BEHAVIOR Vision Suite: Customizable Dataset Generation via Simulation

Yunhao Ge* Yihe Tang* Jiashu Xu* Cem Gokmen*
Chengshu Li Wensi Ai Benjamin Jose Martinez Arman Aydin Mona Anvari
Ayush K Chakravarthy Hong-Xing Yu Josiah Wong Sanjana Srivastava Sharon Lee
Shengxin Zha Laurent Itti Yunzhu Li Roberto Martín-Martín Miao Liu Pengchuan Zhang
Ruohan Zhang Li Fei-Fei Jiajun Wu
Stanford University Harvard University Meta AI

Please do not share externally

Abstract

The systematic evaluation and understanding of computer vision models under varying conditions require large amounts of data with comprehensive and customized labels, which real-world vision datasets rarely satisfy. While current synthetic data generators offer a promising alternative particularly for embodied AI tasks, they often fall short for computer vision tasks due to low asset and rendering quality, limited diversity, and unrealistic physical properties. We introduce the BEHAVIOR Vision Suite (BVS), a set of tools and assets to generate fully customized synthetic data for systematic evaluation of computer vision models, based on the newly developed embodied AI environment, BEHAVIOR-1K. BVS supports a large number of adjustable parameters at the scene level (e.g., lighting, object placement), the object level (e.g., joint configuration, attributes such as "filled" and "folded"), and the camera level (e.g., field of view, focal length). Researchers can arbitrarily vary these parameters during data generation to perform controlled experiments. We showcase three example application scenarios: systematically evaluating the robustness of models across different continuous axes of domain shift, evaluating scene understanding models on the same set of images, and training and evaluating simulation-to-real transfer for a novel vision task: unary and binary state prediction. All code and data will be made public.

1. Introduction

Large-scale datasets and benchmarks have fueled computer vision research in the past decade [2, 9, 10, 15–18, 24, 32, 36, 45, 53]. Driven by these datasets and benchmarks, thousands of models and algorithms tackling different perception challenges are being proposed every year, on the topics of object detection [60], segmentation [22], action recognition [49], video understanding [30] and beyond. Despite their success, these real-world datasets have a few inherent limitations. First, the ground-truth object/pixel-level labels are either prohibitively expensive to acquire (e.g. segmentation masks) [31] or inaccurate/noisy (e.g. depth sensing) [38]. As a result, each real dataset often only offers a limited set of labels, thus hindering the development and evaluation of computer vision models that perform a wide range of perception tasks on the same image inputs. Even if annotation is free and accurate, real-world datasets are bounded by the availability of source images. For example, images of rare events such as traffic accidents or low-light conditions might be difficult to acquire from the Internet or real-world sensors. Lastly, once collected, these real-world datasets have a fixed data distribution and cannot be easily customized. This last limitation makes it almost impossible for researchers to conduct customized experiments and forces them to adopt the experimental setup decided by dataset creators, often leading models to overfit to the datasets and the eventual obsoletion of the benchmarks.

To circumvent this limitation, researchers and practitioners have come up with a variety of ways to generate synthetic datasets that complement the real ones. In the realm of indoor scene understanding, 3D reconstruction datasets [4, 41, 52] provide a promising avenue to generate source images from arbitrary viewpoints and free (geometric) annotations. Due to the imperfect nature of 3D reconstruction techniques, however, the rendered images are not very realistic. With each entire scene as a static mesh, these datasets also offer very limited customizability other than camera trajectories. Recent synthetic indoor datasets (often designed by 3D artists) [11, 28, 29, 43] come to the rescue because they not only provide free annotations (both geometric and semantic) but also allow for object layout reconfiguration (since the objects are usually independent CAD models). However, these datasets do not guarantee physical plausibility (object penetration and levitation happen often)

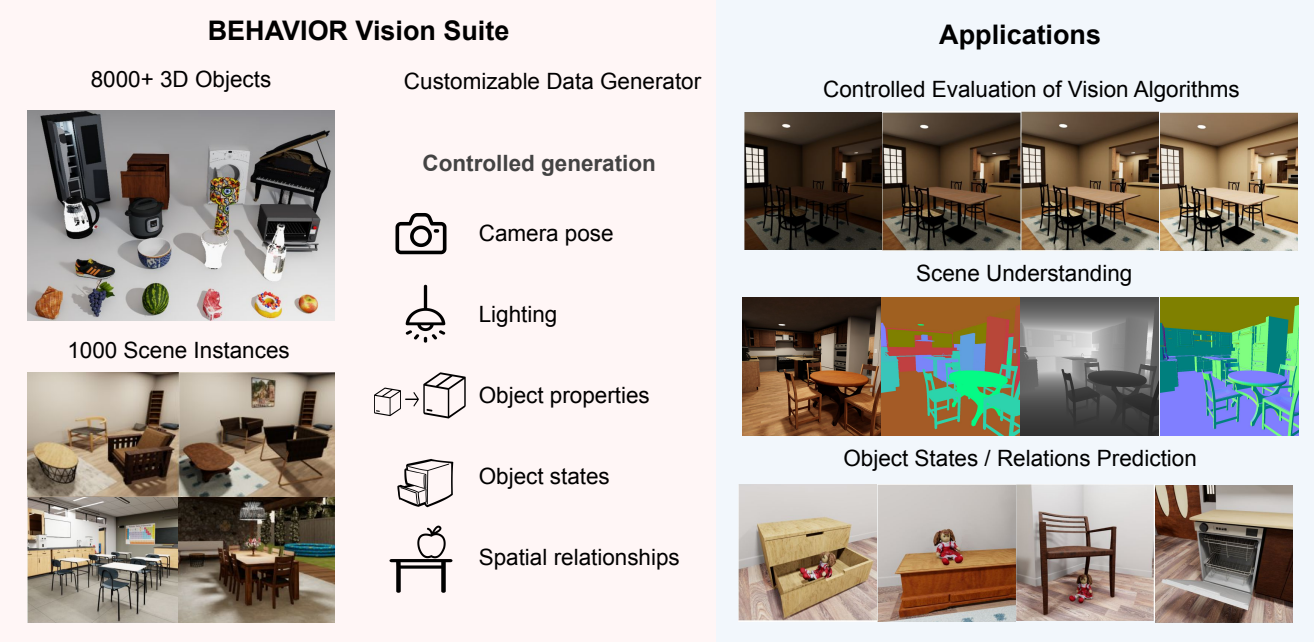


Figure 1. Overview of BEHAVIOR Vision Suite (BVS), our proposed toolkit for computer vision research. BVS builds upon extended object assets and scene instances from BEHAVIOR-1K [26], and provides a customizable data generator that allows users to generate photorealistic, physically plausible labeled data in a controlled manner. We demonstrate BVS with three representative applications.

and do not provide any customization capability beyond changing object poses. 3D simulators [7, 12, 23, 25, 44, 48], on the other hand, guarantee physical plausibility because of their underlying physics engines. They also allow users to customize the joint configuration of articulated objects and even more advanced object states such as "cooked" or "sliced" [23, 25]. Yet these 3D simulators generally cater to Embodied AI and robotics researchers, and as a result, they lack photorealism compared to the synthetic datasets mentioned before (usually due to speed constraints) and they don't provide off-the-shelf tooling to generate customized image/video datasets for computer vision researchers.

To overcome the aforementioned challenges, we propose BEHAVIOR Vision Suite (BVS), a customizable data generation tool that allows for systematic evaluation and understanding of computer vision models (see Fig. 1 for an overview). First, we expand the 3D asset library in BEHAVIOR-1K [26], focusing on enhancing both object diversity and scene variety as well as adding features to increase value of the assets for vision tasks. Then, we introduce Customizable Dataset Generator, which leverages the simulator from the BEHAVIOR-1K benchmark [26, 47] to generate custom vision datasets. We build a versatile and customizable toolbox to generate high-quality synthetic data for systematic model evaluation and understanding.

In a nutshell, BEHAVIOR Vision Suite has the following unique combination of desirable features:

1. offers exhaustive image/object/pixel-level labels for free (scene graph, point cloud, depth, segmentation, etc)
2. covers a wide variety of indoor scenes and objects (8K+ objects, 1K scene instances, fluid, soft bodies, etc)
3. guarantees high physical plausibility and photorealism
4. provides maximum customization capability in terms of object models, poses, joint configurations, semantic states, lighting, texture, material, camera setting, etc.
5. includes easy-to-use tooling to generate customized data for new use cases.

To demonstrate the usefulness of BVS, we showcase three example applications: 1) parametrically evaluating model robustness across different conditions such as lighting and occlusion, 2) evaluating different types of representative computer vision models on the same set of images, and 3) training and evaluating sim2real transfer for object states and relations prediction. By showing these three examples, we hope that BVS can unlock more possibilities for the computer vision community.

2. Related works

In this section, we will compare BEHAVIOR Vision Suite against other real RGB-D datasets, 3D reconstruction datasets, synthetic datasets and 3D simulators in terms of customizability and visual quality (see Tab. 1).

Dataset Category	Customizability				Visual Quality
	Camera View	Obj Pose	Obj State	CV Toolkit	
Real RGB-D Datasets	✗	✗	✗	N/A	Real
3D Reconstruction Datasets	✓	✗	✗	N/A	Medium
Synthetic Datasets	✓	✓	✗	~	High
3D Simulators	✓	✓	✓	✗	Low
BEHAVIOR Vision Suite (Ours)	✓	✓	✓	✓	High

Table 1. Comparison of real and different types of synthetic datasets to BEHAVIOR Vision Suite. Camera View indicates whether images can be rendered from arbitrary viewing angles. Obj Pose indicates whether object layout can be modified. Obj State indicates whether object physical states (e.g. open/close, folded) and semantic states (cooked, soaked, etc) can be modified. CV toolkit indicates whether utility functions are provided to sample camera poses that satisfy certain constraints (those that capture half-open kitchen cabinets filled with grocery items, for instance). Visual Quality indicates how photorealistic the images are. For more detailed comparison, see Appendix.

2.1. Real Indoor Scene RGB-D Datasets

RGB-D image datasets of real indoor scenes [1, 5, 38, 46, 56] have driven advancement in 3D perception and holistic scene understanding. Recent works include ARKitScenes [1] and ScanNet++ [56] that provide dense semantic and 3D annotations. While these real datasets capture image distribution from the real world, they are expensive to annotate and inherently static: users are unable to generate images from new camera views, acquire new types of annotations, or modify the scenes in any way. Our work is thus complementary, offering users a fully customizable generator of photorealistic synthetic data.

2.2. 3D Reconstruction Datasets

3D reconstruction datasets like Gibson and Matterport [4, 52] allow rendering of novel views. HM3DSem [41, 55] scales up to 1,000 scenes, improves the reconstruction quality, and provides more accurate dense semantic annotations. While these datasets have tremendously benefited the embodied navigation community, their application to computer vision is limited. Each scene is a single 3D mesh and hence prohibits further customization such as object layout. Furthermore, the visual quality of novel view rendering highly depends on reconstruction fidelity—artifacts like glasses still exist. Semantic label acquisition is also very expensive. Our work, in contrast, is capable of generating images with customized object layouts, consistent visual quality, together with free, comprehensive labels.

2.3. Synthetic Datasets

Synthetic datasets offer an alternative approach that saves the cost of semantic labeling. Hypersim [43], 3D-FUTURE [11] and InteriorNet [28] render virtual images from artist-created scenes where objects are independent models. OpenRooms [29] generates scene layouts from real scans and provides configurable rendering options. Objaverse [6, 8] also provides a number of interior scenes along with large-scale 3D object models. However, despite being photorealistic, these datasets do not guarantee physical plausibility: objects often penetrate with each other or slightly levitate in the air. Also, most, if not all, objects are non-articulated and don't support any semantic state changes. Our work, on the other hand, has more customization capability (e.g. joint configuration, semantic states such as "cooked" or "filled") and physical plausibility.

2.4. 3D Simulators

A large number of 3D simulators with physical realism have been developed recently. iGibson [25, 44] and Habitat 2.0 [48] introduce reconfigurable indoor scenes with articulated assets, while the former also highlights their support for extended object states such as wetness level. ThreeDWorld [12] emphasizes physical interaction modeling, especially with non-rigid objects. ProcTHOR [7] automates large-scale generation of semantically plausible virtual environments. Since these 3D simulators cater towards the Embodied AI and robotics community, their visual quality is limited. In contrast, we leverage a new 3D simulator called OmniGibson that has been shown to be significantly more photorealistic than all the ones mentioned before, according to a human survey conducted in [26], making our work a better candidate for computer vision research. Furthermore, our work offers many utility functions that allow users to easily generate near-infinite images that suit their specific needs, which most existing 3D simulators lack.

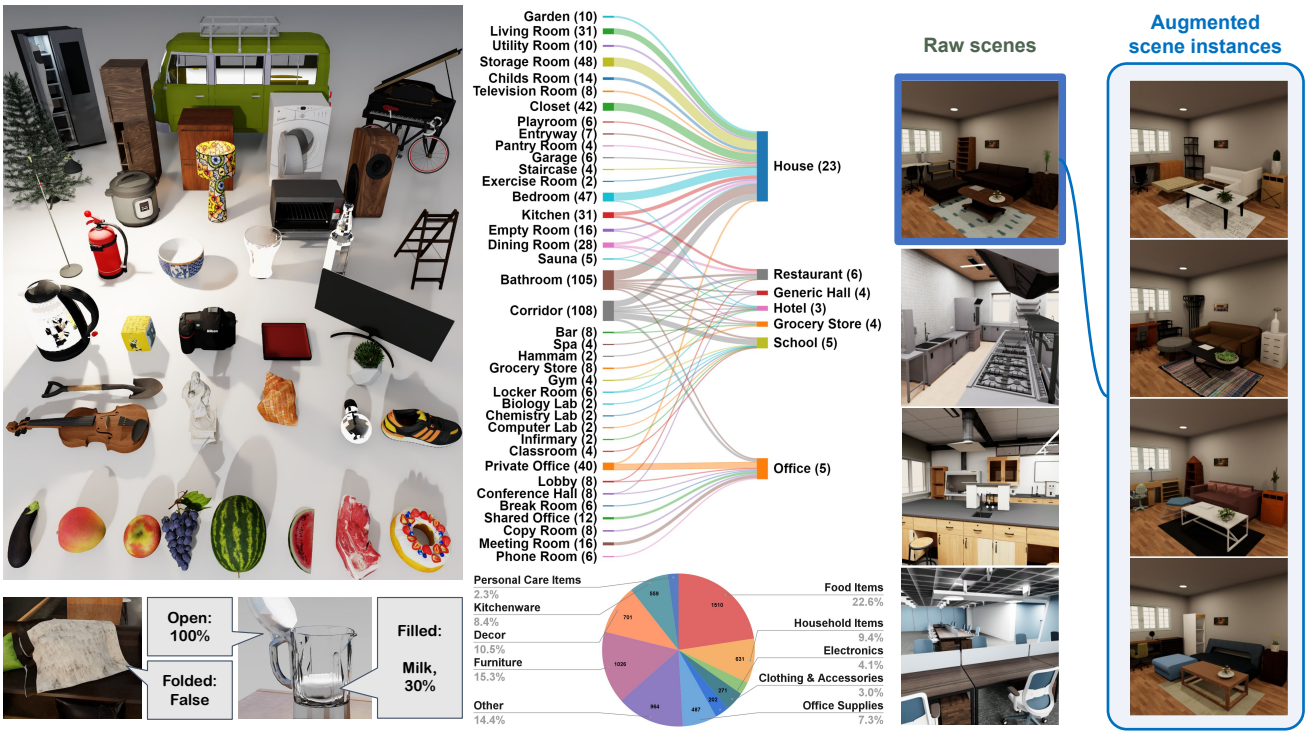
3. BEHAVIOR Vision Suite

BEHAVIOR Vision Suite is composed of two main components: the Extended BEHAVIOR-1K Assets and the Customizable Dataset Generator. The assets serve as the foundation, while the generator uses the assets to generate vision datasets that suit downstream tasks of interest.

3.1. Extended BEHAVIOR-1K Assets

The Extended BEHAVIOR-1K Assets are a collection of 8,841 object models and 1,000 scene instances that are variations of 51 artist-designed raw scenes. Out of the 8,841 objects, 2,156 are structural elements such as walls, floors, ceilings, and 6,685 are non-structural objects belonging to 1,937 categories. These categories are of great variety such as food, tools, electronics, clothing, office supplies, and others. The distribution of the objects into these semantic categories can be seen in Fig. 2. The 51 raw scenes consist of houses (23), offices (5), restaurants (6), grocery stores (4), hotels (3), schools (5), and generic halls (4) as well as a simulated twin of a mock apartment at our research lab.

This collection of assets is the result of a year-long effort to extend the BEHAVIOR-1K [26] assets to increase their usability and value for computer vision applications.



(a) Examples of 3D objects and the semantic properties they support

(b) Distribution of scenes, room types and objects

(c) Examples of raw scenes and augmented scene instances

Figure 2. Overview of Extended BEHAVIOR-1K Assets: Covering a wide range of object categories and scene types, our 3D assets have high visual and physical fidelity, and rich annotations of semantic properties, allowing us to generate 1,000+ realistic scene configurations.

In terms of quantity, we increased the object count from 5,215 to 8,841 through 1) acquisition of more everyday objects, 2) segmentation of building structure into individual objects (e.g. walls are segmented into linear components to make 3D bounding box labels more useful), 3) procedural generation of sliced fruits and vegetables. We also procedurally generate 1,000 diverse scene instances from the original 51 raw scenes by varying the object models for furniture and inserting additional everyday objects into the scenes.

To improve physical realism, we significantly improved the collision mesh quality by first applying V-HACD [34] and CoACD [51] with different parameters and manually selecting the best option, balancing physical accuracy, affordance preservation and simulation efficiency. For over 2,000 objects, this pipeline failed to generate satisfying candidates, so we manually designed their collision meshes.

In terms of lighting, we annotate realistic light sources on objects like lamps and ceilings so that the scene is lit up the same way as it will be in the real world. And in terms of semantic property annotation, we further annotate appropriate fillable volumes for containers (e.g. cups, pots) and fluid source/sink locations (e.g. faucets, drains, sprayers) so that we can spawn fluids in the scene realistically. Scene objects were annotated as non-randomizable when necessary, e.g. when they physically support other objects. Similarly, clutter objects in the scenes were annotated as such, allowing them to be removed and replaced with alternative clutter.

Altogether, we designed the assets to form a strong basis for custom data generation (discussed in the next section), with a functional organization that allows accurate object randomization, and the annotations to provide a large number of modifiable parameters at both the object and scene levels.

3.2. Customizable Dataset Generator

The Customizable Dataset Generator is the software component of the BEHAVIOR Vision Suite designed to generate synthetic datasets with specific characteristics. Built on OmniGibson [26], it leverages NVIDIA Omniverse's photorealistic, real-time ray-tracing renderer and OmniGibson's procedural sampling functions for object states to generate custom images and videos that satisfy arbitrary requirements. The produced datasets include rich, comprehensive annotations (segmentation masks, 2D/3D bounding boxes, depth, surface normals, flows, point clouds) for free. More importantly, users take full control of the dataset generation process by configuring specific scenes, objects, states, camera angles, and lighting conditions, while physical plausibility is guaranteed by the underlying physics engine.

3.2.1 Capabilities

At the core of the generator are its generative capabilities:

- **Scene Object Randomization:** The generator can swap the objects in a particular scene with other objects in the same category. In the assets, the objects are organized into categories that consist of objects that share similar visual and affordance characteristics. By randomizing the objects, we can drastically change the scene appearance while keeping the semantic realism of object layout.
- **Physically Realistic Pose Generation:** The generator can procedurally change the physical states of the objects to satisfy certain predicates. This includes 1) placing objects with respect to other objects in the scene in a certain way (e.g. inside, on top of, under), 2) opening articulated objects, 3) filling containers with fluids, and 4) folding/unfolding pieces of cloth. The generator can generate various valid configurations for the same predicates and ensure physical plausibility.
- **Predicate-Based Rich Labelling:** Beyond providing the usual set of labels (semantic & instance segmentation, bounding boxes, surface normals, depth, etc.), the generator can also label unary predicates for an object (e.g. whether an articulated object is open, or an appliance is toggled on), binary predicates between two objects (e.g. whether or not an object is touching, on top of, next to, etc. another object), binary predicates between an object and a substance (e.g. is an object filled/covered/soaked with a substance), as well as continuous labels (joint openness fraction for articulated objects, filledness fraction for containers, current temperature, etc).
- **Camera Trajectory Generation:** To go from 3D scenes to 2D images, it is necessary to render from a camera, and the placement of such a camera in a 3D scene is challenging: the camera must point at an interesting subject and not be too occluded by any object. The generator uses occupancy grids and hand-crafted heuristics to generate not only static camera poses that satisfy these constraints, for use in image models, but also physically plausible camera trajectories for video/scene understanding models.
- **Configurable Rendering:** The generator also provides an easy API to manipulate rendering parameters, such as lighting and camera intrinsics like aperture and FOV.

3.2.2 Dataset Generation Process

To generate a BVS dataset, we repeat the following steps:

1. **Scene Sampling:** We select one of the 51 raw scenes from the user-configured scene category (say, an office).
2. **Object Randomization:** We randomize the scene objects with similar objects in the same category.
3. **Object Insertion:** We decide what additional objects need to be added into the scene, again depending on the user configuration. We place the objects into the scene using the pose generation capabilities, based on requirements specified by the user. This might include cluttering certain areas (e.g. filling fridges with perishables) or individually manipulating certain objects' states (making a cabinet open or a table covered with water) for downstream predicate prediction.
4. **Updating Camera Pose and Rendering Environment:** We then generate a camera pose (or a sequence of poses as a camera trajectory) as well as randomizing the scene's lighting parameters and the camera's intrinsics based on the user's specification.
5. **Rendering with Ground Truth Labels:** We then render an image (or a sequence of images) and record it alongside all relevant labels requested by the user, including additional modalities (depth/segmentation/etc.), bounding boxes, and predicate and object state values.

Axis	#Scenes	#Video clips
Articulation	17	237
Lighting	16	441
Visibility	14	211
Zoom	9	215
Pitch	16	268

Table 2. We generate up to 200-500 short video clips with diverse scene configurations for parametric evaluation (§4.1). Each video clip varies along one continuous axis with respect to a single target object. On average, each video has 300 frames.

4. Applications and Experiments

We present three applications and their corresponding experiments to demonstrate the utility of BVS: first, systematically evaluating model robustness against various continuous domain shifts like lighting condition (§4.1); second, assessing various scene understanding models using a consistent set of images with comprehensive annotations (§4.2); and third, training and testing the efficacy of simulation-to-real transfer for a new vision task, specifically focusing on object states and relations prediction (§4.3).

4.1. Parametric Model Evaluation

Parametric model evaluation is essential for developing and understanding perception models, as it systematically assesses model performance against various domain shifts.

Task design and dataset generation. We concentrate on five critical parameters that significantly impact model performance but are challenging to rigorously control in real-world datasets: object **articulation**, **lighting**, object **visibility**, camera **zoom** and camera **pitch**. For each parameter, we vary along a continuous axis and evaluate our baseline models along the way. For instance, object visibility varies from the target object being fully occluded to fully visible.

We generate 200 to 500 videos for each axis, featuring diverse target objects from our over 8,000 3D assets. Each video includes a target object with changes focused on a single parameter under examination. Fig. 3 shows exam-

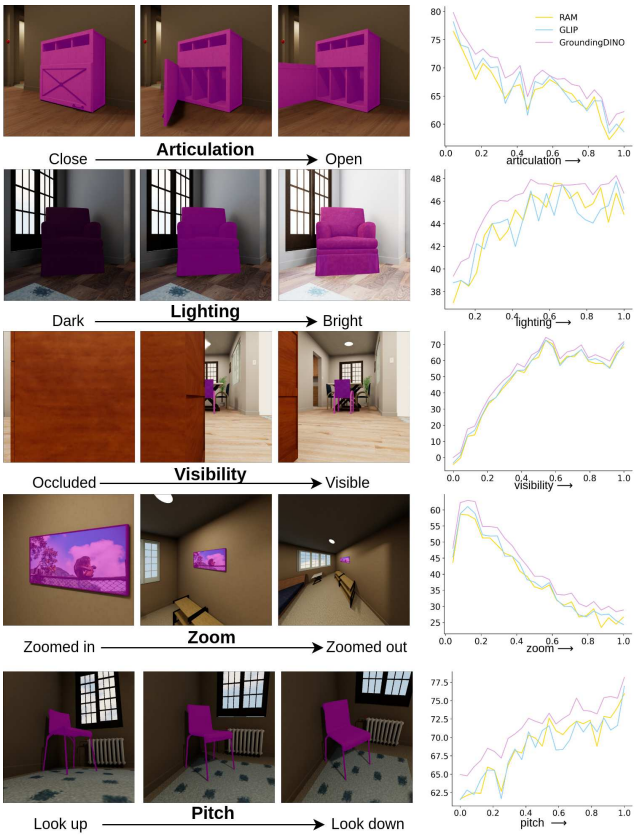


Figure 3. Parametric evaluation of object detection models on five example video clips. Selected frames from the clips are shown on the left, with the target object highlighted in magenta. Average Precision (AP) for our baseline models in §4.2 are plotted on the right. Since BVS allows for full customization of scene layout and camera viewpoints, we can systematically evaluate model robustness to changes in object articulation, lighting conditions, visibility, zoom (object proximity), and pitch (object pose). As we can see, current SOTA models are far from robust to these axes of variation, and we encourage researchers that develop new vision models to use BVS for debugging and parametric evaluation.

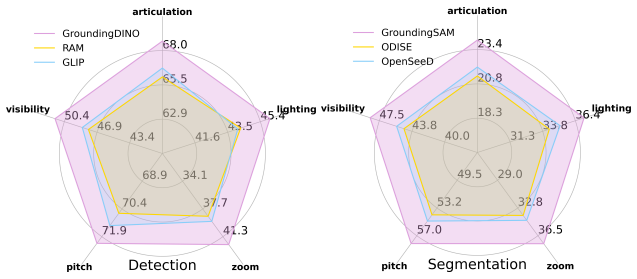


Figure 4. Mean performance of open-vocab object detection and segmentation models across five axes. The larger the colored envelope is for a model, the more robust it is. With the help of BVS, new vision models can be systematically tested for their robustness along these five dimensions and beyond: our users can easily add new axes of domain shift with only a few lines of code.

ples of the target objects with variations in each parameter. (Tab. 2). We control the remaining aspects of the environment and systematically synthesize images while varying the main parameter of interest alone.

Baselines and metrics. We conduct experiments on two representative object-centric vision tasks: *open-vocabulary detection* and *open-vocabulary segmentation*. We believe models developed for these tasks might be sensitive to the object-centric domain shift that we inject. For baselines, we consider the current SOTA models on real datasets: GLIP [27], RAM [58] and Grounding DINO [33] for detection and ODISE [57], OpenSeeD [54] and Grounding SAM [20] for segmentation.

Results and analysis. In Fig. 3 and Fig. 4 we show example images when varying each parameter as well as respective detection Average Precision (AP) performance. To measure the model's ability to recognize the target object (highlighted in magenta), we compute AP solely for the target object as the single ground truth. The following are more detailed analysis of the results.

- **Articulation** varies the joint angles of the articulated target object, ranging from fully closed to fully open. Examples include the opening and closing of drawers, refrigerators, and doors, as well as the folding and unfolding of laptops, etc. Interestingly, we observe a negative correlation between model performance and the degree of articulation. This trend might be attributed to the fact that in existing benchmarks, articulated objects are predominantly depicted in a closed state (e.g. washing machines and microwaves). Consequently, the models are less exposed to scenarios with open articulated objects, leading to decreased performance.
- **Lighting** varies the global illumination of the environment, ranging from being dark to being bright. We observe an increasing trend in model performance until a brightness level of 0.5, indicating that while current models suffer from low light conditions, their performance saturates once the brightness level surpasses a certian threshold.
- **Visibility** varies the visibility of the target object, ranging from fully occluded to fully visible. Visibility is computed as the ratio of the target object's visible pixels over its total pixels. We observe that the model performance quickly degrades when visibility goes below 0.5, leaving a large room for improvements for future models.
- **Zoom** varies how zoomed-in the camera is facing the target object, ranging from very zoomed-in to very zoomed-out. When the view is very zoomed in, with the entire image occupied by a partial view of the target object, the model performance is poor. This suggests that models rely on surrounding semantic context for detection. As the target object is increasingly zoomed out, it also becomes more challenging to detect due to its reduced size. As we expect, the peak model performance lies somewhere in the middle.
- **Pitch** varies the pitch angle of the camera facing the

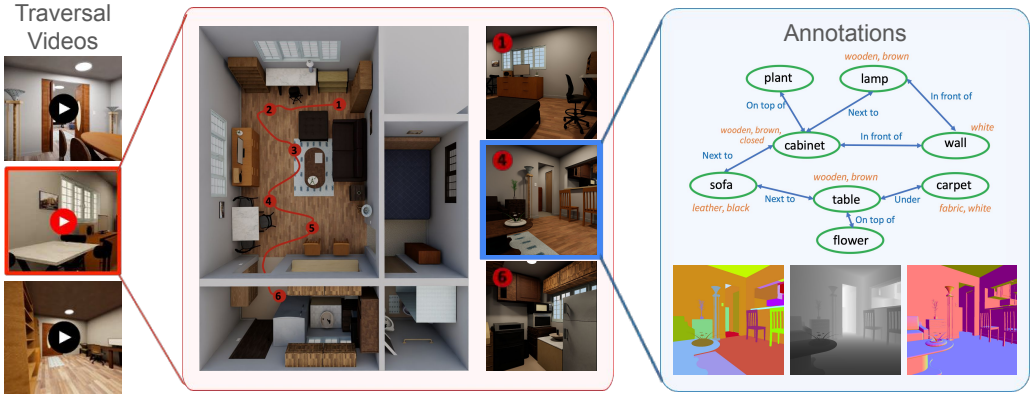


Figure 5. Holistic Scene Understanding Dataset. We generate 10,000 videos across 1,000 scene instance, each scene instance with 10 different camera trajectories. For each image, BVS generates a wide variety of labels (scene graphs, segmentation masks, depth, etc) shown on the right. On average, each video is 1 minute long with 3,000+ frames.

Open-vocab Detection	AP ↑	AP_{small} ↑	AP_{medium} ↑	AP_{large} ↑
GLIP [27]	41.4	7.0	27.5	61.8
RAM [58]	41.3	6.4	27.8	63.9
Grounding DINO [33]	**44.7**	**11.9**	**31.2**	**66.3**

Open-vocab Segmentation	AP ↑	AP_{small} ↑	AP_{medium} ↑	AP_{large} ↑
ODISE [54]	57.1	41.0	53.2	65.0
OpenSeeD [57]	57.3	42.0	54.1	64.8
Grounding SAM [20]	**59.2**	**42.9**	**54.4**	**65.1**

Depth Estimation	RMS ↓	AbsRel ↓	Log10 ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
DPT [42]	0.66	0.14	0.05	0.09	0.15	0.20
NVDS [50]	0.58	**0.13**	**0.04**	0.10	0.15	0.21
iDisc [39]	**0.49**	**0.13**	**0.04**	**0.12**	**0.19**	**0.22**

Point Cloud Reconstruction	Completion Ratio ↑	Completion ↑	Accuracy ↓
GradSLAM [21]	50.0	**14.8**	29.8
NICE-SLAM [59]	**66.3**	12.0	**23.5**

Table 3. A comprehensive evaluation of SOTA models on four vision tasks. Our synthetic dataset can be a faithful proxy for real datasets as the relative performance between different models closely correlates to that of the real datasets.

target object, ranging from looking up to looking down. Our results indicate that the models are not robust to seemingly benign changes in camera viewpoint and tend to perform better if the camera looks down at the target objects. One potential explanation is that in large-scale real datasets, where the models are trained on, it's more common for objects to be slightly below the camera.

To summarize, we observe significant performance variances across three models on all five axes, indicating the lack of robustness of the current SOTA models on extreme or out-of-distribution test environments. By generating large-scale synthetic datasets with controlled variability, BVS provides a unique and powerful test bed to evaluate model performance. Furthermore, our findings align with the observations in §4.2 — relative performance across different models is generally consistent across five axes.

4.2. Holistic Scene Understanding

One of the major advantages of synthetic datasets including BVS is that they offer various types of labels (segmentation masks, depth maps, bounding boxes) for the same sets of input images. We believe this feature can fuel the development of versatile vision models in the future that can perform multiple perception tasks at the same time. Since such models are not currently available, we instead evaluate the current SOTA methods on a subset of the tasks that BVS supports (see below). This will also serve as a validation of the photorealism of our datasets, i.e. models trained on real datasets should perform reasonably without fine-tuning.

Task design and dataset generation. Equipped with BVS's powerful generator (see §3.2), we generate an extensive dataset of 10,000 videos in 1,000 scene instances with per-frame ground truth annotations in multiple modalities. Fig. 5 shows an overview of the generated dataset.

Baselines and metrics. In Tab. 3 we assess 11 models across four tasks. Specifically, we consider *Detection* and *Segmentation* tasks, both in the challenging open vocabulary setting [33, 57]. We also evaluate *Depth Estimation* and *Point Cloud Reconstruction* tasks. Standard metrics are used for all tasks.

Results and analysis. We summarize all our evaluation results in Tab. 3. We observe that the relative performance of these models on our synthetic dataset has high correlation with that on the real datasets such as MS COCO [31] or NYUv2 [38], indicating that our generated synthetic datasets can be a faithful proxy for real datasets.

In summary, we provide a comprehensive benchmark to score and understand a wide range of existing models for each of the four tasks on exactly the same images. While the majority of current vision models focus on single output modality, we hope BEHAVIOR Vision Suite could moti-

Method	Precision	Recall	F1
Zero-shot CLIP	0.293	0.282	0.271
Ours	**0.863**	**0.817**	**0.839**

Table 4. Classification results on the real test set. Task-specific training on syntactic data boosts performance on real images.

Test on		Open	Close	Ontop	Inside	Under	Avg
Synthetic	Precision	0.962	0.897	0.947	0.989	0.874	0.932
	Recall	0.822	0.978	0.913	0.995	0.949	0.929
Real	Precision	0.943	0.958	0.545	0.906	0.948	0.863
	Recall	0.757	0.915	0.913	0.776	0.703	0.817

Table 5. Classification results on held-out synthetic eval set and real test set for our method adapted from [37].

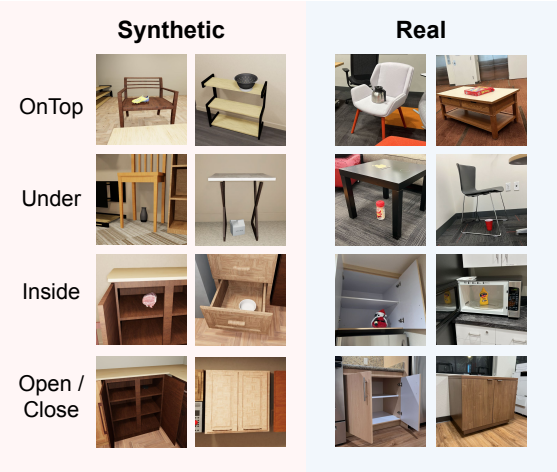


Figure 6. Sample images of each class from our generated synthetic and collected real datasets.

vate researchers and practitioners to develop versatile models that concurrently predict multiple modalities in the future, where our benchmarking results of single-task SOTA methods in this section could serve as a useful reference.

4.3. Object States and Relations Prediction

BVS's capabilities extend beyond model evaluation shown in §4.1 and §4.2. Users can also leverage BVS to generate training data with specific object configurations that are hard to collect in large quantities in the real world or difficult to label. In this section, we showcase one practical application of using BVS to generate a synthetic dataset of diverse kinematic object states and relations, and then training a vision model capable of zero-shot transferring to real-world images on the task of object states and relations prediction [3, 13, 14, 19].

Task design and dataset generation. The task of predicting object states and relations, such as `open` and `inside`, is an important perception task [24, 35, 37]. Yet, in the real world, it is challenging to collect such data, let alone the costly annotations. We leverage our generator to synthesize 12.5k images with five labels (`open`, `close`, `ontop`, `inside`, `under`). Each image contains one or more desired labels, e.g. a toy inside an open cabinet. In addition, we manually collected and labeled 850 real images, with unseen object instances and scenes to test for sim2real performance. Examples are shown in Fig. 6.

Baselines and metrics. Adapting from [37], our model takes an image and the bounding boxes for the target objects as input, and outputs a five-way classification over the five labels. We define `open/close` as a binary relationship between the movable link and the unmovable base of an articulated object. For example, the model can be queried whether each of the drawer of a cabinet is open separately, offering very fine-grained understanding of the object state. More details about model architecture are in the Appendix.

We compare our model with zero-shot CLIP, which doesn't train on the synthetic dataset. Specifically, by harnessing CLIP's zero-shot capabilities [40], this baseline outputs a five-way classification prediction by comparing the image embeddings with the five verbalized prompts' text embeddings. We evaluate our model and zero-shot CLIP baseline in terms of precision, recall, and F1, on the synthetic eval set and the real test set.

Results and analysis. Tab. 5 shows the quantitative results on the held-out synthetic dataset and the real dataset for our method. Although there is some performance gap, our model trained on only synthetic data can zero-shot transfer to real images with good overall accuracy. This indicates that BVS offers a promising way to obtain realistic synthetic data that researchers can use not only for evaluation (as shown in §4.1 and §4.2), but also for training models that can then be transferred to the real world. In fact, from Tab. 4, we observe that task-specific training on synthetic data is crucial for good performance on real images.

5. Conclusion

We introduced the BEHAVIOR Vision Suite (BVS), a novel toolset designed to help systematic evaluation and understanding of computer vision models under varying conditions. BVS allows researchers to control a multitude of parameters at various levels—scene, object, and camera—thereby enabling the creation of highly tailored datasets for specific computer vision tasks. Our experiments highlight the versatility and effectiveness of BVS with three key application scenarios. Firstly, we demonstrated its capability in evaluating model robustness against a range of domain shifts, showcasing its utility in helping understand how models perform under diverse and challenging conditions. Secondly, we provided comprehensive benchmarking results of various scene understanding models on a single, common dataset, to show the potential of developing a multi-task method using a single BVS dataset. Finally, we explored the potential of BVS in facilitating

sim2real transfer for novel vision tasks, object states and relations prediction. We aim to provide the computer vision community with a powerful tool that addresses the current data generation challenges. BVS demonstrates the potential of synthetic data in advancing the field, offering researchers a means to generate high-quality, diverse, and realistic datasets tailored to their specific needs.

References

[1] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 3

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1

[3] Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, et al. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20155–20165, 2023. 8

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 1, 3

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 3

[6] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 3

[7] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022. 2, 3

[8] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1

[11] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 1, 3

[12] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020. 2, 3

[13] Yunhao Ge, Harkirat Behl, Jiashu Xu, Suriya Gunasekar, Neel Joshi, Yale Song, Xin Wang, Laurent Itti, and Vibhav Vineet. Neural-sim: Learning to generate training data with nerf. In *European Conference on Computer Vision*, pages 477–493. Springer, 2022. 8

[14] Yunhao Ge, Jiashu Xu, Brian Nlong Zhao, Neel Joshi, Laurent Itti, and Vibhav Vineet. Beyond generation: Harnessing text to image models for object detection and segmentation. *arXiv preprint arXiv:2309.05956*, 2023. 8

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1

[16] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[18] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 1

[19] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and XIAOJUAN QI. Is synthetic data from generative models ready for image recognition? In *The Eleventh International Conference on Learning Representations*, 2022. 8

[20] IDEA-Research. Grounding sam. https://github.com/IDEA-Research/Grounded-Segment-Anything, 2023. 6, 7

[21] Krishna Murthy Jatavallabhula, Soroush Saryazdi, Ganesh Iyer, and Liam Paull. gradslam: Automagically differentiable slam. *arXiv preprint arXiv:1910.10672*, 2019. 7

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[23] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 2

[24] 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. *Int J Comput Vis*, 123:32–73, 2017. 1, 8

[25] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021. 2, 3

[26] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, Mona Anvari, Minjune Hwang, Manasi Sharma, Arman Aydin, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Silvio Savarese, Hyowon Gweon, Karen Liu, Jiajun Wu, and Li Fei-Fei. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Proceedings of The 6th Conference on Robot Learning*, pages 80–93. PMLR, 2023. 2, 3, 4

[27] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 6, 7

[28] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. *arXiv preprint arXiv:1809.00716*, 2018. 1, 3

[29] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, et al. Openrooms: An end-to-end open framework for photorealistic indoor scene datasets. *arXiv preprint arXiv:2007.12868*, 2020. 1, 3

[30] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019. 1

[31] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 1, 7

[32] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[33] 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. *arXiv preprint arXiv:2303.05499*, 2023. 6, 7

[34] Khaled Mamou. Volumetric approximate convex decomposition. In *Game Engine Gems 3*, chapter 12, pages 141–158. A K Peters / CRC Press, 2016. 4

[35] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*, 2018. 8

[36] Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 1

[37] Toki Migimatsu and Jeannette Bohg. Grounding predicates through actions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 3498–3504, 2022. 8

[38] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1, 3, 7

[39] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. idisc: Internal discretization for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21477–21487, 2023. 7

[40] 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. 8

[41] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021. 1, 3

[42] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 7

[43] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In

Proceedings of the IEEE/CVF international conference on computer vision, pages 10912–10922, 2021. 1, 3

[44] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7520–7527. IEEE, 2021. 2, 3

[45] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[46] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[47] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on Robot Learning*, pages 477–490. PMLR, 2022. 2

[48] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems*, 34:251–266, 2021. 2, 3

[49] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013. 1

[50] Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9466–9476, 2023. 7

[51] Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3d meshes with collision-aware concavity and tree search. *ACM Transactions on Graphics (TOG)*, 41(4):1–18, 2022. 4

[52] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018. 1, 3

[53] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 1

[54] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 6, 7

[55] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. *arXiv preprint arXiv:2210.05633*, 2022. 3

[56] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 3

[57] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023. 6, 7

[58] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 6, 7

[59] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 7

[60] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023. 1