Detect Anything via Next Point Prediction

Qing Jiang[1,2‡], Junan Huo[1,2], Xinyu Chen[2,3,4], Yuda Xiong[1,2], Zhaoyang Zeng[1,2]
Yihao Chen[1,2], Tianhe Ren[1,2], Junzhi Yu[3], Lei Zhang[1,2†]
[1] South China University of Technology
[2] International Digital Economy Academy (IDEA)
[3] Peking University [4] Zhongguancun Academy
https://rex-omni.github.io/

A. More Details on Training Data

A.1. Public Datasets

In Table 1, we enumerate the publicly available datasets leveraged for Rex-Omni's training across various subtasks, including Object Detection, Object Referring, Visual Prompting, OCR, Layout Grounding, GUI Grounding, Pointing, Affordance Grounding, Spatial Referring, and Keypointing. For each of these tasks, a set of question templates was defined to construct corresponding question-answer (QA) pairs. In total, approximately 8.9 million public data samples were utilized.

A.2. Illustration of the Data Engines

We present illustrations of our Grounding Data Engine and Referring Data Engine in Figure 1. For grounding data engine, our specific annotation process is as follows:

- **Image Captioning:** We begin by generating descriptive captions for each image using Qwen2.5-VL-7B-Instruct. These captions provide natural language descriptions of the visual content, typically covering multiple objects within the scene. The prompt used in this stage is *Please describe this image in details.*
- **Phrase Extraction:** We then apply the SpaCy[1] NLP toolkit to extract noun phrases from the generated captions. These phrases may include basic class names (e.g., tabletop, lemon) as well as more specific descriptions (e.g., sliced yellow lemons, green lemons).
- **Phrase Filtering:** This step marks a key departure from prior approaches. To minimize data ambiguity, we remove noun phrases containing descriptive attributes such as adjectives (e.g., green lemon is discarded, while lemon is retained). The rationale is that current grounding models struggle to accurately interpret such descriptive expressions, often detecting all instances of a category regardless of the modifier. For instance, the phrase green lemon may incorrectly trigger detections of all lemons, thereby introducing significant labeling errors.
- **Phrase Grounding:** Finally, we use DINO-X [30], an open-vocabulary object detector, to produce bounding boxes corresponding to the filtered phrases.

For referring data engine, our specific annotation process is as follows:

- **Expression Generation:** Given an image annotated with bounding boxes and corresponding category labels, we prompt Qwen2.5-VL-7B with the image and category information to generate a set of referring expressions. Each expression is designed to naturally describe an object category present in the image, mimicking human-like descriptions. The prompt used in this stage is shown in Figure 2
- **Pointing:** For each generated referring expression, we employ Molmo [7], a state-of-the-art referring model, to produce the corresponding spatial point. Although Molmo outputs only point-level predictions, it exhibits strong performance in understanding and grounding referring expressions.
- **Mask Generation:** We apply SAM [13] to generate a mask for each ground-truth bounding box in the image.
- **Point-to-Box Association:** Each point produced by Molmo is aligned with a SAM-generated mask. When a point lies within a mask, the corresponding bounding box is linked to the referring expression, thereby grounding the language in the object region.

B. More Details on Training Pipelines

B.1. Implementaion Details for SFT Stage

We adopt the standard cross-entropy loss for training. The model is trained on 8 nodes, each equipped with 8 A100 GPUs, and the total training time is approximately 8 days. All model parameters are updated during training. We

This work was done during Qing, Junan's internship and Xinyu's academic visit at IDEA. † Corresponding author. ‡Project Lead.

[1] https://spacy.io/

Task	Output Format	Question Template Example	Datasets
Object Detection	Box	Detect [PHRASE] in this image	APTv2 [43], BDD100K [44], DeepFashion [22] DOTAv2 [40], EgoObjects [50], FAIR-1M [36] HumanParts [19], ImageNet-Part [9] NuImages [4] PACO [29], OpenImages [14], O365 [32] V3Det [38], VisDrone [8]
Object Referring	Box	Detect [PHRASE] in this image	HumanRef, RefCOCOg
Visual Prompting	Box	Given reference boxes [BOX] indicating one or more objects, find all objects with the same category	O365 [32], OpenImages [14], HierText [23] CrowdHuman [31], SA-1B [13], VisDrone [8] FSCD147 [27]
OCR	Box / Polygon	Detect all the text in box/polygon format and recognize them	Art [6], HierText [23], ICDAR2013 [25] ICDAR2015 [12], LSVT [37], RCTW [33] ReCTS [46] SROIE [10], TextOCR [34] IDLOCR [3], WildReceipt [35]
Layout Groudning	Box	Detect [PHRASE] in this image	DocLayNet [28], PubLayNet [48], TableBank [18] M6Doc [5], CDLA [16], TabRecSet [41]
GUI Grounding	Box / Point	Detect/Point to element [PHRASE]	Os-Atlas [39], UI-RefExp [2], ShowUI [20]
Pointing	Point	Point to [PHRASE]	Pixmo-point [7]
Affordance	Point	Point to [PHRASE]	AGD20K [26]
Spatial Referring	Point	Point to [PHRASE]	RefSpatial [49]
KeyPointing	Box & Point	Can you detect each [PHRASE] in the image in box format, and then provide the coordinates of its [KEYPOINT] as [x0, y0]? Output the answer in JSON format.	AP10K [45], APT36K [42], COCO-Keypoint [21] MacaquePose [15], HumanArt [11], MPII [1] OCHuman [47], CrowdPose [17]

Table 1. Publicly available training datasets used by Rex-Omni, covering tasks such as object detection, referring, prompting, OCR, grounding, pointing, affordance, and keypointing, with outputs including boxes, points, polygons, and JSON-formatted keypoints.

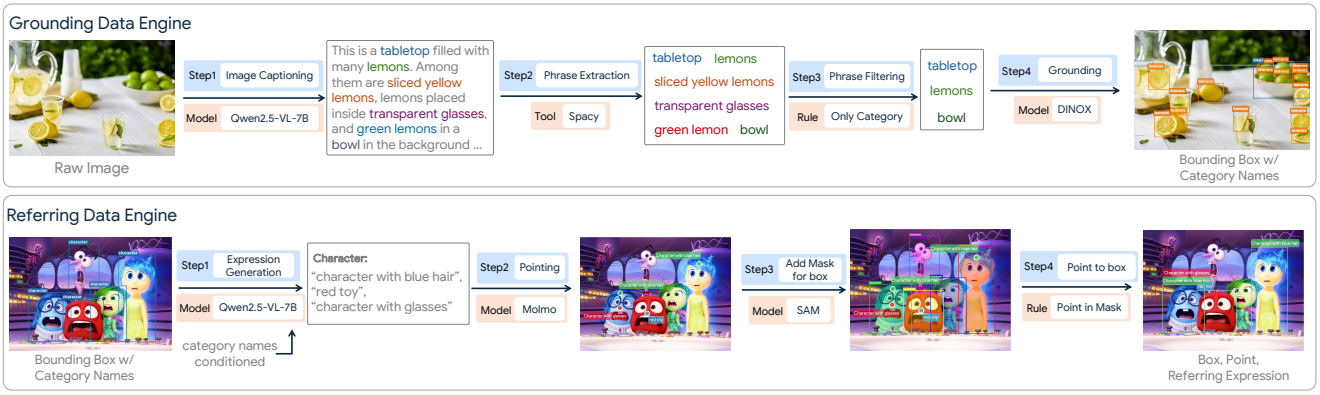

Figure 1. Pipelines of our two primary data engines. The figure illustrates the processes of the Grounding Data Engine (top) and the Referring Data Engine (bottom), which are custom-designed to produce extensive, high-quality grounding and referring data for Rex-Omni's training.

use separate learning rates for different components: 2e-6 for the vision encoder, and 2e-5 for both the projection layer and the LLM. Optimization is performed using the AdamW [24] optimizer with a learning rate warm-up of 3% and a weight decay of 0.01. Following the architecture of Qwen2.5-VL, Rex-Omni also employs a native resolution Vision Transformer as its vision encoder. We constrain the number of input pixels to range from a minimum of $16 \times 28 \times 28$ to a maximum of $2560 \times 28 \times 28$. Given a ViT patch size of 28, this limits the number of image tokens between 16 and 2560.

B.2. Implementation Details for GRPO Stage

Geometry-aware Rewards: To provide informative feedback on the spatial quality of predictions, we design three geometry-aware reward functions tailored to different tasks: box IoU reward, point-in-mask reward, and point-in-box reward. These reward types reflect the structural correctness of the predicted outputs with respect to ground-truth annotations.

Box IoU Reward. This reward is applied to tasks requiring bounding box predictions, including object detection, grounding, referring, and OCR. The reward encourages both accurate localization and correct object-category

You are given an image and a list of object categories to focus on.

Your job is to generate a dictionary of referring expressions for the specified categories only.

📩 Input:

You will receive:

An image (to be visually analyzed)

A list of target categories, e.g. ["apple", "person", "helmet"]

📌 Output Requirements:

You must generate a Python dictionary where:

Each key is one of the provided category names

Each value is a list of mutually exclusive referring expressions that describe subsets of objects within that category

📌 Referring Expression Rules:

Each referring expression must:

Refer only to objects of the given category (no cross-category mixing)

Describe a visually meaningful subset (e.g., "red apples", "people sitting")

Be mutually exclusive with other expressions under the same category

No object should be described by more than one expression for its category

Avoid vague or overlapping phrases (e.g., don't mix "red helmets" and "red and yellow helmets")

Avoid describing all instances unless appropriate (be discriminative when possible)

Be visually grounded, natural language expressions

✅ Output Format:

```
{
 "category1": ["referring expression 1", "referring expression 2", ...],
 "category2": ["referring expression 1", ...]
}
```

Only include the categories provided in the input list.

✅ Example:

Given Categories:

["apple", "person"]

Image Description:

3 red apples, 2 green apples, 2 people sitting, 2 people standing.

Expected Output:

```
{
 "apple": ["red apples", "green apples"],
 "person": ["people sitting", "people standing"]
}
```

❌ Bad Examples:

❌ Overlapping Referring Expressions:

```
{
 "bottle": ["green bottles", "bottles on the table"]  # if the same green bottles are on the table
}
```

❌ Cross-Category Mixing:

```
{
 "person": ["cameramen and salesmen"]
}
```

❌ Mixed Types in One Phrase:

```
{
 "helmet": ["red and yellow helmets"]
}
```

Now, analyze the image and generate referring expressions only for the provided categories, formatted as described. The provided categories are [<CATEGORIES>].

Figure 2. The prompt used for Qwen2.5-VL-7B to generate referring expressions.

alignment.

Given a set of predicted boxes $\hat{B} = \{\hat{b}_1, \ldots, \hat{b}_m\}$ and the ground-truth boxes $B^* = \{b_1^*, \ldots, b_n^*\}$, we perform a ground-truth-guided matching. For each ground-truth box b_j^*, we find the predicted box $\hat{b}_i$ that maximizes the IoU with b_j^*:

$$\text{IoU}(b_j^*, \hat{b}_i) = \max_{\hat{b}_i \in \hat{B}} \text{IoU}(b_j^*, \hat{b}_i). \tag{1}$$

If the category label of $\hat{b}_i$ matches that of b_j^*, we assign the IoU value as the reward r_j for that ground-truth box. Otherwise, $r_j = 0$. Let $R = \{r_1, \ldots, r_n\}$ be the reward set

for all GT boxes. We then compute recall and precision as follows:

$$\text{Recall} = \frac{\sum_{j=1}^{n} r_j}{n}, \quad \text{Precision} = \frac{\sum_{j=1}^{n} r_j}{m}, \quad r^{\text{IoU}} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall} + \epsilon}, \tag{2}$$

where ϵ is a small constant to prevent division by zero. This formulation rewards both spatial accuracy and label correctness. It penalizes unmatched or misclassified predictions and balances over- and under-prediction through the F1-style reward signal.

Point-in-Mask Reward. This reward is applied to tasks where the model localizes objects via point predictions, such as pointing-based detection, grounding, and referring. It evaluates whether a predicted point lies within the object mask.

Given a set of ground-truth bounding boxes $B^* = \{b_1^*, \ldots, b_n^*\}$, we apply SAM to extract a binary mask M_j for each ground-truth box b_j^*. Let $\hat{P} = \{\hat{p}_1, \ldots, \hat{p}_m\}$ denote the predicted points, each associated with a category label. For each ground-truth mask M_j, we determine whether there exists a predicted point $\hat{p}_i$ that lies inside M_j:

$$\exists \hat{p}_i \in \hat{P}, \quad \text{s.t.} \quad \hat{p}_i \in M_j. \tag{3}$$

If such a point exists and its associated category label matches that of M_j, we assign a reward of 1 to the corresponding ground-truth instance; otherwise, the reward is 0. Precision, recall, and F1 reward are then computed using the same formulation as in the Box IoU Reward.

Point-in-Box Reward. This reward is specifically designed for the GUI Grounding task, where the model is expected to predict a point indicating the clickable position (e.g., a button) on a graphical user interface. If the predicted point falls within the ground-truth bounding box of the target GUI element, a reward of 1 is assigned; otherwise, the reward is 0. This simple binary reward effectively encourages precise point-level interaction behavior required in GUI scenarios.

We sample 66K data from the SFT dataset to serve as training data for the GRPO stage. We reuse the same dialogue templates from the SFT phase. The GRPO training is conducted on 8 A100 GPUs for approximately 24 hours. We set the rollout size to 8, the KL penalty coefficient β to 0.01, and use a batch size of 64. All model parameters are updated during this stage.

C. Dense200 Benchmark

We introduce Dense200, a manually collected dataset consisting of 200 densely annotated images covering 109 categories. In Dense200, each image contains an average of 91.2 bounding boxes, with an average size of 66.8×64.5 pixels. It poses significant challenges due to the combination of small object sizes and high object density, demanding precise spatial reasoning and accurate localization.

D. Visualization Results

We visualize the performance of Rex-Omni and other models in various scenarios in Figure 3, Figure 4, Figure 5, Figure 6, Figure 7, Figure 8, Figure 9. We also show some visualization results of Rex-Omni in different tasks:

- Common and Long-tailed Object Detection (Figure 10)
- Dense Object Detection (Figure 11)
- Object Referring (Figure 12)
- Object Pointing (Figure 13)
- Layout Grounding (Figure 14)
- OCR (Optical Character Recognition) (Figure 15)

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 2

[2] Chongyang Bai, Xiaoxue Zang, Ying Xu, Srinivas Sunkara, Abhinav Rastogi, Jindong Chen, et al. Uibert: Learning generic multimodal representations for ui understanding. *arXiv preprint arXiv:2107.13731*, 2021. 2

[3] Ali Furkan Biten, Ruben Tito, Lluis Gomez, Ernest Valveny, and Dimosthenis Karatzas. Ocr-idl: Ocr annotations for industry document library dataset. In *European Conference on Computer Vision*, pages 241–252. Springer, 2022. 2

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2

[5] Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15138–15147, 2023. 2

[6] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1571–1576. IEEE, 2019. 2

[7] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1, 2

Figure 3. Visualization of detection predictions from different models on common and long-tailed object detection benchmarks, using COCO and LVIS, respectively.

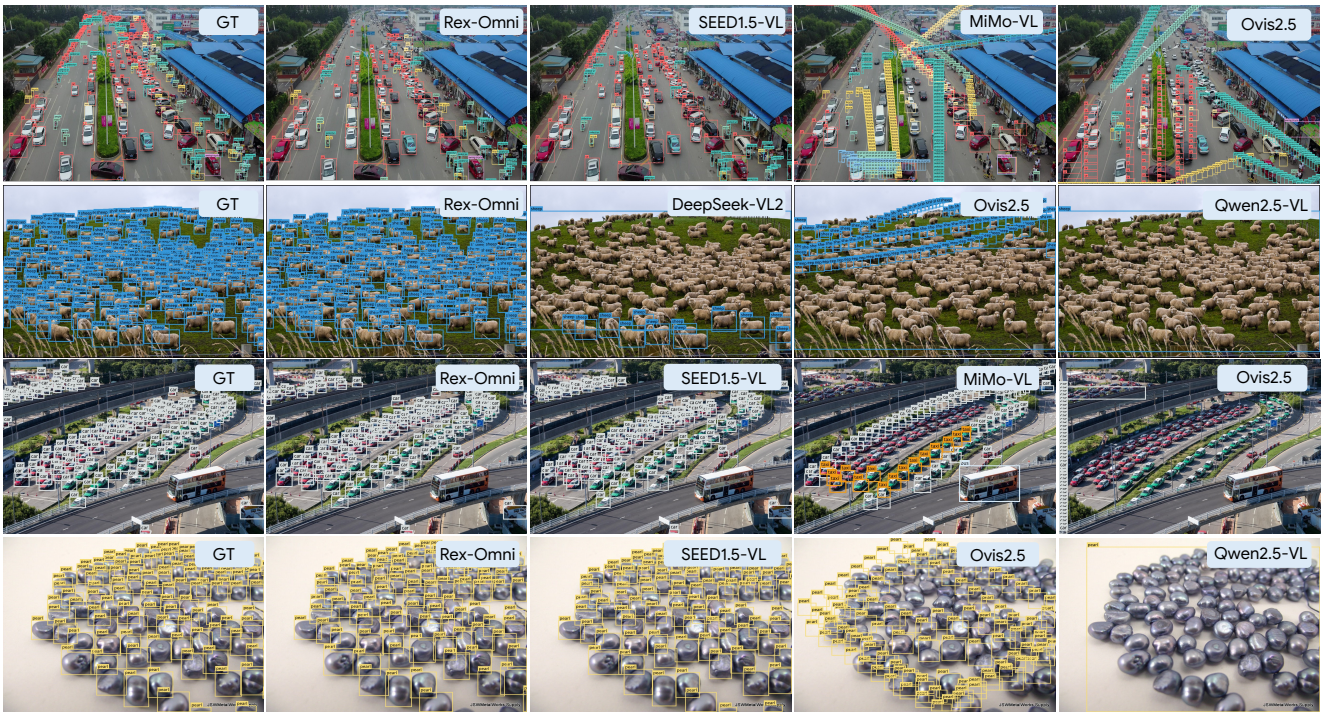

Figure 4. Visualization of dense and tiny object detection predictions. This figure presents a qualitative comparison of various models on the VisDrone and Dense200 datasets.

[8] Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 2

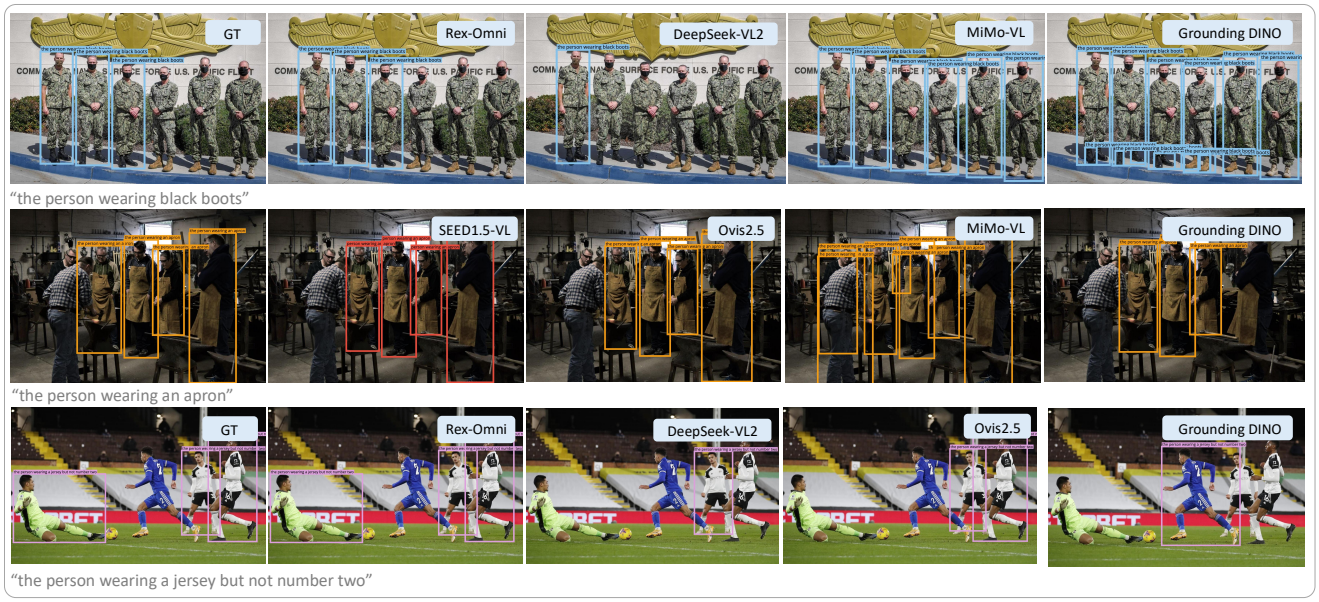

Figure 5. Visualization of model predictions on referring object detection benchmarks.

Figure 6. Qualitative comparison of visual prompting predictions between T-Rex2 and Rex-Omni.

[9] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022. 2

[10] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar 2019 robust reading challenge on scanned receipts ocr and information extraction. In *International conference on document analysis recognition*, 2019. 2

[11] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 618–629, 2023. 2

[12] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 2

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *arXiv: 2304.02643*, 2023. 1, 2

[14] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R.

Figure 7. Qualitative comparison of object pointing predictions from different models.

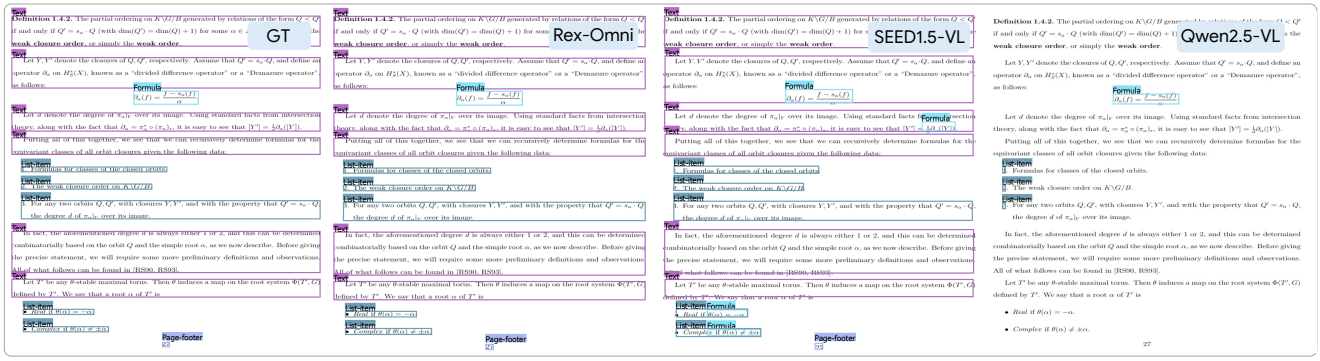

Figure 8. Qualitative comparison of layout grounding predictions from different models. The figure illustrates the models' ability to localize and interpret various layout elements.

Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *arXiv: 1811.00982*, 2018. 2

[15] Rollyn Labuguen, Jumpei Matsumoto, Salvador Blanco Negrete, Hiroshi Nishimaru, Hisao Nishijo, Masahiko Takada, Yasuhiro Go, Ken-ichi Inoue, and Tomohiro Shibata. Macaquepose: a novel "in the wild" macaque monkey pose dataset for markerless motion capture. *Frontiers in behavioral neuroscience*, 14:581154, 2021. 2

[16] Hang Li. Cdla: A chinese document layout analysis (cdla) dataset, 2021. 2

[17] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019. 2

[18] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. Tablebank: Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1918–1925, 2020. 2

[19] Xiaojie Li, Lu Yang, Qing Song, and Fuqiang Zhou.

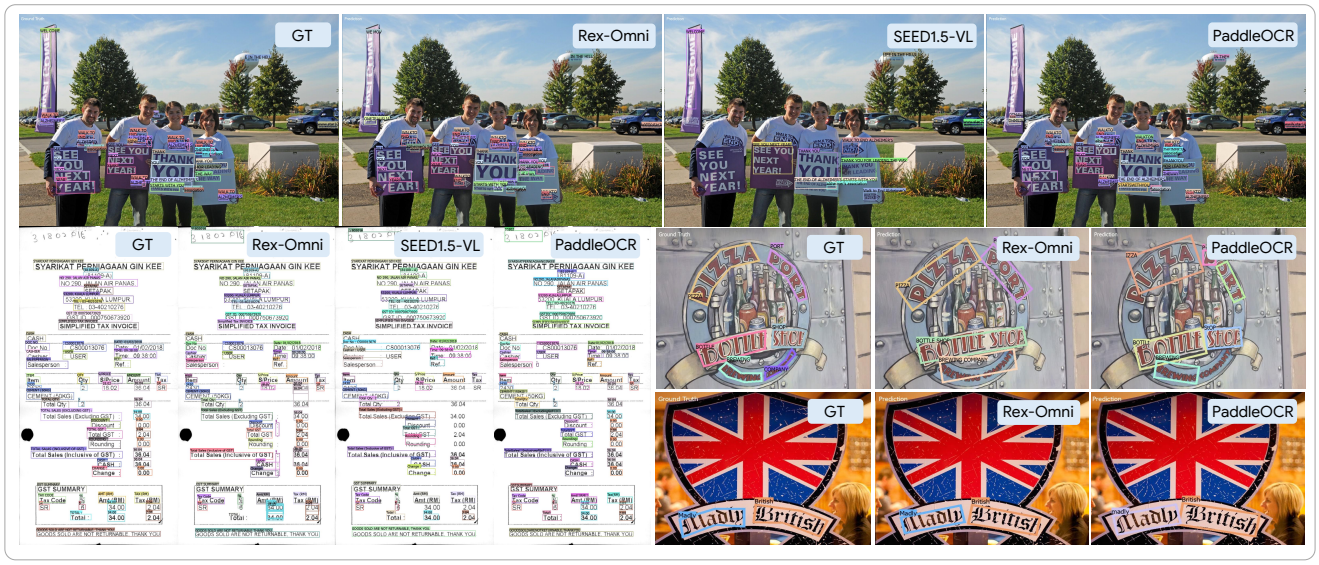

Figure 9. Visualization of OCR results across models.

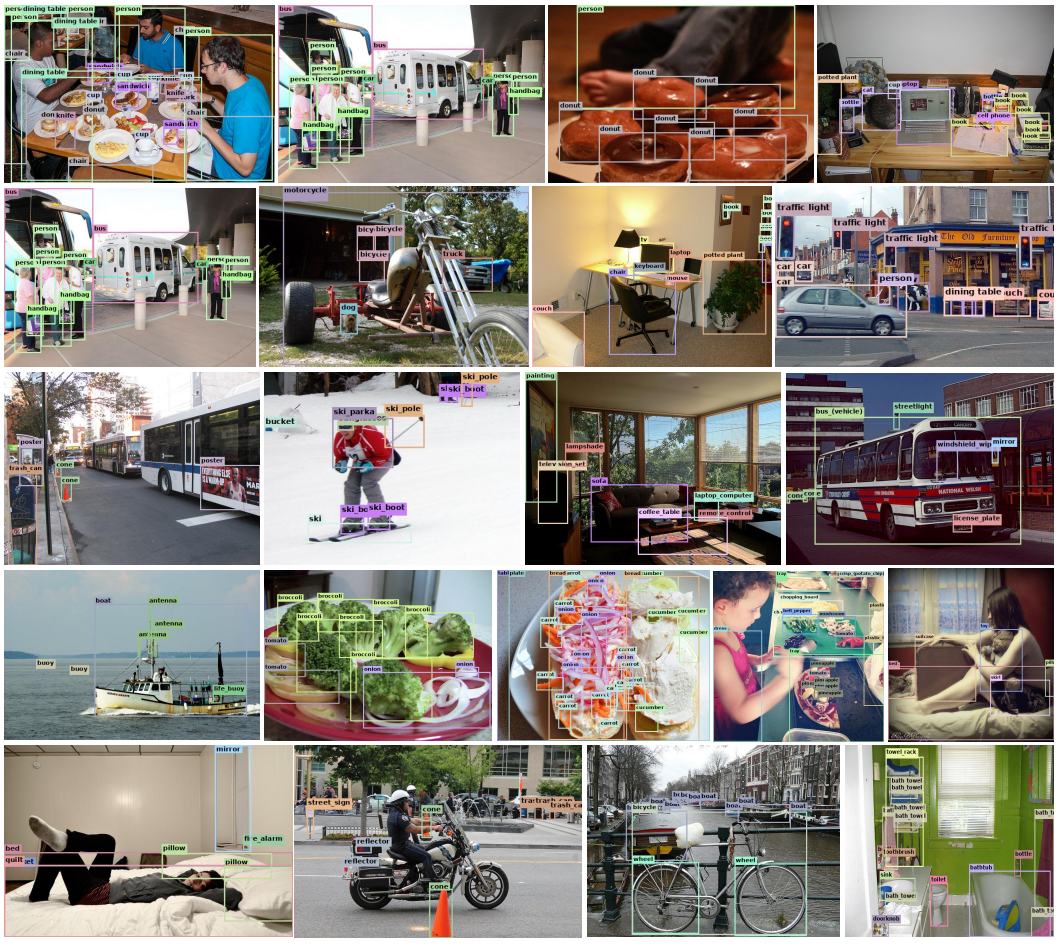

Figure 10. Visualization results of Rex-Omni on common and long-tailed object detection task.

Figure 11. Visualization results of Rex-Omni on dense object detection task.

Detector-in-detector: Multi-level analysis for human-parts. In *Asian Conference on Computer Vision*, pages 228–240. Springer, 2018. 2

[20] Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Zechen Bai, Weixian Lei, Lijuan Wang, and Mike Zheng Shou. Showui: One vision-language-action model for generalist gui agent. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024. 2

[21] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 2

[22] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 2

[23] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis.

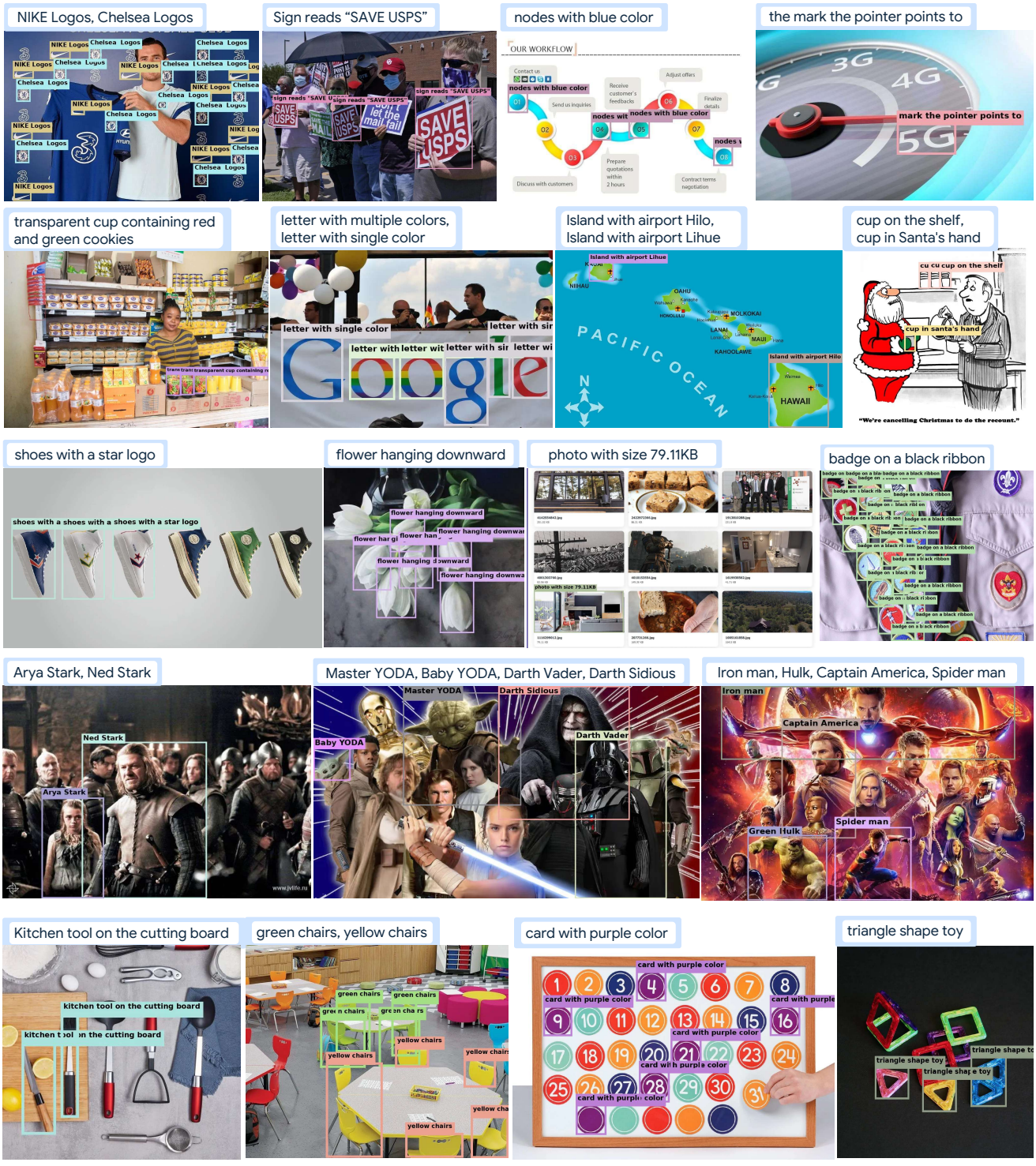

Figure 12. Visualization results of Rex-Omni on object referring task.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2022. 2

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[25] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai,

Figure 13. Visualization results of Rex-Omni on object pointing task.

Masayuki Okamoto, Hiroaki Yamamoto, et al. Icdar 2003 robust reading competitions: entries, results, and future directions. *International Journal of Document Analysis and Recognition (IJDAR)*, 7(2):105–122, 2005. 2

[26] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022. 2

[27] Thanh Nguyen, Chau Pham, Khoi Nguyen, and Minh Hoai. Few-shot object counting and detection. In *European Conference on Computer Vision*, pages 348–365. Springer, 2022. 2

[28] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3743–3751, 2022. 2

[29] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023. 2

Figure 14. Visualization results of Rex-Omni on layout grounding task.

[30] Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, et al. Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024. 1

[31] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 2

[32] 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. 2

[33] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017 14th iapr international conference on document analysis and recognition (ICDAR)*, pages 1429–1434. IEEE, 2017. 2

[34] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale

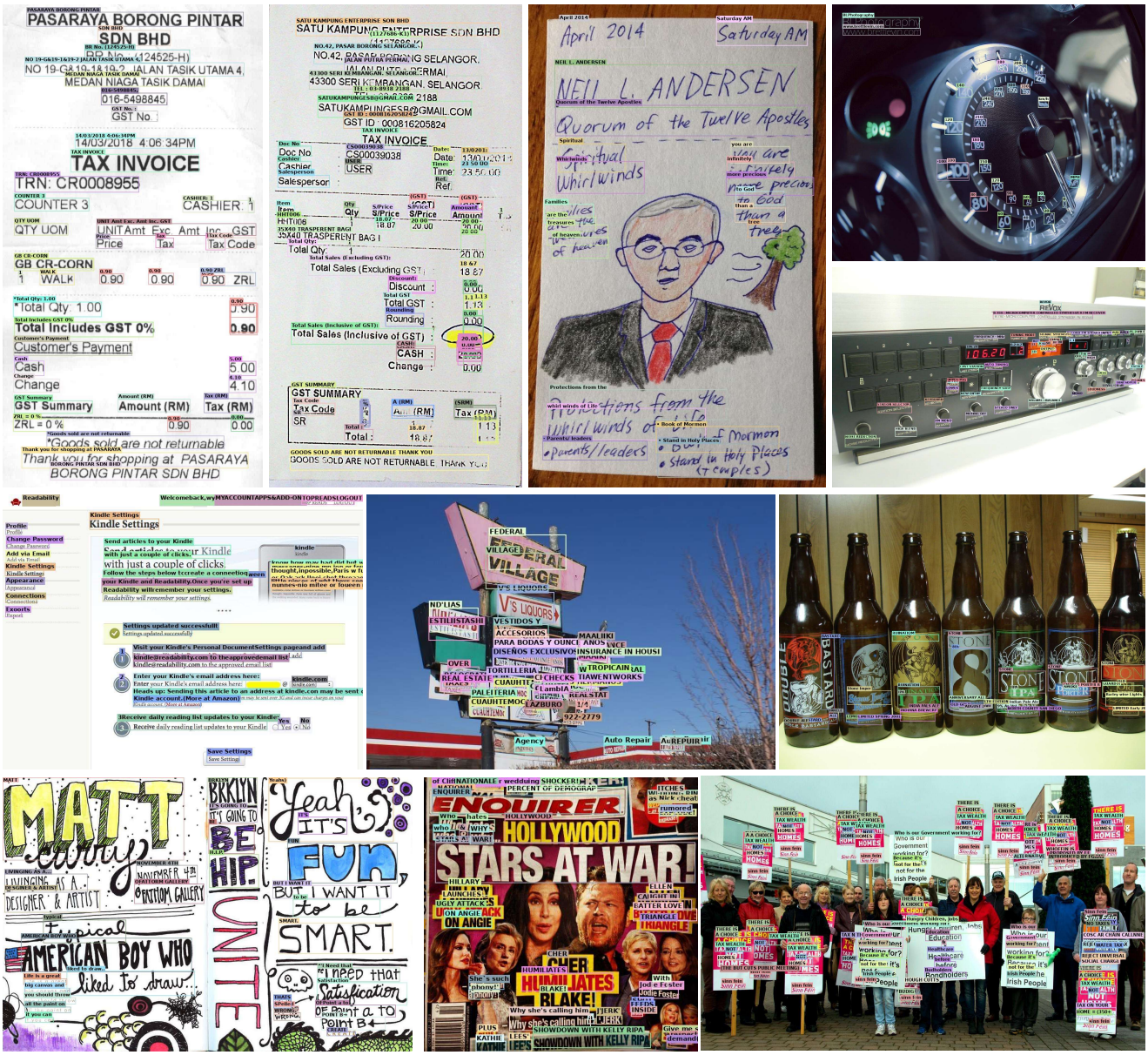

Figure 15. Visualization results of Rex-Omni on OCR task.

end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8802–8812, 2021. 2

[35] Hongbin Sun, Zhanghui Kuang, Xiaoyu Yue, Chenhao Lin, and Wayne Zhang. Spatial dual-modality graph reasoning for key information extraction. *arXiv preprint arXiv:2103.14470*, 2021. 2

[36] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. 2

[37] Yipeng Sun, Jiaming Liu, Wei Liu, Junyu Han, Errui Ding, and Jingtuo Liu. Chinese street view text: Large-scale chinese text reading with partially supervised learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9086–9095, 2019. 2

[38] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19844–19854, 2023. 2

[39] Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. Os-atlas: A foundation

action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*, 2024. 2

[40] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 2

[41] Fan Yang, Lei Hu, Xinwu Liu, Shuangping Huang, and Zhenghui Gu. A large-scale dataset for end-to-end table recognition in the wild. *Scientific Data*, 10(1):110, 2023. 2

[42] Yuxiang Yang, Junjie Yang, Yufei Xu, Jing Zhang, Long Lan, and Dacheng Tao. Apt-36k: A large-scale benchmark for animal pose estimation and tracking. *Advances in Neural Information Processing Systems*, 35:17301–17313, 2022. 2

[43] Yuxiang Yang, Yingqi Deng, Yufei Xu, and Jing Zhang. Aptv2: benchmarking animal pose estimation and tracking with a large-scale dataset and beyond. *arXiv preprint arXiv:2312.15612*, 2023. 2

[44] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 2

[45] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. *arXiv preprint arXiv:2108.12617*, 2021. 2

[46] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1577–1581. IEEE, 2019. 2

[47] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 889–898, 2019. 2

[48] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1015–1022. IEEE, 2019. 2

[49] Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. In *Adv. Neural Inform. Process. Syst.*, 2025. 2

[50] Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, and Zhicheng Yan. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20110–20120, 2023. 2