Supplement to Objaverse: A Universe of Annotated 3D Objects

A. Instance Segmentation with CP3D

Model. We use the Mask-RCNN [5] model of [1] with a ResNet-50 backbone [6]; no additional changes to their model are made. Instead of a softmax activation, the model uses a Gumbel activation, given by the formula $\eta(q) = \exp(-\exp(-q))$, to transform logits into probabilities. More details about the model and activation can be found in [1].

Training. We take the pretrained ResNet-50 Mask-RCNN checkpoint of [1] and finetune the model for 24 epochs with the CP3D augmentation integrated into the training pipeline. We use a batch size of 64 and a learning rate of 0.002.

Additional Results Here we report detection metrics in addition to the segmentation results reported in the paper in Table 1. Notably, we see an impressive gain of two points on AP for rare categories.

Method	AP	APr	APc	APf
GOL [1]	27.5	19.8	27.2	31.2
GOL + 3DCP	**28.9**	**21.8**	**28.7**	**32.2**

Table 1. **Detection results for bounding box AP category metrics.** APr, APc, and APf measure AP for categories that are rare (appear in 1-10 images), common (appear in 11-100 images), and frequent (appear in $>$100 images), respectively.

B. Open-Vocabulary ObjectNav

Model. The agent's embodiment is a simulated LoCoBot [2]. The action space consists of six actions: MoveAhead, RotateLeft, RotateRight, End, LookUp, and LookDown. Given the excellent exploration capabilities of EmbCLIP [3,7], we opt to keep the same overall architecture, just replacing the learned embedding for target types in prior work by a linear projection of the text branch output of CLIP for the target description, as shown in Fig. 1. Additionally, in order to provide more information about the target and the current visual input, we increase the respective internal representations for each modality from the original 32-D to 256-D. Note that our model does not employ the alternative zero-shot design described in [7], where the target description is not observed by the agent's RNN. Given the scale of Objaverse-LVIS, we can train agents with good generalization following a more standard design.

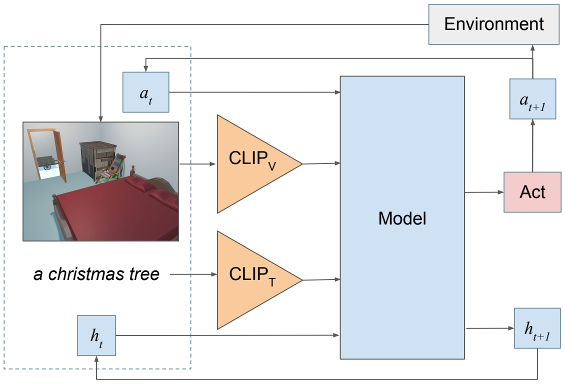


Figure 1. **Open-Vocabulary ObjNav Model overview.** The ObjectNav model (employing an RNN) uses the high-level architecture illustrated here, where it receives features from the visual and target object description encoders, besides previous hidden units and actions as input, and outputs the next action.

Training. For training, we use ProcTHOR to procedurally generate 10,080 houses. Each house has up to three rooms, entirely populated with Objaverse-LVIS assets except for structural components like doors and windows, which are inherited from ProcTHOR [3]. We sample targets corresponding to LVIS categories for which a single instance is present in the scene, resulting in a total of 9,421 unique assets corresponding to 262 categories targeted during training. Training uses DD-PPO [11] and is distributed across 28 GPUs on 7 AWS g4dn.12xlarge machines, with each GPU hosting 360 houses and the subset of Objaverse-LVIS assets populating them. The training hyperparameters are identical to the ones in [3], and the 262 training target categories are listed in Table 2 and Table 3, respectively.

Testing. For testing, we sample 150 episodes for each of 30 target categories, which are a subset of the training target categories. The resulting 4,500 episodes are sampled

Hyperparameter	Value
Discount factor (γ)	0.99
GAE parameter (λ)	0.95
Value loss coefficient	0.5
Entropy loss coefficient	0.01
Clip parameter (ϵ)	0.1
Rollout horizons	32, 64, 128
Rollout timesteps	20
Rollouts per minibatch	1
Learning rate	$3 \cdot 10^{-4}$
Optimizer	Adam [8]
Gradient clip norm	0.5

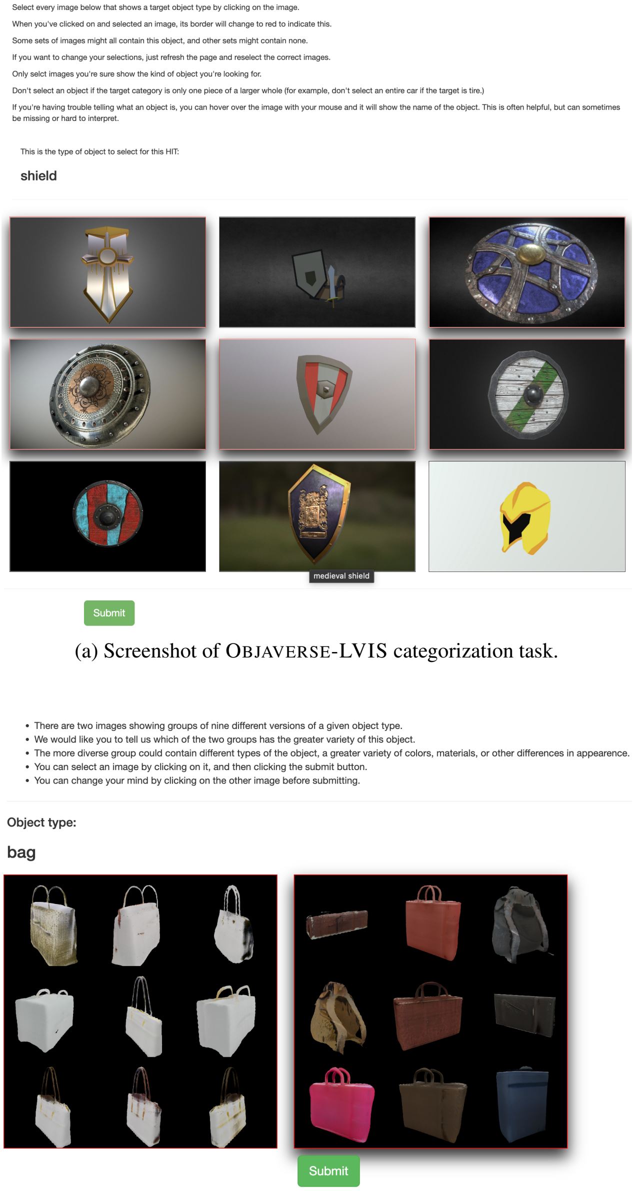
Table 2. **Training hyperparameters for Open-Vocabulary ObjectNav.**

from 151 procedural houses not seen during training. The 30 testing target categories are listed in Table 4. For the results provided in the main paper, the agent is trained for just 18 million simulation steps, but the resulting policy already shows reasonable performance given the variety of targets and scenes. Improved performance can be achieved with extended training (e.g., after approx. 460 million steps, the success rate is 33.0%).

C. Composition

Human subjects data. A portion of the data included in OBJAVERSE is generated by human subjects (*i.e.* crowdworkers recruited through Amazon's Mechanical Turk platform) as outlined in Section 3 and detailed below. The collection process has been reviewed and approved for release by an Institutional Review Board.

Data collection interfaces. Human annotators were used to provide the category labels for OBJAVERSE-LVIS as described in Section 3. This task was accomplished by first creating sets of 500 candidate objects for each LVIS category. These candidate sets included objects visually resembling the target category (as ranked by the CLIP features of their thumbnail images), as well as instances whose metadata contained terms with a high similarity to the target category (as ranked by their GloVe vector similarity [9]). Candidate objects were shown to crowdworkers nine at a time, and they were asked to mark objects that were members of the category, as shown in Figure 2 a. In addition to the visual reference for each object, annotators also had access to the object's name and were encouraged to use this when helpful. Human annotators were also used to rate the relative diversity of of 3D objects generated by models trained using OBJAVERSE and ShapeNet. The user interface and instructions for this task are shown in Figure 2 b. Two sets of nine objects generated by each model were shown with random left-right orientations, and workers were asked to choose the set exhibiting the greater variety in appearance.



Select every image below that shows a target object type by clicking on the image.

When you've clicked on and selected an image, its border will change to red to indicate this.

Some sets of images might all contain this object, and other sets might contain none.

If you want to change your selections, just refresh the page and reselect the correct images.

Only selct images you're sure show the kind of object you're looking for.

Don't select an object if the target category is only one piece of a larger whole (for example, don't select an entire car if the target is tire.)

If you're having trouble telling what an object is, you can hover over the image with your mouse and it will show the name of the object. This is often helpful, but can sometimes be missing or hard to interpret.

This is the type of object to select for this HIT:

shield

(a) Screenshot of OBJAVERSE-LVIS categorization task.

- There are two images showing groups of nine different versions of a given object type.
- We would like you to tell us which of the two groups has the greater variety of this object.
- The more diverse group could contain different types of the object, a greater variety of colors, materials, or other differences in appeerence.
- You can select an image by clicking on it, and then clicking the submit button.
- You can change your mind by clicking on the other image before submitting.

Object type:

bag

(b) Screenshot of relative diversity rating task.

Figure 2. **Data collection interfaces.**

D. Estimating Coverage

We use OpenAI's CLIP ViT-B/32 model to estimate the categorical coverage of the objects in OBJAVERSE. Specifically, for each object, we compute the CLIP image embedding from the thumbnail and the cosine similarity between an text embedding of each WordNet entity [4]. The entity

Bible, Christmas tree, Rollerblade, alligator, ambulance, amplifier, arctic (type of shoe), armor, banner, barbell, barrel, barrow, baseball bat, basketball, bat (animal), bath mat, beachball, bear, bed, beetle, bench, beret, bicycle, binder, binoculars, bird, blackberry, bookcase, boot, bottle, bowling ball, bullhorn, bunk bed, bus (vehicle), butterfly, cab (taxi), cabinet, canoe, cape, car (automobile), card, cardigan, carnation, cart, cassette, cat, chair, chaise longue, chicken (animal), clothes hamper, coatrack, coffee table, cone, convertible (automobile), cornice, cow, cowboy hat, crab (animal), crate, crossbar, cube, cylinder, deck chair, deer, desk, dinghy, dirt bike, dog, dollhouse, doormat, dove, drawer, dresser, duckling, dumbbell, dumpster, easel, elephant, elk, fan, ferret, file cabinet, fireplace, fireplug, fishing rod, flag, flagpole, flamingo, flip-flop (sandal), flipper (footwear), foal, football (American), footstool, forklift, frog, futon, garbage, gargoyle, giant panda, giraffe, golf club, golfcart, gondola (boat), goose, gorilla, gravestone, grill, grizzly, grocery bag, guitar, handcart, hat, heater, hockey stick, hog, horse, horse carriage, jeep, kayak, keg, kennel, kitchen table, kitten, knee pad, ladder, ladybug, lamb (animal), lamp, lamppost, lawn mower, legging (clothing), lion, lizard, locker, log, loveseat, machine gun, mailbox (at home), manhole, mascot, mast, milk can, minivan, monkey, mop, motor, motor scooter, motor vehicle, motorcycle, mushroom, music stool, nut, ostrich, owl, pajamas, parasail (sports), parka, penguin, person, pet, pew (church bench), piano, pickup truck, pinecone, ping-pong ball, playpen, pole, polo shirt, pony, pool table, power shovel, propeller, pug-dog, pumpkin, rabbit, radiator, raincoat, ram (animal), rat, recliner, refrigerator, rhinoceros, rifle, road map, rocking chair, router (computer equipment), runner (carpet), saddle (on an animal), saddle blanket, saddlebag, sandal (type of shoe), scarecrow, scarf, sculpture, seabird, shark, shepherd dog, shield, shirt, shoe, sink, skateboard, ski parka, skullcap, snake, snowmobile, soccer ball, sock, sofa, sofa bed, solar array, sparkler (fireworks), speaker (stero equipment), spear, spider, sportswear, statue (sculpture), step stool, stepladder, stool, subwoofer, sugarcane (plant), suit (clothing), suitcase, sunhat, surfboard, sweat pants, sweater, swimsuit, table, tape measure, tarp, telephone pole, television camera, tennis ball, tennis racket, tights (clothing), toolbox, tote bag, towel, trailer truck, trampoline, trash can, tricycle, trousers, truck, trunk, turtle, tux, underdrawers, vacuum cleaner, vending machine, vest, wagon wheel, water ski, watering can, wet suit, wheel, window box (for plants), wok, wolf, and *wooden leg.*

Table 3. **Training target types for Open-Vocabulary ObjectNav.**

Christmas tree, bed, bench, blackberry, chair, chicken (animal), dog, easel, elk, fireplug, forklift, garbage, gargoyle, guitar, mascot, motor, penguin, pony, pool table, radiator, rifle, scarf, sock, speaker (stero equipment), sportswear, sweat pants, trash can, trunk, wet suit, and *wheel.*

Table 4. **Testing target types for Open-Vocabulary ObjectNav.**

with the maximum cosine similarity is then assigned as the object's entity. The WordNet entities are textually encoded in the form, "a {entity} is a {definition}", which is loosely inspired by CuPL [10]. For instance, we might have "a *bat* is a nocturnal mouselike mammal with forelimbs modified to form membranous wings and anatomical adaptations for echolocation by which they navigate" or "a *bat* is a club used for hitting a ball in various games". Computing the nearest WordNet entity for each object gave us an estimated coverage of 20.8K entities.

References

[1] Konstantinos Panagiotis Alexandridis, Jiankang Deng, Anh Nguyen, and Shan Luo. Long-tailed instance segmentation using gumbel optimized loss. *arXiv preprint arXiv:2207.10936*, 2022. 1

[2] Carnegie Mellon University. Locobot: an open source low cost robot. http://www.locobot.org/. 1

[3] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, et al. Procthor: Large-scale embodied ai using procedural generation. *Conference on Neural Information Processing Systems*, 2022. 1

[4] Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010. 2

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[7] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14809–14818, 2022. 1

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 2

[9] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[10] Sarah Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*, 2022. 3

[11] Erik Wijmans, Abhishek Kadian, Ari S. Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *ICLR*, 2020. 1