WeDetect: Fast Open-Vocabulary Object Detection as Retrieval

Supplementary Material

Table S1. WeDetect training dataset.

#Sample	Dataset
5.0M	OpenImagesV6 (1.74M) [11], Objects365 V2 (1.74M) [27] V3Det (0.18M) [28], ImageNetBox (0.54M) [4] GoldG (0.8M) [14]
15.0M	self-collected data (15.0M)

Table S2. WeDetect-Ref training dataset.

Stage	#Sample	Task	Dataset
Stage1	685k	image caption	LLaVA Pretrain (558k) [18]
		region caption	refcoco/+/g (127k) [9, 22, 31]
Stage2	1685k	image QA	LLaVA-OneVision-SI (835k) [12]
		grounded conversation	VCR (213k) [33], Osprey (145k) [32]
		region caption	refcoco/+/g (56k) [9, 22, 31] DAM-LVIS (90k) [15], VG (346k) [10]
Stage3	3911k	detection	V3Det (1122k) [28], LVIS (661k) [6] COCO (698k) [16]
		REC	refcoco/+/g (127k) [9, 22, 31], flickr (185k) [26] humanref (135k) [8], Ref-L4 (45k) [2] grefcoco (209k) [17], FineCops-Ref (182k) [19] reircoco (547k) [7]

Table S3. WeDetect-Ref training settings.

Config	Stage1	Stage2	Stage3
training module	projector	projector + LLM	projector + LLM
training objective	language modeling loss	language modeling loss	sigmoid focal loss
batch size	256	128	128
learning rate	$1e^{-3}$	$1e^{-5}$	$1e^{-5}$
epoch	1	1	1

A. Limitation

Different from other LMMs with next-token prediction, WeDetect-Ref is a binary classification model, which does not support detecting multiple queries in a single forward pass. Instead, we can detect multiple queries in multiple times and then merge the results, similar to the inference pipeline of Grounding-DINO on LVIS. High performance on COCO and OdinW13 in Table 5 demonstrates the effectiveness of this mechanism. Further, as WeDetect-Ref enjoys a 13x speedup compared to Qwen3-VL, WeDetect-Ref can still run faster than it with a few queries while getting higher performance.

B. Implementation Details

WeDetect is trained with around 20M samples, including 15M self-collected data and 5M open-sourced data. Details are summarized in Table S1. Each class name is encoded separately and is represented as a single embedding. For head and neck initialization, the model is trained for 20 epochs with a learning rate of $5e^{-4}$. For end-to-end training, the model is trained for 30 epochs with a learning rate of $1e^{-5}$. The total batch size is 320.

WeDetect-Ref is finetuned from Qwen3-VL [1] using a three-stage training strategy. The training data and configurations are summarized in Table S2 and Table S3, respectively. As shown in Table 7, we find that incorporating negative detection data is crucial for achieving high detection performance. To this end, for each image in the detection datasets, we select three negative class names as negative training samples. These negative class names are chosen based on the top-3 confidence predictions from WeDetect among the negative classes, serving as hard negative examples. The proposals used for both training and evaluation are the top 100 boxes generated by WeDetect-Base-Uni. During training, if a ground truth box has no overlap with any of the proposed boxes, it is added directly to the proposal set to ensure that every ground truth instance has at least one positive match. The input image size is constrained to range from 900 * 32 * 32 to 1600 * 32 * 32, corresponding to 900–1600 visual tokens. Since WeDetect-Ref can process only one query per forward pass, we perform inference separately for each query and then merge the results when evaluating on object detection benchmarks. For example, on the COCO dataset, the model is run 80 times per image, once for each category.

Table S4. Zero-shot evaluation on the D^3 [29] dataset.

Model	Full	Pres	Abs
OFA-DOD-B [29]	21.6	23.7	15.4
Grounding-DINO-B [20]	20.7	20.1	22.5
OWLv2 [24]	22.8	22.1	24.7
GLIP-T [14]	19.1	18.3	21.5
FIBER-B [5]	22.7	21.5	26.0
Gen-Enhanced-Negs [35]	26.0	25.2	28.1
Weak-to-Strong [25]	30.8	31.0	30.4
InternVL2-8B [3]	9.8	11.0	6.2
Griffon-13B [34]	12.3	12.4	12.2
Groma-7B [21]	16.0	15.9	16.3
InternVL2-76B [3]	25.3	25.7	23.5
ROD-MLLM-7B [30]	29.7	30.0	28.7
WeDetect-Ref 2B	41.8	43.9	35.4
WeDetect-Ref 4B	**42.0**	**44.0**	**35.8**

C. More Experiment Results

C.1. Performance on Language-Based Object Detection

Language-Based Object Detection expands category names to flexible language expressions for open-vocabulary object detection (OVD) and overcomes the limitation of referring

Table S5. Detailed zero-shot results on COCO [16].

Model	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
WeDetect-Tiny	44.9	61.4	48.9	26.6	49.8	62.3
WeDetect-Base	52.1	69.4	57.0	34.8	57.1	69.2
WeDetect-Large	54.5	72.9	59.7	42.2	59.9	66.5

expression comprehension (REC) only grounding the pre-existing object, which is a combination task of them. We evaluate our model on the commonly used D^3 [29] dataset. As shown in Table S4, WeDetect-Ref significantly outperforms other models by more than 10 AP, thanks to its high performance on both OVD and REC.

C.2. Detailed WeDetect Evaluation Results

In Table S5, Table S6, and Table S7, we provide detailed evaluation results of WeDetect on COCO [16], ODinW35 [13], and COCO-O [23], respectively. The high and balanced performance across all benchmarks demonstrates the superior open-vocabulary capacities of the WeDetect series.

D. Visualization

D.1. Visualizations of Self-Collected Dataset

In this work, we develop a data engine to curate a high-quality dataset characterized by balanced concepts, exhaustive annotations, and multi-granularity labels. Examples are shown in Figure S1. For example, in the first picture, the streetlight will be annotated with a hierarchical label list "[Urban facilities, Lighting equipment, streetlight]".

D.2. Visualizations of Inference Results of WeDetect-Uni

In Figure S2, we visualize some top-scoring proposals extracted by WeDetect-Large-Uni. The model can extract proposals containing both the whole objects and the main parts of the objects with a high recall rate.

D.3. Visualizations of Inference Results of WeDetect-Ref

In Figure S3, we visualize several referring expression comprehension examples. We demonstrate that WeDetect-Ref is capable of handling: (a) queries with compositional descriptions, (b) queries involving spatial directions, (c) queries requiring high-level semantic understanding, and (d) OCR-related queries. Interestingly, although WeDetect-Ref is trained exclusively on English data, it can still process multilingual queries (f), thanks to the strong multilingual foundation provided by Qwen3-VL [1].

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1, 2

[2] Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *CVPR*, 2025. 1

[3] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[5] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *NeurIPS*, 2022. 1

[6] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 1

[7] Xiangzhao Hao, Kuan Zhu, Hongyu Guo, Haiyun Guo, Ning Jiang, Quan Lu, Ming Tang, and Jinqiao Wang. Referring expression instance retrieval and a strong end-to-end baseline. In *ACM MM*, 2025. 1

[8] Qing Jiang, Lin Wu, Zhaoyang Zeng, Tianhe Ren, Yuda Xiong, Yihao Chen, Liu Qin, and Lei Zhang. Referring to any person. In *ICCV*, 2025. 1

[9] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 1

[10] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 1

[11] 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

Table S6. Detailed zero-shot results on ODinW35 [13].

Dataset	ODinW13	ODinW35	WeDetect-Tiny	WeDetect-Base	WeDetect-Large
AerialMaritimeDrone large	✓	✓	10.8	13.5	11.0
AerialMaritimeDrone tiled		✓	19.2	19.7	25.4
AmericanSignLanguageLetters		✓	2.7	4.5	5.1
Aquarium	✓	✓	22.6	30.0	33.3
BCCD		✓	5.2	4.0	9.4
boggleBoards		✓	0.6	0.5	1.2
brackishUnderwater		✓	4.2	5.8	7.5
ChessPieces		✓	6.8	3.9	11.9
CottontailRabbits	✓	✓	78.3	78.8	79.1
dice		✓	2.0	1.4	2.5
DroneControl		✓	2.9	1.7	5.3
EgoHands generic	✓	✓	58.2	62.4	65.0
EgoHands specific		✓	17.7	21.2	20.3
HardHatWorkers		✓	7.2	9.3	10.3
MaskWearing		✓	1.2	1.8	2.4
MountainDewCommercial		✓	33.0	41.9	44.2
NorthAmericaMushrooms	✓	✓	39.8	55.0	41.3
openPoetryVision		✓	0.0	0.2	1.1
OxfordPets by breed		✓	0.7	1.9	2.6
OxfordPets by species		✓	1.9	11.8	8.9
PKLot		✓	0.0	0.0	0.8
Packages	✓	✓	67.9	68.3	71.3
PascalVOC	✓	✓	61.1	64.9	66.6
pistols	✓	✓	52.1	65.2	69.5
plantdoc		✓	1.7	2.2	4.2
pothole	✓	✓	13.6	24.5	24.3
Raccoons	✓	✓	53.1	50.9	52.3
selfdrivingCar		✓	8.2	9.3	10.4
ShellfishOpenImages	✓	✓	45.7	59.9	60.7
ThermalCheetah		✓	0.8	0.1	2.9
thermalDogsAndPeople	✓	✓	40.7	53.5	52.8
UnoCards		✓	0.0	0.0	1.1
VehiclesOpenImages	✓	✓	59.9	63.3	67.3
WildfireSmoke		✓	14.0	25.4	24.3
websiteScreenshots		✓	3.3	3.6	5.1
ODinW13 Average			46.4	53.1	53.4
ODinW35 Average			21.1	24.6	25.8

Table S7. Detailed zero-shot results on COCO-O [23].

Model	Cartoon	Handmake	Painting	Sketch	Tattoo	Weather	Average
WeDetect-Tiny	43.5	28.8	42.9	44.7	27.1	44.7	38.6
WeDetect-Base	52.7	37.4	52.0	50.7	22.4	49.6	44.1
WeDetect-Large	53.8	37.4	51.5	50.5	35.7	52.5	47.0

detection, and visual relationship detection at scale. *IJCV*, 2020. 1

[12] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1

[13] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng

Figure S1. Visualizations of self-annotated data. Each object is annotated with hierarchical labels.

Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. In *NeurIPS*, 2022. 2, 3

[14] 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 *CVPR*, 2022. 1

[15] Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao,

Figure S2. Visualizations of top-scoring proposals generated by WeDetect-Large-Uni.

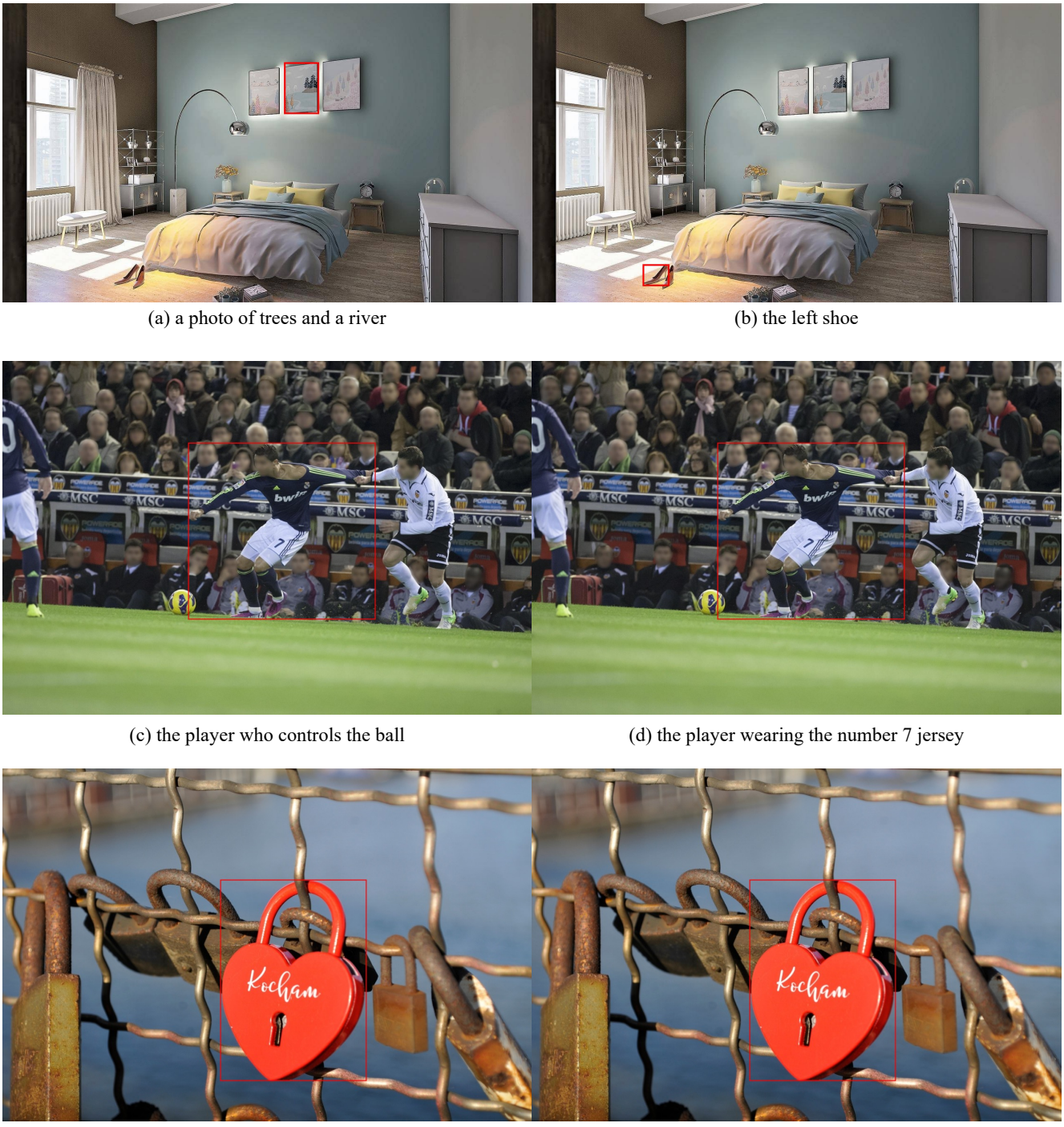
Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, et al. Describe anything: Detailed localized image and video captioning. In *ICCV*, 2025. 1

[16] 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 *ECCV*, 2014. 1, 2

[17] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, 2023. 1

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[19] Junzhuo Liu, Xuzheng Yang, Weiwei Li, and Peng Wang. Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. *arXiv preprint arXiv:2409.14750*, 2024. 1



(a) a photo of trees and a river

(b) the left shoe

(c) the player who controls the ball

(d) the player wearing the number 7 jersey

(e) the lock with a heart shape

(f) 心形的锁

Figure S3. Visualizations of referring expression comprehension results of WeDetect-Ref 4B.

[20] 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 *ECCV*, 2024. 1

[21] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *ECCV*, 2024. 1

[22] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In

CVPR, 2016. 1

[23] Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, and Hui Xue. Coco-o: A benchmark for object detectors under natural distribution shifts. In *ICCV*, 2023. 2, 3

[24] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *NeurIPS*, 2023. 1

[25] Kwanyong Park, Kuniaki Saito, and Donghyun Kim. Weak-to-strong compositional learning from generative models for language-based object detection. In *ECCV*, 2024. 1

[26] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *CVPR*, 2015. 1

[27] 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 *ICCV*, 2019. 1

[28] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *ICCV*, 2023. 1

[29] Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described object detection: Liberating object detection with flexible expressions. In *NeurIPS*, 2023. 1, 2

[30] Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. Rod-mllm: Towards more reliable object detection in multimodal large language models. In *CVPR*, 2025. 1

[31] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1

[32] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *CVPR*, 2024. 1

[33] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 1

[34] Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *ECCV*, 2024. 1

[35] Shiyu Zhao, Long Zhao, Yumin Suh, Dimitris N Metaxas, Manmohan Chandraker, Samuel Schulter, et al. Generating enhanced negatives for training language-based object detectors. In *CVPR*, 2024. 1