Can OOD Object Detectors Learn from Foundation Models?
Supplementary Material

Jiahui Liu, Xin Wen, Shizhen Zhao, Yingxian Chen, and Xiaojuan Qi[⋆]

The University of Hong Kong

1 Extended Experiments

1.1 Study on the Source of Novel Categories

In the primary experiments, we employ GPT-4 [1] as the large language model (LLM) to generate novel objects for image editing tasks. To broaden the exploration of models capable of producing novel objects, we additionally incorporate Llama 2 [10] and WordNet [7], substituting GPT-4 with these alternatives for imagining novel objects.

Llama 2 The same prompts from the main experiments are introduced into Llama 2 [10] (70B) to collect a set of associative novel objects for driving subsequent image editing processes. Our observations indicate a distinct divergence in the responses of Llama 2 compared to those of GPT-4 [1] when the same ID object is given, for example:

> **ID Object:** 'dog'.
> **GPT-4:** 'wolf', 'fox', 'coyote', 'jackal', 'hyena'.
> **Llama 2:** 'stuffed animal (dog)', 'dog toy', 'dog bed', 'dog food bowl', 'dog leash'.

Different from GPT-4, while Llama 2 is capable of associating relevant concepts, it imagines some objects that lack volumetric or property alignment with the corresponding ID object (More imagined objects are presented in Tab. 3). Consequently, this discrepancy may lead to synthetic images that do not contain enough visually similar novel objects. It is further supported by the experiment results shown in Tab. 1, where the synthetic data derived from Llama 2 fails to match the performance levels achieved by GPT-4.

WordNet We use WordNet [7] based on its structured object relationship graph to retrieve novel objects *parallel (other hyponyms of the same hypernyms)* to ID objects to drive image editing (Retrieved objects are shown in Tab. 3). It better ensures the selection of novel objects that share similar properties and applicability compared to those provided by Llama 2. As shown in Tab. 1, synthetic

[⋆] corresponding author

Table 1: Ablation on different sources of novel categories (ID: PASCAL-VOC [3]).

Method	MS-COCO		OpenImages	
	FPR95↓	**AUROC↑**	**FPR95↓**	**AUROC↑**
Llama 2	39.21	85.49	14.53	95.05
WordNet	38.86	85.27	**13.30**	95.30
GPT4	**36.44**	**86.52**	13.34	**95.37**

Table 2: Ablation on LLMs' Prompts (ID: PASCAL-VOC [3]).

Method	MS-COCO		OpenImages	
	FPR95↓	**AUROC↑**	**FPR95↓**	**AUROC↑**
baseline	50.86	78.15	23.60	91.42
Attributes	40.50	85.16	14.01	95.09
Objects	**36.44**	**86.52**	**13.34**	**95.37**

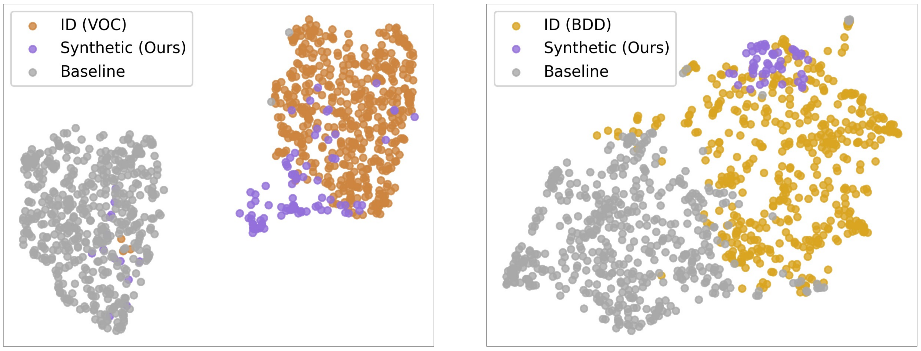

Fig. 1: Visualization with UMAP [6] of ID samples and OOD samples for optimizing ID/OOD decision boundaries. ID samples are marked as orange (PASCAL-VOC [3]) and yellow (BDD-100K [12]) points, while the synthetic OOD samples from our method and baseline method [11] are marked as purple and gray points, respectively.

images powered by WordNet provide better supervision than Llama 2 but worse than GPT-4. The superior logic and world knowledge of GPT-4 support the high performance of our synthetic data, which also means that a more powerful LLM in the future has greater potential to promote our method.

1.2 Study on Prompts for LLMs

From Objects to Attributes In addition to prompting the LLM to imagine novel objects, we further explore using the LLM to associate *normal objects with novel attributes.* We use the new prompt for GPT-4:

> Now if I provide you an object name, you should return to me words or short phrases that describe the abnormal state of the object to describe that this object is in a very abnormal state or has suffered an accident, so that the object appears visually abnormal. For example, I give you the word: car, you should response and only response: 'wrecked car', 'car with shattered glass', 'charred car', 'car with severe rust', 'car with exposed engine'.

The imagined concepts are presented in Tab. 3 and the fair experiment results are shown in Tab. 2. Compared with the baseline method [11], the introduction of novel attributes provides better OOD samples for optimizing the decision boundary. Nonetheless, novel attributes are not as effective as novel objects. As shown

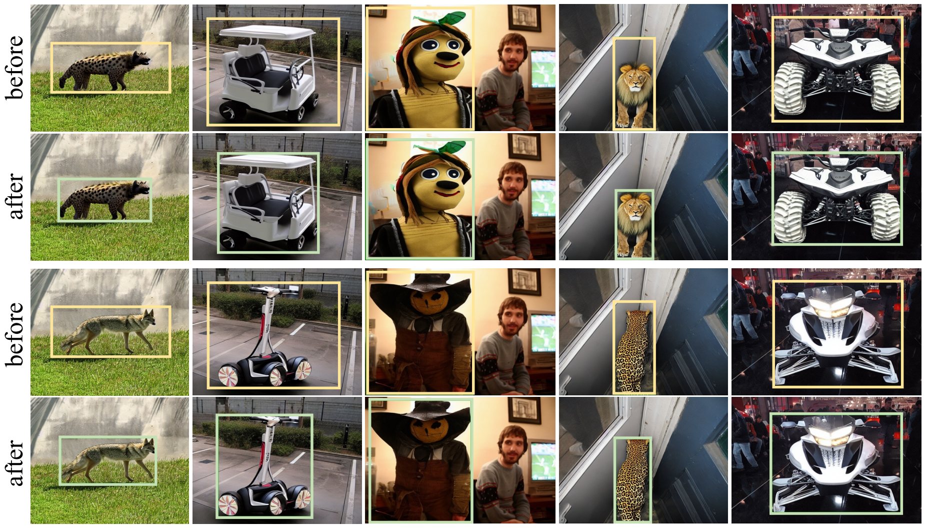

Fig. 2: Edited images are shown with annotations bounding boxes on novel objects. The edited images with novel objects are firstly annotated with editing masks as fuzzy boxes (shown in the lines of "before", with yellow bounding boxes). Then the SAM-based refiner refines the boxes and obtains the accurate boxes (shown in the lines of "after", with green bounding boxes).

in Tab. 3, rather than novel objects, novel attributes tend to be extended descriptive phrases or sentences. It can be seen that handling long descriptive phrases is a greater challenge for subsequent Stable-Diffusion-Inpainting [9], which encourages us to further explore image editing in the future.

1.3 Feature Visualization

We visualize the features based on UMAP [6] in the reduced dimension spaces, as shown in Fig. 1. As mentioned in the main manuscript, our synthetic samples are located in close proximity to the ID samples, resulting in excellent performance. In contrast, although the samples augmented by the baseline method [11] are large in number, they are far away from the ID samples, which is not conducive to optimizing accurate decision boundaries.

2 Synthetic Data

2.1 Imagined Concepts

To complement the details of the synthetic data, we present the imagined concepts used in our synthetic data and associated experiments as shown in Tab. 3. During the experiment implementation, we ensure that all synthetic concepts do not have any overlap with the OOD categories that are used in model evaluations. As illustrated in Tab. 3, although imagined categories are numerous and almost endless, these categories from different models (GPT-4 [1], Llama 2 [10], and WordNet [7]) show great differences, resulting in completely different help for the OOD object detection task (see Tab. 1).

2.2 Extended Visualization of Refined Annotations

The annotation refinement process based on SAM [4] efficiently helps novel objects obtain accurate bounding boxes. We present some qualitative results in Fig. 2. As shown in the figure, when an image containing the novel object is edited, the bounding box inherited from its editing mask often cannot wrap the novel objects tightly. After refining, the obtained new bounding box achieves a superior fit around the novel object well. This will facilitate the extraction of more effective instance-level features with the object detector. The ablation study results in the main manuscript also confirm the effectiveness of the refiner.

2.3 Extended Visualization of High Similarity OOD Samples

The selection of synthetic samples is predicated on their similarities. A large number of experimental results and Fig. 1 indicate the appropriateness of our selection criteria. Here, we extensively visualize high-similarity synthetic images in Fig. 3. These selected images not only maintain consistent contexts but also contain novel objects with annotated bounding boxes. Moreover, they exhibit high cosine similarities with the corresponding initial samples in the latent space, thereby facilitating the optimization of decision boundaries with high precision.

3 Limitations

While our controllable synthetic data leverages superior knowledge within foundation models to achieve excellent performance in OOD object detection, our data synthesis and application strategy is tailored to this specific task. The adaptability of our method across different domains [2, 5, 8] or tasks is still an area for further exploration. This realization motivates our continued pursuit of more robust data synthesis methodologies in the future, aimed at facilitating the completion of more challenges in the real open world.

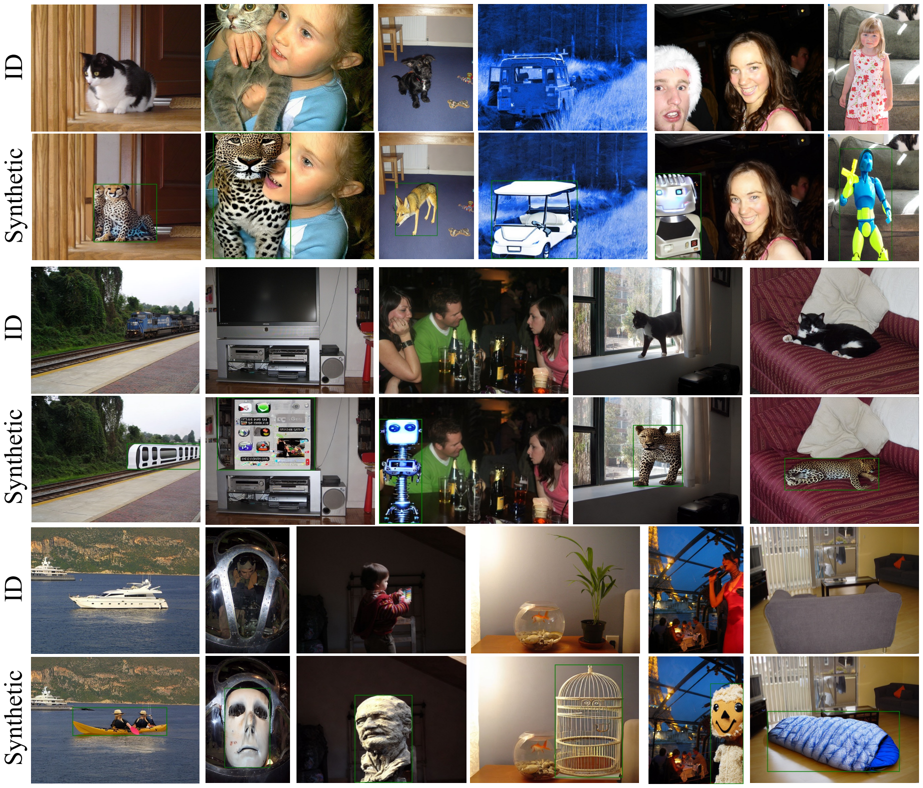

Fig. 3: High similarity ID-Synthetic image pairs are shown in the figure, and edited novel objects are shown with green bounding boxes.

Table 3: Examples of ID categories, OOD categories, and synthetic categories/attributes from different models [1, 7, 10].

ID Categories (PASCAL-VOC, BDD-100K)
couch, bus, dog, cat, boat, car, airplane, horse, dining table, chair, cow, potted plant, motorcycle, bird, sheep, train, bottle, person, tv, bicycle, traffic light, truck, motorcycle, bus, bicycle, rider, train, pedestrian, traffic sign.
OOD Categories (MS-COCO, OpenImages)
book, sink, vase, baseball bat, tennis racket, bed, knife, kite, frisbee, orange, giraffe, skateboard, mouse, remote, wine glass, cup, zebra, fork, fire hydrant, spoon, snowboard, bowl, apple, truck, suitcase, elephant, banana, carrot, handbag, cell phone, stop sign, hair drier, broccoli, hot dog, baseball glove, scissors, oven, skis, parking meter, pizza, clock, toilet, toothbrush, microwave, donut, tie, toaster, backpack, cake, ...
Synthetic Categories (GPT-4)
altar, bonsai tree, monorail, semi-trailer, elk, holstein, flamingo, floor cushion, texel, manikin, suffolk, ram, hot air balloon, beagle, ATV (All-Terrain Vehicle), cinema screen, teapot, island, cargo van, projector, mug, lectern, rollerblader, street lamp, milestone, go-kart, wheelbarrow, locomotive, digital display board, panther, persian, pedestal, tractor, dorper, vicuña, floor pillow, dalmatian, loveseat, sedan, shuttle, robin, dummy, jersey, stallion, merino, herb garden, puppet, action figure, e-reader screen, sculpture, computer screen, swing, statue, chopper, golf cart, aquarium, clydesdale, foal, sledge, angus, billboard, buffalo, counter, doll, window box, beacon, canteen, siamese, cheviot, hammock, ferry, coach, arabian, monitor, ottoman, space shuttle, minivan, mare, bison, labrador, pitcher, mannequin, hearse, zeppelin, terrarium, emergency siren, stool, desk, eagle, robot, scarecrow, scarecrows, alpaca, workbench, convertible, guernsey, equestrian, steer, ...
Synthetic Categories (Llama 2)
dining table-patterned throw blanket, toy bottle, dog bed, sculpture, airplane-themed mug, toy boat, boat-themed mug, hay bale, sheep-shaped pillow, couch-shaped pillow, tv-themed mug, bicycle-patterned throw blanket, couch leg, horse figurine, dog food bowl, car keychain, horse grooming tools, dining table leg, toy chair, dining table-themed mug, couch-patterned throw blanket, dog leash, flight simulator, sheep-patterned throw blanket, sheep wool yarn, stuffed animal (bird), sheep-themed mug, motorcycle helmet, car-themed mug, airplane-shaped pillow, train-shaped pillow, boat-shaped pillow, birdhouse, toy bicycle, boat-patterned throw blanket, chair-themed mug, stuffed animal (sheep), toy bus, stuffed animal (dog), cheese block, train-themed mug, bottle-shaped pillow, doll, bottle-themed mug, horse feed bucket, scratching post, train-patterned throw blanket, drone, bus-shaped pillow, chair leg, toy couch, cat toy, catnip toy, bus ticket, butter dish, airplane-patterned throw blanket, toy motorcycle, motorcycle-themed mug, dog toy, chair-shaped pillow, tv-shaped pillow, ...
Synthetic Categories (WordNet)
heterotroph, strike_out, yenta, spindle_horn, schlepper, secretaryship, sugar_bowl, drone, sir, sled, murphy_bed, looker, dispose, true_cat, cartesian, brace, frame, piggyback, crateful, autogiro, screeching, platform_bed, hooray_henry, mass_meeting, snowmobile, hue_and_cry, coaster_wagon, bike, stall, whirlybird, weather_ship, butter_dish, teaspoonful, direction_finder, tense, cast, wireless, muster, canis_familiaris, doodlebug, horsemeat, drawing_room, pillow_block, boob_tube, lamarckian, senatorship, inhibit, glider, park, she-devil, cadetship, cart, radio_receiver, ostensorium, coney, waveguide, groom, inspectorship, racer, peri, orthopter, bench_hook, veau, electromotive_series, congener, pelmet, welcome_wagon, pep_pill, king, shillyshally, vital_principle, vassal, radio-controlled_aircraft, hosanna, radio_set, doe, warplane, grille, bunk, farm_animal, ...
Synthetic Attributes (GPT-4)
person with exaggerated proportions, person with disproportionate limbs, bird with oversized beak, cat with two tails, airplane with mismatched wings, airplane with overly long nose cone, bicycle with a transparent frame, boat with a transparent hull, couch with a patchwork fabric design, bottle levitating above the ground, chair with disproportionally long legs, boat floating in mid-air, bus with mismatched wheels, ...

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Chen, H., Wang, S., Li, G., Nie, L., Wang, X., Ning, Z.: Distributed orchestration of service function chains for edge intelligence in the industrial internet of things. IEEE Transactions on Industrial Informatics **18**(9), 6244–6254 (2021)
3. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. International journal of computer vision **88**, 303–338 (2010)
4. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
5. Lyu, X., Chang, C., Dai, P., Sun, Y.t., Qi, X.: Total-decom: Decomposed 3d scene reconstruction with minimal interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20860–20869 (2024)
6. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
7. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
8. Ning, Z., Chen, H., Ngai, E.C., Wang, X., Guo, L., Liu, J.: Lightweight imitation learning for real-time cooperative service migration. IEEE Transactions on Mobile Computing **23**(2), 1503–1520 (2023)
9. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
10. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
11. Wilson, S., Fischer, T., Dayoub, F., Miller, D., Sünderhauf, N.: SAFE: Sensitivity-aware features for out-of-distribution object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23565–23576 (2023)
12. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2636–2645 (2020)