Appendix

We provide an overview of the Appendix below.

Transfer to downstream (Sec. A). We elaborate on the additional details to transfer our MQ-Det to downstream tasks, including finetuning-free, few-shot, and full-shot settings. The introduction is organized as followed.

- Different ways to acquire the vision queries in Sec. A.1. This is an elaborative description of Sec. 3.1.2 in the main text.
- A customized tuning approach of MQ-Det in Sec. A.3, which achieves similar performance as full-model tuning while only fine-tuning very few parameters.

Experiments (Sec. B). We provide additional training and evaluation details, including:

- Detailed results on 35 downstream tasks in ODinW are provided in Tab. III and Tab. VII, described in Sec. B.1.
- An explicit evaluation on an open-vocabulary detection setting is conducted to further investigate the generalization of our multi-modal queries, as presented in Sec. B.2.
- Some visualization results of MQ-Det in Sec. B.1 and Fig. I.
- The training hyper-parameters are illustrated in Sec B.3 and Tab. VI.

Further discussion (Sec. C). We provide a comprehensive discussion on our work, including limitations and broader impacts in Sec C.

A Transfer to downstream

A.1 Different ways to acquire vision queries

During finetuning-free evaluation, we extract 5 instances as vision queries for each category from the downstream training set without any finetuning. We also provide two alternative strategies and observe similar performance, *i.e.*, retrieval and test-time online update, where the former obtains vision queries from heterogeneous external data like ImageNet [1], and the latter dynamically stores high-confidence instances as vision queries during evaluation. These two additional approaches are proposed to simulate realistic scenarios:

- Retrieval (user-provided exemplars): a small number of exemplars are provided by the users without any fine-tuning. We retrieve 5 exemplars as vision queries for each category from ImageNet-21K [1] to simulate the user-provided exemplars. These samples are heterogeneous from the downstream test data, *e.g.*, domains. The results are provided in MQ-GLIP-T-Retrival of Tab. I.
- Online updating: the model dynamically stores high-confidence instances as vision queries during evaluation. No vision queries are provided at the initial stage of evaluation. The results are illustrated in MQ-GLIP-T-Online of Tab. I. We provide detailed description in Sec. A.2.

The results are illustrated in Tab. I. We select 3 downstream datasets from ODinW [2] to verify the effectiveness of each approach. Generally, all three approaches to acquire vision queries demonstrate similar performance. We observe that vision queries from online updating hold relatively lower quality, thus leading to slight performance drop, since no manual annotations are provided. Meanwhile, exemplars retrieved from ImageNet are object-centric and contain little noise (*e.g.*, other irrelevant objects), thus improving the performance.

A.2 Test-time online update

Our test-time online update strategy is conducted via the following steps: 1) only utilize language queries to conduct detection at the initial stage of evaluation. 2) Store detected instances with high confidence as the vision queries of corresponding categories. 3) Use both language queries and stored

vision queries for evaluation and seek for more vision queries. Tab. I verifies the effectiveness of our approach.

A.3 Partial tuning rivals full-model tuning

Since the GCP modules are interleaved into the frozen detector, we provide a customized fine-tuning strategy, partial tuning, namely, only tuning the newly added GCP modules and freezing all other parameters. The results in Tab. II indicate that our partial tuning strategy achieves comparable performance with traditional full-model tuning (*i.e.*, only -0.4 % AP on ODinW-13). Partial tuning accounts for much fewer learnable parameters, thus friendly to training time and memory costs. The experimental results in the main text are all based on the partial tuning.

Table I: Different ways to acquire vision query. We report the finetuning-free performance. All models should be compared with the MQ-GLIP-T model at the top of the table.

Model	AerialDrone	Aquarium	Rabbits
GLIP-T	12.5	18.4	70.2
MQ-GLIP-T	15.8	23.5	75.4
MQ-GLIP-T-Online	15.5	23.2	74.8
MQ-GLIP-T-Retrieval	16.0	23.6	75.1

Table II: Different fine-tuning strategies of MQ-Det under the 5-shot setting. The implementation used in our evaluation is highlighted in color.

Strategy	ODinW-13 (%)	
	AP_{avg}	AP_{mid}
Partial tuning	59.1	62.4
Full-model tuning	59.5	64.5

B Experiments

B.1 All results

We report the per-dataset performance under various settings in ODinW-13 and ODinW-35, shown in Tab. III and Tab. VII, respectively. We also provide some visualized results in Fig. I.

Table III: Per-dataset AP performance (%) on ODinW-13. We report results on 0, 1, 3, 5, 10-shot detection. MQ-GD-T denotes MQ-GroundingDINO-T. Specifically, the "zero-shot" here actuall stands for the finetuning-free setting with 5 vision queries.

Dataset	MQ-GLIP-T					MQ-GLIP-L					MQ-GD-T
	0	1	3	5	10	0	1	3	5	10	0
PascalVOC	59.8	52.9	58.9	59.3	59.6	64.7	64.7	67.4	68.1	68.8	57.5
AerialDrone	15.8	22.1	29.8	31.0	31.6	17.4	30.7	36.1	37.0	36.1	13.6
Aquarium	23.5	31.7	36.1	40.1	42.4	30.3	39.2	45.8	47.0	49.7	18.5
Rabbits	75.4	76.2	77.4	75.6	75.5	71.8	76.0	75.0	74.3	75.3	79.9
EgoHands	41.2	64.3	66.0	66.7	68.2	57.2	68.8	68.1	71.6	72.6	65.4
Mushrooms	61.0	89.7	89.0	91.8	89.0	63.9	87.4	91.6	90.7	92.2	68.2
Packages	68.5	71.9	72.8	73.7	74.4	53.0	70.6	71.2	72.0	73.5	64.1
Raccoon	41.6	61.2	64.8	65.5	61.9	58.1	70.9	73.3	72.0	76.7	49.2
Shellfish	26.6	27.9	34.2	41.9	40.0	63.0	61.1	60.1	62.8	60.7	29.2
Vehicles	57.2	60.6	59.5	65.7	65.6	63.2	68.3	70.2	71.2	72.4	56.7
Pistols	59.6	56.5	60.3	61.4	61.7	74.4	73.6	72.7	74.3	74.8	69.2
Pothole	14.7	26.7	28.0	33.7	36.4	27.0	30.9	30.8	36.9	38.7	25.2
Thermal	48.0	59.4	64.2	62.4	72.7	58.7	68.5	72.2	72.5	74.5	64.9
Average	45.6	53.9	57.0	59.1	59.9	54.1	62.4	64.2	65.4	66.6	50.9

B.2 Explicit evaluation on an open-vocabulary detection setting

To further investigate the transferability of MQ-Det, we evaluation our models on a clear separation of base and novel classes, which is similar to previous open-vocabulary object detection [6, 4]. We first construct a novel category set from 1,203 LVIS categories. Specifically, we remove the LVIS categories that exist in the 365 classes of Objects365 and finally obtain 986 novel categories that

did not appear during our modulated pre-training. The remaining 217 categories are represented as base categories. Then, we conduct finetuning-free inference with 5 vision queries on the separated categories to verify the generalization of multi-modal query learning. Tab. IV shows the results. The results indicate that multi-modal queries generalize well to novel classes that do not exist in the modulated pre-training. Specifically, +4.1%, +5.7%, and +6.3% AP on novel classes of MQ-GroundingDINO-T, MQ-GLIP-T, and MQ-GLIP-L over their baselines, respectively.

Table IV: Finetuning-free detection with explicit open-vocabulary category separation on LVIS.

Model	AP_{novel}	AP_{base}	AP_{all}
GroundingDINO-T	22.1	36.7	25.6
GLIP-T	20.8	42.0	26.0
GLIP-L	35.4	45.5	37.9
MQ-GroundingDINO-T	26.2	43.0	30.2
MQ-GLIP-T	26.5	42.8	30.4
MQ-GLIP-L	41.7	51.3	44.0

It is worth noting that the separation of base and novel classes differs from previous works on open-vocabulary detection (OVD) [5]. The reason is that the testing categories of previous separation are partially included in our pre-training dataset Objects365 [3]. Therefore, we represent the classes in LVIS that do not exist in our modulated pre-training dataset Objects365 as novel classes. The frequency distribution of the separated LVIS dataset is shown in Tab. V:

Table V: Frequency distribution of the separated LVIS for open-vocabulary evaluation.

Class	#Rare	#Common	#Frequent
Novel	326	404	256
Base	11	57	149
All	337	461	405

B.3 Training hyper-parameters

We report the hyper-parameter settings of the modulated pre-training of MQ-Det in Tab. VI. Other settings are the same with corresponding language-queried detectors.

Table VI: Hyper-parameters of modulated pre-training.

Item	Value	Item	Value
optimizer	AdamW	max vision query num (K)	5000
lr of GCP	1e-5	vision query num (k)	5
lr of gate	5e-3	mask rate	40%
weight decay	1e-4	layer with GCP	6$\sim$12

C Further discussion

Limitations. First, multi-modal queries make limited contribution with sufficient training data for each category. This may because the foundation models learn enough accurate classification boundaries, thus reducing the effectiveness of language and vision queries. Second, the applications of MQ-Det on other dense prediction tasks such as segmentation remain unexplored.

Broader impacts. MQ-Det shows strong downstream transfer ability with highly flexible category vocabularies. This allows inexperienced users to easily use MQ-Det models (*e.g.*, MQ-GLIP) for their own needs by simply providing some visual examples and corresponding text descriptions. However, this also raises concerns about how our MQ-Det models with a large vocabulary could be used inappropriately in the community, such as for large-scale illegal video surveillance. The

open-set detection capabilities could be manipulated through specialized visual or textual cues to facilitate targeted detections instead of generic ones. This manipulation could introduce biases in the detector and result in unfair predictions.

Table VII: Per-dataset AP performance (%) on ODinW-35. We report results on 0, 3-shot detection. MQ-GD-T denotes MQ-GroundingDINO-T. Specifically, the "zero-shot" here actuall stands for the finetuning-free setting with 5 vision queries.

Dataset	MQ-GLIP-T		MQ-GLIP-L	MQ-GD-T
	0	3	0	0
AerialMaritimeDrone_large	15.8	29.8	17.4	13.6
AerialMaritimeDrone_tiled	18.3	27.1	20.8	21.9
AmericanSignLanguageLetters	1.8	19.9	3.0	0.1
Aquarium	23.5	36.1	30.3	18.5
BCCD_BCCD	4.3	57.3	12.4	12.3
ChessPiece	10.7	64.8	2.8	14.9
CottontailRabbits	75.4	77.4	71.8	79.9
DroneControl_Drone_Control	6.4	40.9	9.3	1.2
EgoHands_generic	41.2	66.0	57.2	65.4
EgoHands_specific	3.6	32.1	6.7	0.1
HardHatWorkers	5.4	37.4	5.5	4.8
MaskWearing	0.3	46.3	1.0	0.0
MountainDewCommercial	49.2	46.4	17.7	39.9
NorthAmericaMushrooms	61.0	89.0	63.9	68.2
OxfordPets_by-breed	0.4	31.5	0.5	0.4
OxfordPets_by-species	1.6	68.1	0.4	0.5
PKLot_640	0.9	30.3	2.6	0.0
Packages	68.5	72.8	53.0	64.1
Raccoon_Raccoon	41.6	64.8	58.1	49.2
ShellfishOpenImages	26.6	34.2	63.0	29.2
ThermalCheetah	2.3	41.1	11.1	7.1
UnoCards	0.2	35.4	0.8	0.0
VehiclesOpenImages	57.2	59.5	63.2	56.7
WildFireSmoke	13.2	22.5	20.5	14.6
boggleBoards	0.1	76.5	0.1	0.0
brackishUnderwater	4.5	31.3	5.3	3.3
dice_mediumColor	0.4	14.6	0.6	0.1
openPoetryVision	0.0	3.1	0.0	0.1
pistols	59.6	60.3	74.4	69.2
plantdoc	1.7	12.5	1.5	0.2
pothole	14.7	28.0	27.0	25.2
selfdrivingCar	8.2	17.5	9.0	6.9
thermalDogsAndPeople	48.0	64.2	58.7	64.9
websiteScreenshots	1.0	6.6	1.7	1.0
PascalVOC	59.8	58.9	64.7	57.5
Average	20.8	43.0	23.9	22.5

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[2] 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, 2022.

[3] 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*, pages 8430–8439, 2019.

[4] Yifan Xu, Mengdan Zhang, Xiaoshan Yang, and Changsheng Xu. Exploring multi-modal contextual knowledge for open-vocabulary object detection. *arXiv preprint arXiv:2308.15846*, 2023.

[5] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.

[6] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 350–368. Springer, 2022.

Figure I: Visualized results of MQ-GLIP-T on the LVIS benchmark.