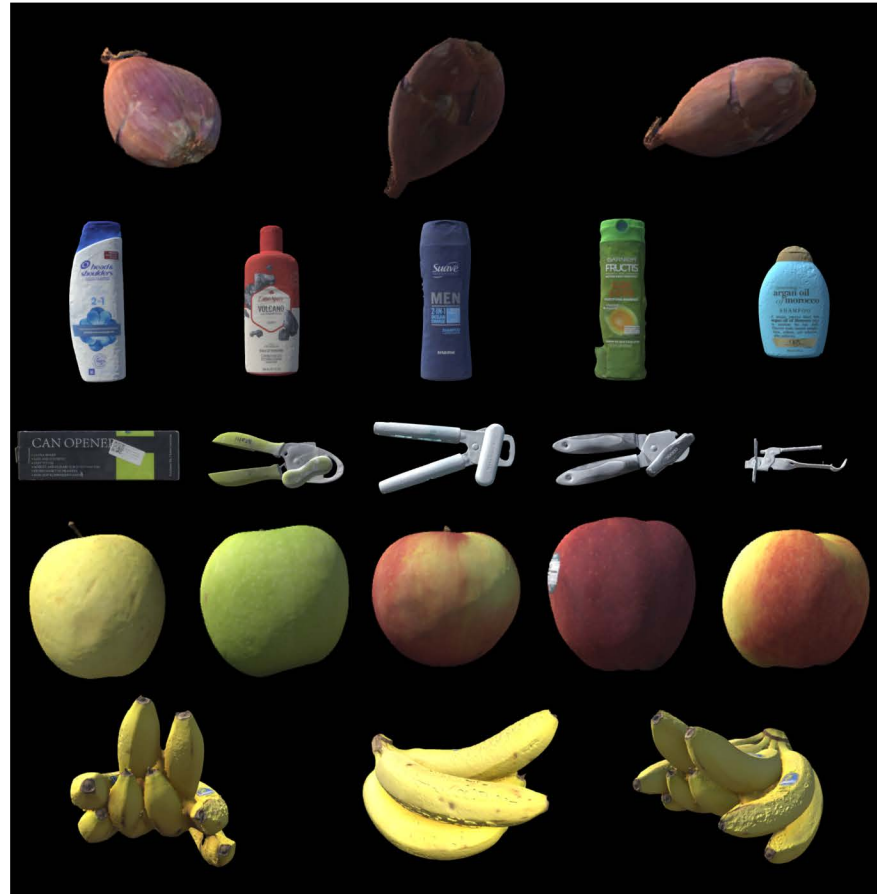


A Large Model's Ability to Identify 3D Objects as a Function of Viewing Angle

Motivation

Identifying and describing objects in the physical world are important tasks with various practical applications, particularly in the fields of mixed reality and robotics. Large multimodal models such as those used in Stable Diffusion, DALL-E, and Midjourney offer significant potential for visual processing, but are trained on images from the Internet.



These images may be from "photogenic" perspectives of objects, rather than being representative of all viewing angles. In this work, we explore the extent to which caption generation varies in accuracy as a function of viewing angle.

Model Capture

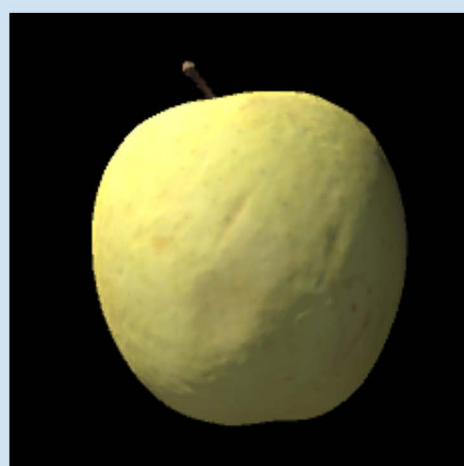
We chose 10 object classes for 3D capture: apple, banana, can opener, gauze, lemon, lime, onion, potato, shampoo, and toothpaste. For most of these, we included five instances per class, with a total of 36 object instances across the 10 classes. For reproducibility of our method, we chose to use photogrammetry for 3D model capture rather than using a less accessible 3D capture methodology (e.g., structured light, LiDAR). Similarly, we pre-treated specular surfaces with dulling spray, but otherwise did not retouch the models provided by off-the-shelf photogrammetry software.

2D Image Generation

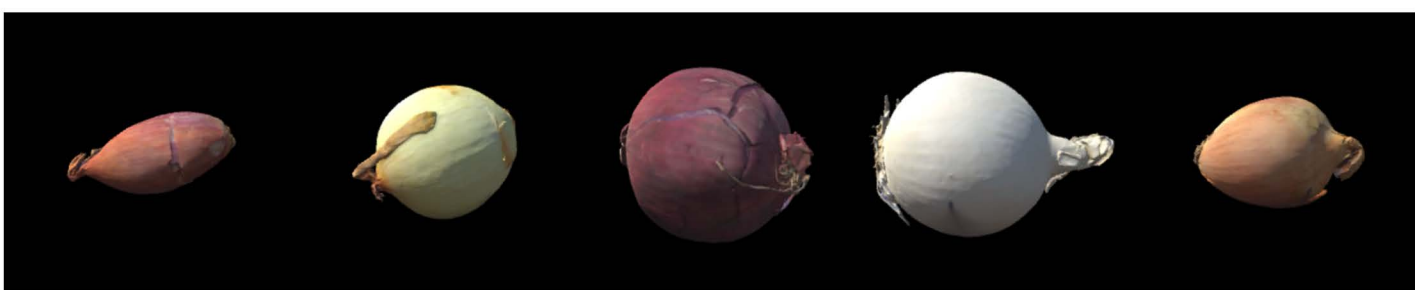
To generate the images of the 3D models from various perspectives, we used the Unity Game Engine. We first created a sphere object at the origin of the scene, made the sphere invisible, and attached the camera object to that sphere. We then moved the camera a distance of 300 units away from the sphere and pointed it towards the origin. With this setup, we were able to rotate the camera to a set of positions on the surface of a sphere of radius 300 while pointing at the origin, allowing us to render images of an object at the origin from many perspectives.

Description Generation

For each input image we provided, Interrogate CLIP provides the string it determines is the most likely text prompt to have resulted in that image being generated by Stable Diffusion. For the purposes of our own research question, we seek to know only if the name of the object class (e.g. **apple**) appears anywhere within the relatively lengthy output prompt. For example, the image of our first apple instance, taken from an inward-facing camera at (0°, 0°) on the surface of the surrounding sphere generated the following prompt shown to the right.



a close up of an **apple** in the dark, cycles4d, phobos, floating planets and moons, octave render, cycles4d render, visiting saturn, rendered in corona, octsne render, inspired by Ma Yuan, charon, spring on saturn, outer wilds, with small object details, pluto, golden **apple**, a raytraced image, saturn



	-180	-162	-144	-126	-108	-90	-72	-54	-36	-18	0	18	36	54	72	90	108	126	144	162	180	
90	4	4	4	5	4	4	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4
81	4	4	4	4	4	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4
72	4	3	4	4	4	4	4	4	4	4	3	4	5	5	5	4	4	4	4	4	4	4
63	3	3	3	3	3	3	4	4	3	3	3	4	4	4	5	5	5	4	3	3	3	3
54	3	2	3	2	2	2	1	2	4	4	4	4	5	5	5	5	5	4	4	3	3	3
45	3	2	3	2	0	0	1	0	3	4	4	4	5	5	5	5	4	4	4	4	3	3
36	3	2	2	1	0	0	0	1	1	4	5	5	5	5	4	4	3	3	4	3	3	3
27	2	0	1	1	1	1	0	1	2	1	5	4	5	5	3	3	2	3	3	3	2	2
18	2	1	0	0	0	1	0	1	0	1	3	5	4	5	3	3	2	3	4	1	2	2
9	1	0	0	0	0	0	0	2	0	1	3	5	5	4	2	2	1	1	4	0	1	1
0	0	0	0	0	0	0	0	0	0	2	4	4	5	3	2	2	1	1	3	0	0	0
-9	0	1	0	0	0	0	0	0	1	0	2	3	4	3	2	2	2	1	2	1	0	0
-18	1	1	1	1	0	1	0	0	1	3	3	4	4	3	3	1	2	1	2	1	1	1
-27	2	1	2	1	1	0	0	0	1	1	3	5	4	3	2	1	2	2	2	2	2	2
-36	1	0	1	4	3	1	0	1	2	2	1	2	4	3	3	3	1	3	2	1	1	1
-45	3	1	2	3	2	2	1	1	2	2	2	2	4	3	3	2	2	3	1	2	3	3
-54	2	1	2	2	3	2	2	3	2	3	2	1	4	2	3	2	2	3	2	1	2	2
-63	2	2	2	2	2	2	3	2	2	3	2	2	1	2	2	3	4	3	3	1	2	2
-72	3	2	2	2	2	2	3	2	1	2	2	2	2	2	2	0	3	3	3	3	3	3
-81	2	1	3	2	2	2	3	2	0	2	0	2	0	2	2	1	1	2	2	2	2	2
-90	1	1	1	1	3	2	3	2	1	2	0	1	2	2	3	1	3	3	2	2	1	1

Results

Despite significant differences between object classes in how easily they can be classified, we found a consistently strong dependence on viewing perspective in determining the likelihood of correctly classifying a given object instance. This dependency could be driven in part by biases in training data. Images of a banana or a tube of toothpaste pulled from the Internet may consistently frame those objects in particular ways as an unconscious design choice of a human photographer. If these models are to be used for real-world image recognition tasks, these weaknesses in the models should be considered.

Acknowledgements

We thank Michael Raphael and Michael Agronin from Direct Dimension for the use of their PASS scanner, used to create the 3D models in this paper.