Appendix: Can We Challenge Open-Vocabulary Object Detectors with Generated Content in Street Scenes?

Annika Mütze[*]
Osnabrück University
Germany
annika.muetze@uos.de

Sadia Ilyas[*]
Aptiv Services Deutschland GmbH
Wuppertal, Germany
sadia.ilyas@aptiv.com

Chrsitian Dörpelkus
University of Wuppertal
Germany
doerpelkus@gmail.com

Matthias Rottmann
Osnabrück University
Germany
matthias.rottmann@uos.de

In this appendix, we provide technical details as well as additional visualizations for the discussions in the main paper.

A. Technical and implementation details

We start with a short overview of all models used in our experiments. In Tab. 1, we list the specification of models and the datasets they were trained on.

Grounding DINO [6] is formed on a Transformer-based detector DINO [12] enabling easy processing of both text and image data. Combining this architecture with grounded pre-training makes Grounding Dino an innovative open-vocabulary model, allowing the detection of various objects based on human prompts. The model consists of three main components: feature enhancer, language-guide query selection, and cross-modality decoder. First, multiscale features are extracted using Swin Transformer [8] as a backbone for image features and BERT [3] for text features. Vanilla features are then fed into the feature enhancer where self-attention along with image-to-text and text-to-image cross-attention are applied. For effectively using the prompt to steer object detection, the language-guided query selection module chooses out of the processed features the ones that are more relevant to the input text as decoder queries. In the cross-modality decoder, both image and text modalities are merged by employing self-attention, cross-attention using the text and image features and a feed-forward neural network.

YOLO-World [2] is an open-set model that extends the well-known YOLO object detection framework [10] by incorporating open-vocabulary detection through vision-language modeling. The open-set capabilities are based on a new re-parameterizable Vision-Language Path Aggregation Network (RepVL-PAN) and region-text contrastive loss to learn interaction between region-text pairs and thus visual and linguistic information. Other parts of the architecture are a YOLOv8 [4] backbone for image feature extraction and CLIP [9] as a model to convert the nouns of a given prompt into embeddings. Then both feature representations are fed to the RepVL-PAN and fused at multiple levels. Ultimately a text-contrastive head regresses bounding boxes and object embeddings.

MDETR [5] is an end-to-end text-modulated detection method derived from DETR [1]. It utilizes a convolutional backbone for visual feature extraction and a language model, RoBERTa [7], for text feature extraction.

OmDet-Turbo [13] introduces an Efficient Fusion Head (EFH) module, aiming for real-time performance. The overall OmDet-Turbo model features a text backbone, an image backbone and an EFH module. EFH includes an Efficient Language-Aware Encoder (ELA-Encoder) and Decoder (ELA-Decoder). ELA-Encoder selects top-K initial queries, fused with prompt embedding for language-guided multi-modality queries in ELA-Decoder.

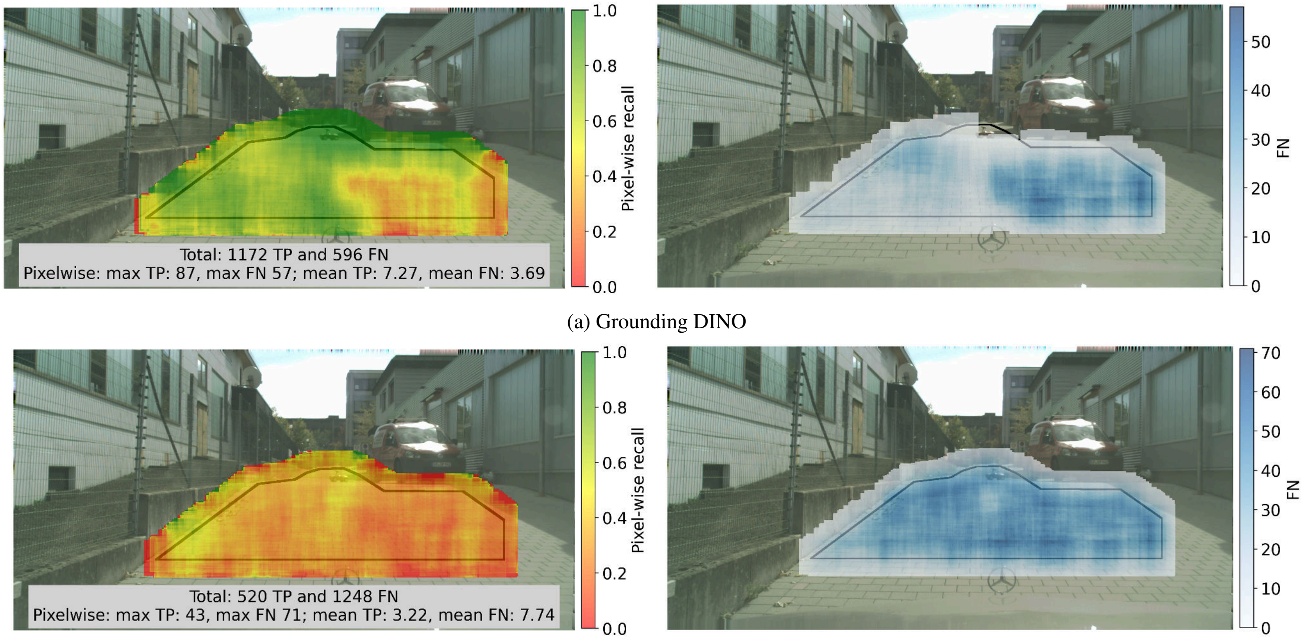
We use the above-mentioned open-vocabulary models by providing them with different text prompts and explore to what extent they could be utilized for object detection on synthetic data. We also use a classical closed-vocabulary object detection model, FasterRCNN [11], as reference to compare against the open-vocabulary models.

B. Model performance across image locations

As described in the main paper, to investigate our hypothesis that the location of an object has a significant influence on the detection, we generate 2000 copies of selected scene with different inpaintings at different locations.

Table 1. Information about the models belonging grouped according to different categories i.e., open-vocabulary and classical (object) detection. It also specifies the backbone and pre-trained data used for the models in our experiments.

Category	Model	Backbone	Pre-trained Data
open-vocabulary	YOLO-World-L	YOLOv8-L	O365, GoldG
	Grounding DINO-T	Swin-T	O365, GoldG, Cap4M
	MDETR	EfficientnetB5	VG, COCO, Flickr30k
	OmDet-Turbo	Swin-T	O365, GoldG
class. detection	FasterRCNN	ResNet50	COCO



(a) Grounding DINO

(b) YOLO-World

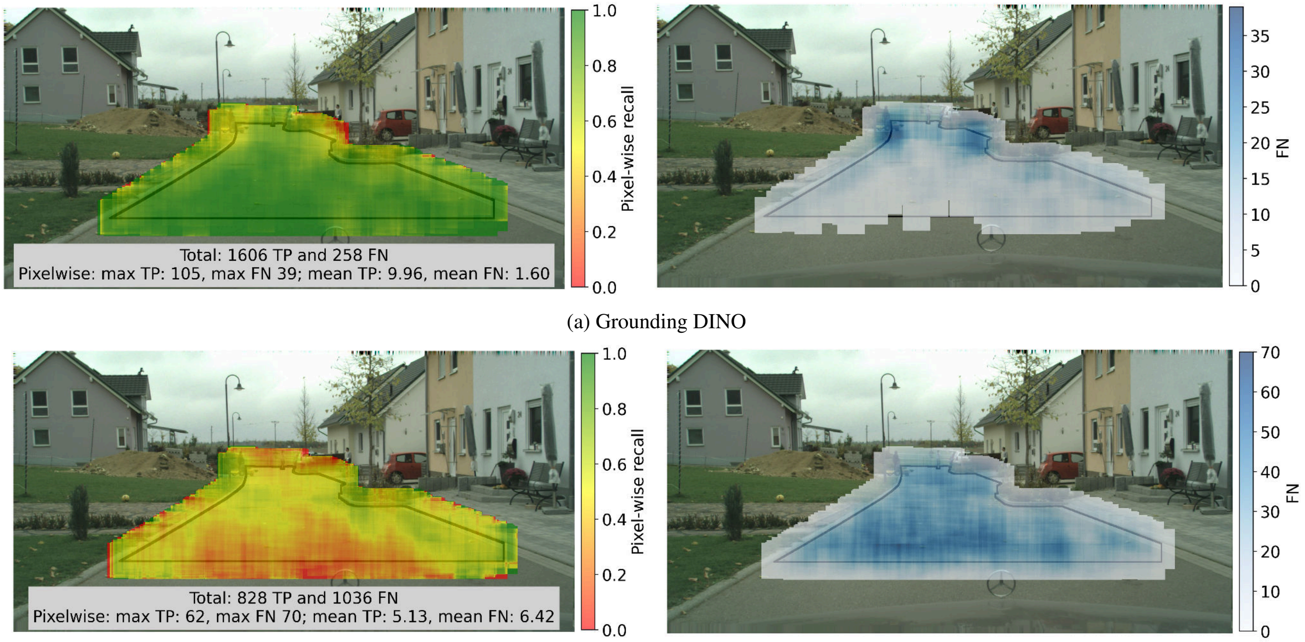
Figure 1. Pixel-wise recall heatmaps for the same scene with varying models used for detection. The open vocabulary models Grounding DINO and YOLO-World were tested based on the prompt “object on the street”.

When analyzing the prediction capability on random image locations across models, depicted exemplary in Figs. 1 and 2, we again observe FN clusters. However, those differ across the models. While YOLO-World can be challenged in most of the image regions, Grounding DINO has dedicated blind spots and performs more reliable on the remaining area. We fixed the prompt to test the models to “object in the street” in this experiment.

C. Additional experiments

Before and in between conducting the experiments in the main paper, we did some tests on a small image subset on similar detection tasks. This way, we wanted to rule out misleading errors or correlations. However, these experiments have not been realized on big sets of data. Therefore, we just explain them briefly in the following part. After our first detection tests on generated hybrids, we realized that synthetic content seemed to be relatively easy to detect for Grounding DINO. Therefore, we wanted to ensure that not the sake of synthetic content, rather than the actual appearance of objects or the underlying semantics results in TP detections. Considering that, we altered street scenes in different ways to see how this affects the detection performance.

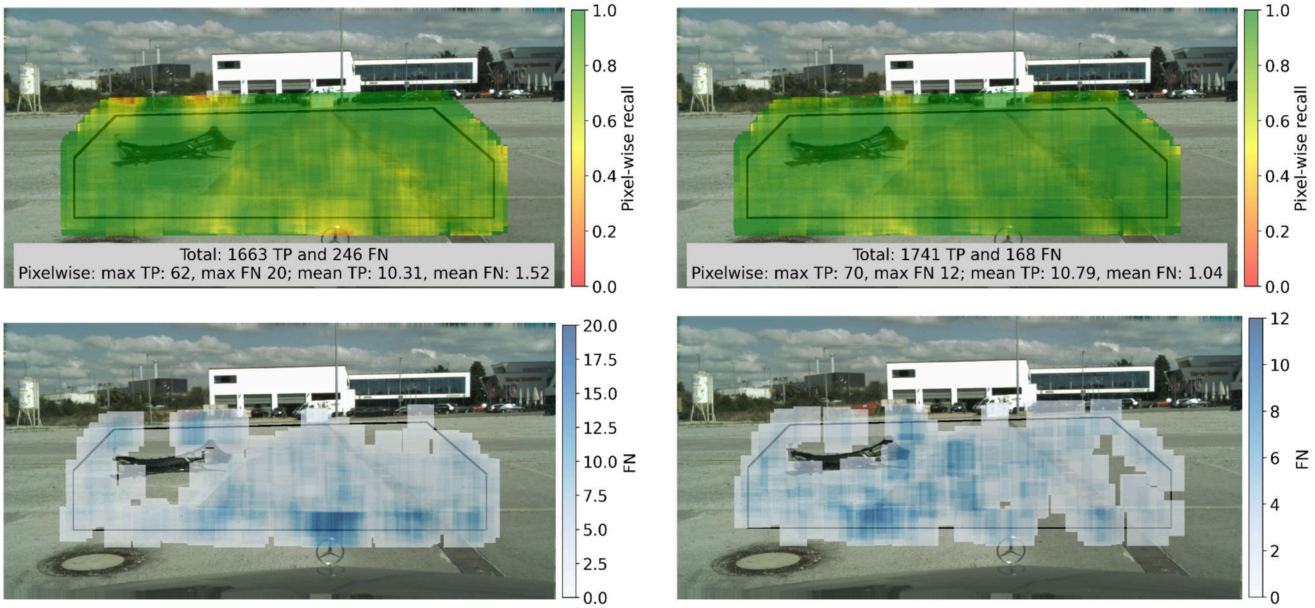
Noise. First, we wanted to see how object detectors react to noise in the form of white and gray ovals. For that we placed ovals with the according color based on the location of the ground truth bbox into the street scene images as shown in a) of Figure 4. This leads to sharp edges between the noise oval and the original content. Since object detectors are trained on common objects, these noise patterns should be hard to detect. However, if correctly predicted anyway, that would indicate a high relevance of the sharp



(a) Grounding DINO

(b) YOLO-World

Figure 2. Pixel-wise recall heatmaps for the same scene with varying models used for detection. The open vocabulary models Grounding DINO and YOLO-World were tested based on the prompt "object on the street".



(a) Prompt p3: *object on the street*

(b) Prompt p2: *object . animal . person*

Figure 3. Prediction quality of Grounding DINO with varying prompt. Heatmaps show pixel-wise recall as well as the number of FN for identical scenes from LostAndFound with different detection prompts.

edge for the detection performance of open-vocabulary object detectors, as the noise object shows only one color.

Pattern. Focused on the sharp edges, we manually placed snippets of the same road pattern onto the ground truth bbox as can be seen in b) of Fig. 4. Similar to the noise of the previous experiment, again we expect the detector to only detect the ‘object’ in the form of the pattern frame, by recognizing the sharp edges, as the frame itself provides no content.

Removed. When visually inspecting the inpaintings of some TP results manually, we could not rarely observe any obvious edges between the original image and the inpainted. However, we wanted to ensure that it is not the inpainted transition between the original image and the inpainted object that leads to a TP detection. In order to prove this, we studied failed inpaintings, i.e. those images we filtered out in the main paper which show no actual object, only artificial pixels that are adapted to the surroundings of the masked area.

Brightness smoothing. After examining some inpaintings, we observed that the inpainted content often seemed noticeably brighter than the rest of the image. Therefore, we suspected the different brightness to cause a TP detection independent of the actual object content. To cope with that, we started by calculating the brightness of each pixel in the masked area by following $Brightness = \frac{R+G+B}{3}$ with $R, G, B \in [0, 255]$ denoting the three color channels of an RGB image. Afterward, each pixel exceeding the threshold of $Brightness > 200$ has been replaced by the mean color $mC \in [0, 255]^{1\times 3}$ of the surrounding masked area $S \in [0, 255]^{n\times m\times 3}$. Herby, S was defined as $S = 2M - M$ with $M \in [0, 255]^{n\times m\times 3}$ being the originally masked area. Accordingly, $mC = mean(S)$ can be calculated easily. Thus, we retrieved objects integrating more smoothly into the street scene as seen in d) of Fig. 4.

D. Findings of additional experiments

All additional experiments have been performed with Grounding DINO using the prompt “object on the street”. As it was possible to automatically insert the noise ovals into the street scenes, we did so for the whole LostAndFound Dataset, obtaining 179 images for both white and gray ovals. On the white oval set, the detection led to 175 TP and 4 FN results. For the set with gray ovals, the detector provided 162 TP and 17 FN results. Next, the pattern dataset was tested. As these images were created manually, there are only 16 images. Running a detection on them delivered 1 TP and 15 FN results. We continued with the set showing a synthetic background in the mask area, as the original objects were removed during inpainting. As during the early stage of this work we just had some of these examples, this set contains 21. All of these were detected as FN from Grounding DINO. For the smoothed images we have not conducted quantitative results. However, we saw for particular examples that the confidential score of the prediction was mostly close or even higher compared to the non-smoothed image. One example for this, along with TP and FN results for the other experiments are shown in Fig. 4. Considering the high percentage of TP results on the noise sets brings the assumption that only the edges occur by adding or inpainting synthetic content matters for detection. However, object detectors being sensitive to this kind of edge in images is not surprising. Reflecting on the obvious noise ovals added in this case, we think retrieving some FN indicates that the actual inpainted content itself matters instead of the detector always detecting artificial content as TP. This is supported by the other experiments. Testing on the pattern and removed objects, which denotes synthetic empty content shows as expected and desired almost only FN results across our small experiments. These experiments indicate that inpainting objects is a reasonable way of systematically challenging open vocabulary object detectors.

E. Additional visualizations

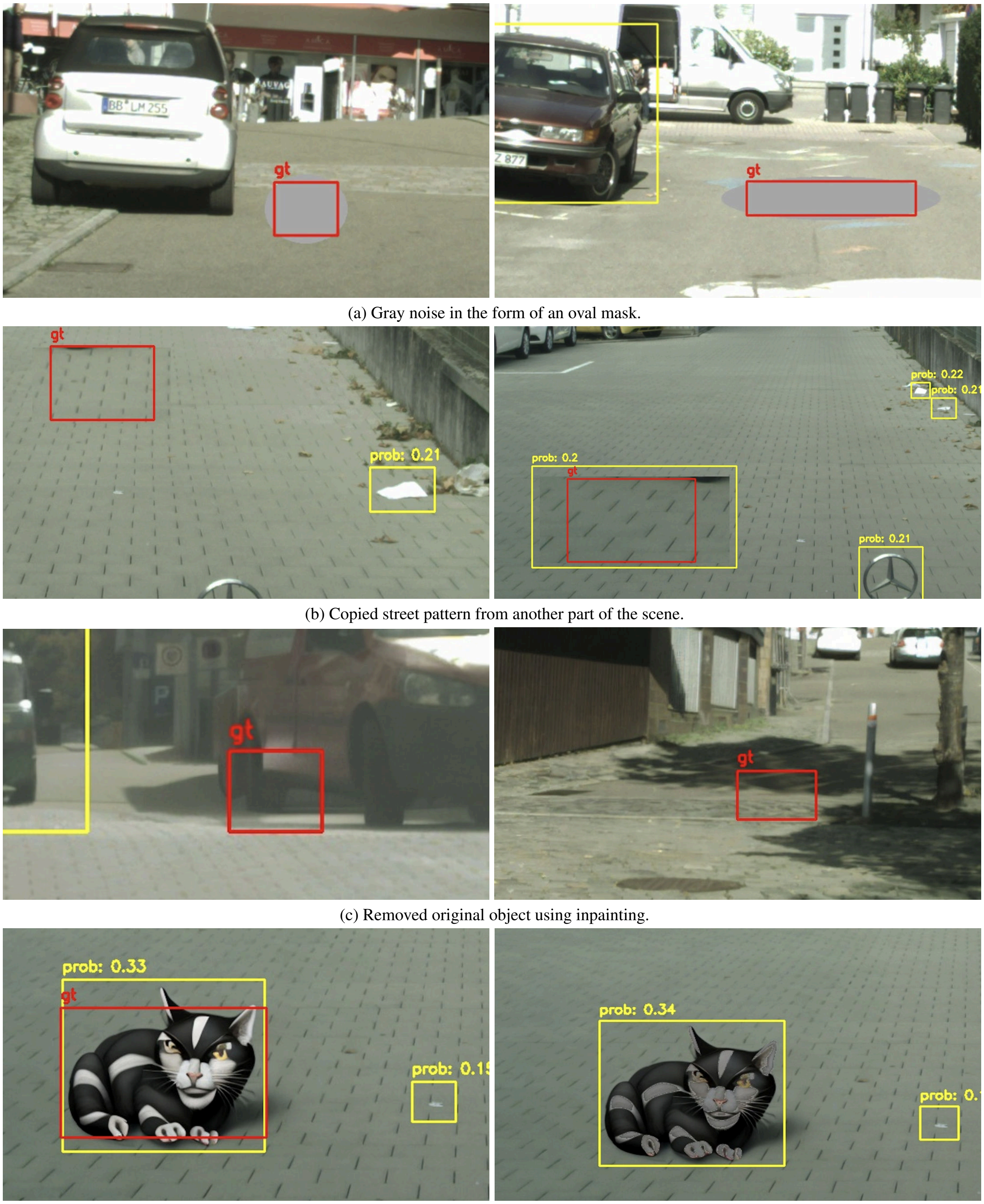
We present random examples of inpainted objects from the generated Hybrid-Concept-Inpainting LostAndFound dataset which were detected correctly by Grounding DINO provided the prompt “object on the street” in Fig. 5. Figure 6 depicts objects from the same dataset overlooked by Grounding DINO. Both figures give an idea of the variety and abnormality of the generated data.

F. Prompt list for Single-Concept-Inpainting

prompt_list = [’robot’, ’helicopter’, ’monster’, ’skateboard’, ’dog’, ’cat’, ’monkey’, ’horse’, ’elephant’, ’lion’, ’tiger’, ’bear’, ’deer’, ’rabbit’, ’squirrel’, ’wolf’, ’fox’, ’sheep’, ’goat’, ’chicken’, ’crocodile’, ’alligator’, ’hamster’, ’gerbil’, ’mouse’, ’rat’, ’guinea pig’, ’ferret’, ’rabbit’, ’cavy’, ’tapir’, ’hedgehog’, ’kangaroo’, ’koala’, ’panda’, ’zebra’, ’giraffe’, ’hippopotamus’, ’rhinoceros’, ’sloth’, ’antelope’, ’bison’, ’buffalo’, ’ostrich’, ’emu’,’penguin’, ’seal’, ’walrus’, ’manatee’, ’platypus’, ’okapi’, ’armadillo’, ’badger’, ’mole’, ’opossum’, ’raccoon’, ’porcupine’, ’weasel’, ’lemur’, ’gorilla’, ’chimpanzee’, ’orangutan’, ’tamarin’, ’sloth bear’, ’sea lion’, ’tortoise’, ’flamingo’, ’robot’, ’helicopter’, ’monster’, ’skateboard’, ”Sofa”, ”Coffee table”, ”Bookshelf”, ”Lamps”, ”Cutting board”, ”Pots pans”, ”Dishes”, ”Desk”, ”Chair”, ”Printer”, ”Vacuum cleaner”, ”Fan”, ”Clock”, ”Shoes”]

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1



(a) Gray noise in the form of an oval mask.

(b) Copied street pattern from another part of the scene.

(c) Removed original object using inpainting.

(d) Smoothed by applying a filter.

Figure 4. Examples of different types of synthetic content added to street scenes. Showing ground truth bboxes in red and predictions in yellow annotated with the confidence score.

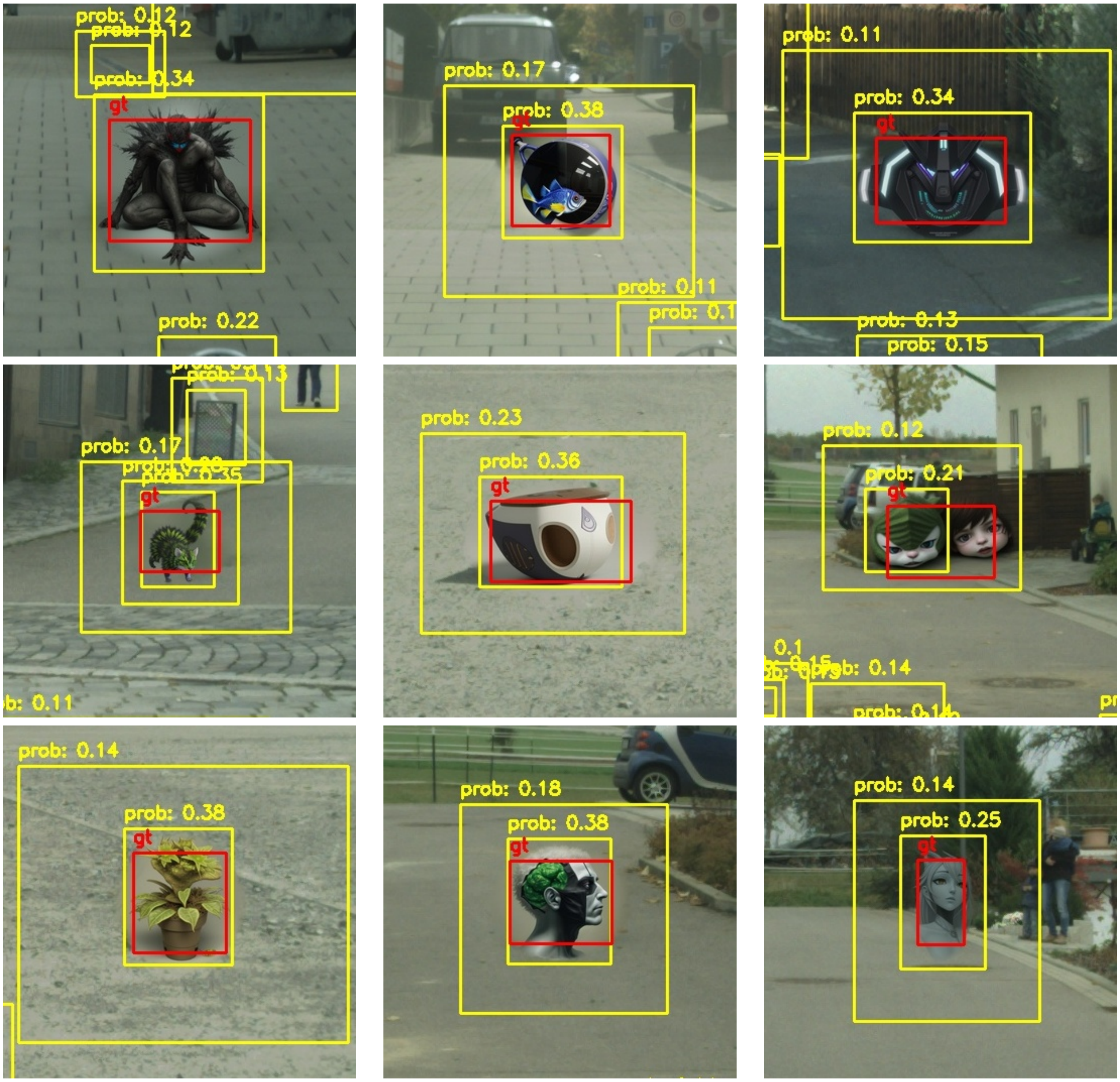


Figure 5. Examples for TP predictions using Grounding DINO with prompt "object on the street" and a confidence score of 0.1. Red bboxes annotated with gt denoting the ground truth of the original object used for inpainting. Yellow bboxes show model predictions with the respective confidence score.

[2] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. *arXiv preprint arXiv:2401.17270*, 2024. 1

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[4] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO. https://github.com/ultralytics/ultralytics, 2023. 1

[5] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 1

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang,

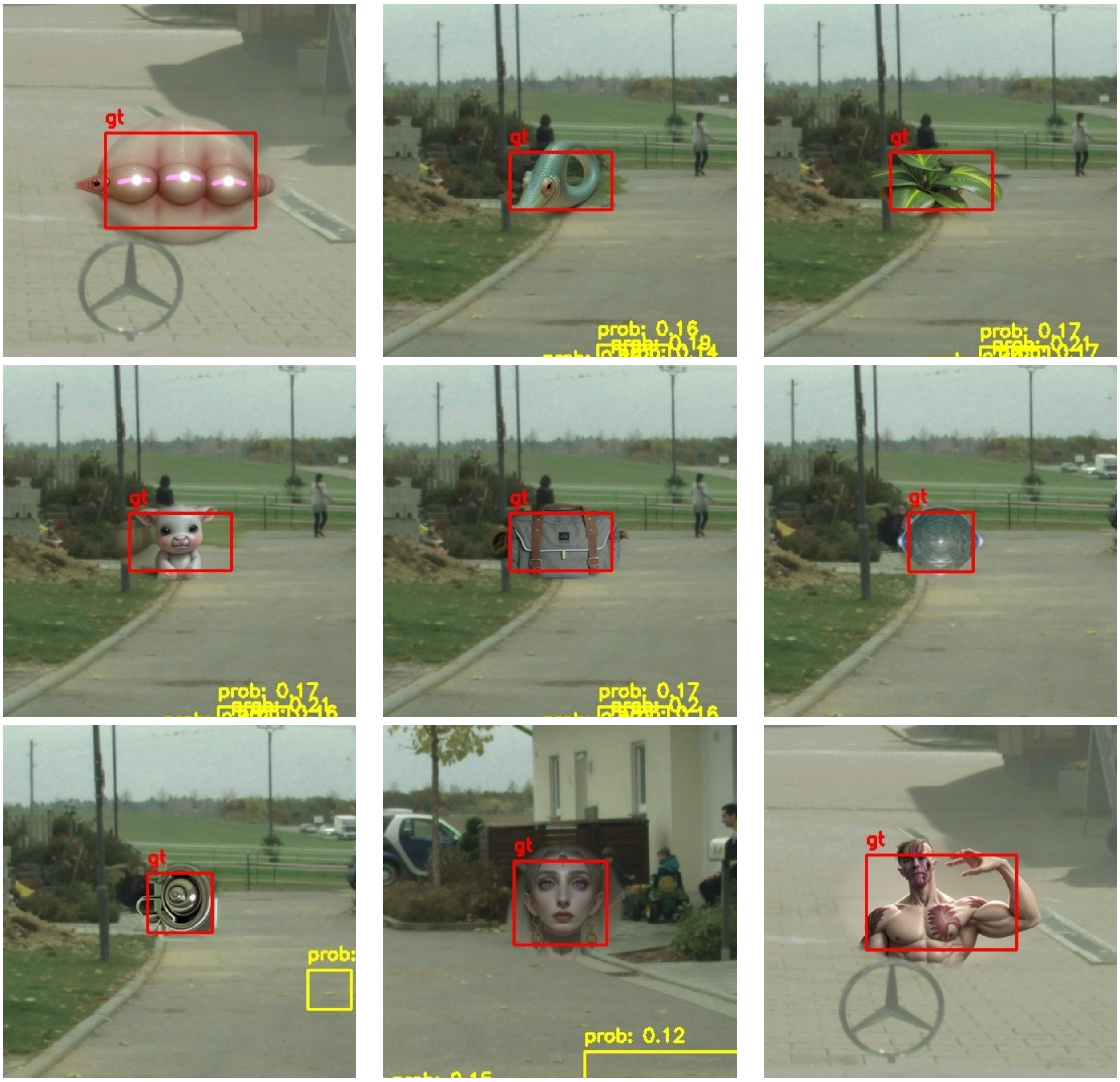


Figure 6. Examples for FN predictions using Grounding DINO with prompt "object on the street" and a confidence score of 0.1. Red bboxes annotated with gt denoting the ground truth of the original object used for inpainting. Yellow bboxes show model predictions with the respective confidence score.

Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[10] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2016. 1

[12] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 1

[13] Tiancheng Zhao, Peng Liu, Xuan He, Lu Zhang, and Kyusong Lee. Real-time transformer-based open-vocabulary detection with efficient fusion head. *arXiv preprint arXiv:2403.06892*, 2024. 1