23.9**	**54.1**
Few-Shot Setting							
DyHead-T [6]	✗	✗	Swin-T	O365	0.66M	37.5	43.1
GLIP-T [25]	✓	✗	Swin-T	O365,GoldG,Cap4M	5.5M	38.9	50.7
DINO-Swin-T [47]	✗	✗	Swin-T	O365	0.66M	41.2	49.0
OmDet [50]	✓	✗	ConvNeXt-B	COCO,O365,LVIS,PhraseCut	1.8M	42.4	48.5
MQ-GLIP-T	✓	✓	Swin-T	O365†	0.66M	**43.0**	**57.0**
Full-Shot Setting							
GLIP-T [25]	✓	✗	Swin-T	O365,GoldG,Cap4M	5.5M	62.6	61.9
DyHead-T [6]	✗	✗	Swin-T	O365	0.66M	63.2	58.7
DINO-Swin-T [47]	✗	✗	Swin-T	O365	0.66M	66.7	-
OmDet [50]	✓	✗	ConvNeXt-B	COCO,O365,LVIS,PhraseCut	1.8M	67.1	65.3
DINO-Swin-L [47]	✗	✗	Swin-L	O365	0.66M	68.8	67.3
MQ-GLIP-T	✓	✓	Swin-T	O365†	0.66M	64.8	62.5

中国科学院大学
University of Chinese Academy of Sciences

Thank you.