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

Jacob Rubinstein, Francis Ferraro, Cynthia Matuszek, Don Engel
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County (UMBC), Baltimore, Maryland, USA
{jrubins1, ferraro, cmat, donengel}@umbc.edu

***Abstract*—Virtual reality is progressively more widely used to support embodied AI agents, such as robots, which frequently engage in 'sim-to-real' based learning approaches. At the same time, tools such as large vision-and-language models offer new capabilities that tie into a wide variety of tasks and capabilities. In order to understand how such agents can learn from simulated environments, we explore a language model's ability to recover the type of object represented by a photorealistic 3D model as a function of the 3D perspective from which the model is viewed. We used photogrammetry to create 3D models of commonplace objects and rendered 2D images of these models from an fixed set of 420 virtual camera perspectives. A well-studied image and language model (CLIP) was used to generate text (i.e., prompts) corresponding to these images. Using multiple instances of various object classes, we studied which camera perspectives were most likely to return accurate text categorizations for each class of object.**

***Index Terms*—Multimodal interaction, Virtual Reality, CLIP, 3D Models**

I. Introduction

The ability to accurately identify and describe 3D objects is an important task with various practical applications, particularly in the field of robotics. Embodied agents in human-centric environments need to be able to handle dynamic settings in which objects and tasks may change quickly, and will need to be able to respond to human instructions pertaining to those settings. In this work, we focus on learning to identify objects that occur in such environments, in order to ultimately be responsive to commands such as, e.g., "pack the apple in the lunch basket." Identifying how language is tied to the physical, perceptual world in which agents operate is a component of the widely studied symbol grounding problem [1].

There is work on using virtual reality (VR) to support such *grounded language learning* [2]–[4], with a focus on sim-to-real learning approaches, in which an agent is trained in VR and then learned behaviors are transferred to the physical world. Such machine learning approaches are powerful, but data-hungry, frequently requiring hundreds or thousands of language descriptions of the world to support learning. However, manually generating descriptions for large collections of 3D models can be prohibitively time-consuming and resource-intensive. The motivation for this project is therefore to explore the use of a large language model in automating this process, with a particular focus on whether the viewing angle used to generate an image has a significant impact on the accuracy of a derived image description. In this way, we use models derived from natural language processing and vision-and-language models to support data collection for world understanding on the part of embodied agents, with virtual reality providing the testbed in which agents can be trained.

To understand how large models can support such learning, we first created 3D models of the everyday objects represented in the Grounded Language Dataset (GoLD) [5], which is designed to support exactly the kind of grounded language learning under consideration [6]. The objects in the original GoLD dataset are represented by a combination of RGB and depth images, but not at a resolution that is sufficient to support VR-based learning. Accordingly, we built suitable models of the objects from that dataset using Direct Dimensions' Part Automated Scanning System (PASS), which uses photogrammetry to create photorealistic models. Each object type (e.g., apple) was represented by multiple physical objects, providing a 3D model of each instance of each object type. We then used a custom Unity script to render 2D images of each 3D model from a fixed set of 420 virtual camera perspectives.

The primary contribution of this work is to investigate whether a large vision-and-language model can be used 'in reverse' to generate language describing objects in the environment. We focused on OpenAI's Contrastive Language-Image Pre-Training (CLIP) language model [7] due to its widespread adoption, and generated prompts from images of objects in the environment. In order to generate CLIP prompts corresponding to each 2D image, we use CLIP Interrogator (aka InterrogateCLIP) [8], due to its integration into the popular web-based user interface by AUTOMATIC1111 [9] for Stable Diffusion [10]. We then searched each perspective's generated prompt for the text label corresponding to the object type (e.g., "apple") and rendered a heat map, summed across all objects of that type, to show where the label was (or was not) included in the prompt text.

II. Related Work

The GoLD dataset [5] has previously been used primarily in settings where vision (screen displays) has been used for human interaction, but not virtual reality. We take objects from GoLD, move them from a 2D to a 3D context, and use a large language model instead of direct human annotation to find class labels for objects. In the creation of GoLD, 207 commonplace objects of 47 object types (e.g. apples) were captured as 2D images (825 per object) using a rotating

platform. Mechanical Turk was used to collect 16,500 text and 16,500 audio descriptions of the dataset. While GoLD's 2D images were generated by rotation around a vertical axis, the work described herein is instead framed around the question of arbitrary 3D perspectives, with photogrammetrically-derived 3D models used to generate 2D images from viewpoints outside a single horizontal plane. We were able to acquire only a subset of the original GoLD objects, as detailed further below. We also differ from the GoLD paper's approach by exploring how accurately these images can be annotated by a large language model (rather than by humans) and differ in exploring the labeling accuracy as a function of viewing angle. Other prior work has been done in the spaces of comparing 3D models from multiple views using a bag-of-features approach, but in that previous work, the features were visual rather than text-based [11].

Recent research has explored other uses of CLIP Interrogator, such as its application to curating works of art [12] and detection of harmful memes [13]. The potential for "prompt stealing" (i.e., reverse engineering the prompts used to generate an image) has been explored by Shen *et al.* [14]. The aforementioned papers focused on applications of CLIP Interrogator, whereas we are interested in studying the functionality of CLIP Interrogator itself. The most closely related work to ours is that of Udo and Koshinaka [15], who explored the relative accuracy of CLIP Interrogator and other prompt generation tools. Our work is similar in that it studies the accuracy of prompt generation, but does so within a single tool, as a function of viewing angle.

Our work is also distinct in that we are interested in generating models that can be used in virtual reality environments to support robotic interaction. There exists extensive work in the intersection of robotics and virtual reality/augmented reality [16]–[18], including in the human-robot interaction space [19], [20], but comparatively few works focus on natural language in such a setting; the work that does exist tends to be focused on specific problem spaces (e.g., teleoperation [21] or swarm robotics [22]), despite interest in the subject [23].

III. Experimental Methodology

A. 3D Model Capture

The initial stage of our work consisted of creating 3D models of a sampling of objects from the Grounded Language Dataset (GoLD) [5] dataset. For reproducibility of our method, we chose to use photogrammetry for 3D model capture rather than using a less accessible 3D capture methodology (structured light, LiDAR, etc.). In principle, photogrammetry allows for models of stationary objects to be captured using any digital camera (e.g., phone camera) by collecting images representing a sufficient number of perspectives and running them through 3D reconstruction software. Many versions of such photogrammetric software exist, with varying degrees of usability and performance. We initially experimented with several cloud and desktop-based services for reconstructing 3D models of photosets captured by our mobile phones. For the sake of repeatability and speed, we instead ended up capturing the models for this paper using the Part Automated Scanning System (PASS) invented by Direct Dimensions. PASS enabled us to capture content more quickly than a piecemeal solution thanks to its use of multiple cameras, a staging platform, lighting, and reconstruction software.

The original GoLD dataset includes five high-level object categories (food, home, medical, office, tool). 47 object classes (e.g., "apple") are spread across these five categories, and each class is represented by four or five instances (e.g., five distinct apples). Whereas the GoLD project used a turntable and then selected about four representative 2D images for each of their 207 object instances, we are particularly interested in having many more images per object, and are more limited by the complications of 3D object capture.

We chose 10 of GoLD's 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 (e.g., five distinct apples), with a total of 36 object instances across the 10 classes.

B. 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 change the rotation the sphere to cause the camera to travel around a sphere of radius 300 while pointing at the origin, allowing us to render images of an object at the origin from many perspectives.

The next step was to add each of the 36 3D models to the scene, making sure to align their centers at the origin and aligning their front to the initial camera view. We created prefabs of each of these object placements, thereby allowing us script their appearances.

Our algorithm used to capture images of the objects is further detailed in algorithm 1. The z angle ranges from -90 degrees to 90 degrees which captures view of the objects from bottom to top; this is equivalent to the camera's latitude on the sphere. The y angle ranges from -180 degrees to 180 degrees which captures views of the object in a horizontal loop; this is equivalent to the camera's longitude on the sphere.

We then created a script to capture the images and attached it to the sphere object. In this script, we created an array of GameObjects and added all of the prefab objects. Other initialization steps included making the prefabs invisible and resetting the sphere's rotation.

The images are saved in a folder with the object's name and are named after the x, y, and z coordinates of the sphere's rotation.

C. Description Generation

To generate descriptions of the object, we used the "Interrogate CLIP" feature of Stable Diffusion. For this task, we used the stable-diffusion-webui GitHub repository. We used

Algorithm 1 Image Capturing

```
for object in objectArray do
    object.visible = true
    for zAng = −90; zAng <= 90; zAng+ = 9 do
        for yAng = −180; yAng < 180; yAng+ = 18 do
            sphere.rotation = (0, yAng, zAng)
            Capture Screenshot
        end for
    end for
    sphere.rotation = (0, 0, 0)
    object.visible = false
end for
```

an extension which allowed for batch clip interrogation and exported these results to a comma-separated values (CSV) file for each image.

For each input image, "Interrogate CLIP" provides the string it determines is the most likely text prompt to have resulted in that image being generated by Stable Diffusion. The tool is open source, but lacks technical documentation and no research papers have been published on it by its authors. This lack of technical documentation is noted by Udo and Koshinaka [15], whose paper includes their own analysis-based explanation of Interrogate CLIP's methodology. For the purposes of our own research question, we seek to know only if the name of the GoLD 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 (Figure 1), generates the prompt:

> 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

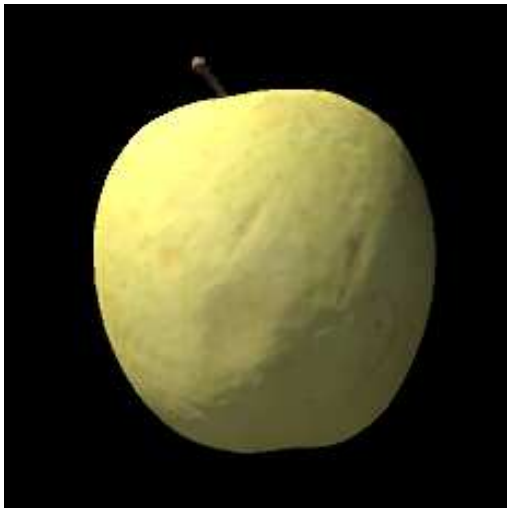


Fig. 1. apple_1, viewed from $(0°, 0°)$

Because the word "apple" occurs at least once in this prompt, we consider this to include the name of the object class.

IV. Results

A. Derived 2D Images

Qualitatively, most of the 36 object instances provided photorealistic views from each of their 420 perspectives, as demonstrated in Figures 2, 3, and 4.

Fig. 2. Example of onion_1 from 3 perspectives.

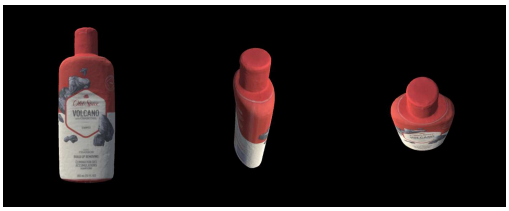


Fig. 3. Example of shampoo_2 from 3 perspectives.

Fig. 4. Example of banana_2 from 3 perspectives.

In general, photogrammetry has performance issues with surfaces that are specular or semitransparent, as reflections interfere with feature detection algorithms' ability to identify a consistent set of points of interest across multiple perspectives [24]. Our collection of scanned objects included several glossy items. We attempted to mitigate this challenge through the application of Krylon Dulling Spray, a transparent coating which reduces the specularity of the surfaces to which it is applied. This pretreatment proved effective for some of our objects, such as metallic can openers (Figure 5), but was an incomplete solution for a few of our objects, most notably our fourth shampoo bottle (Figure 6). Models broken in this way are often fixed by 3D graphics artists before being used in other contexts (e.g., 3D assets for game development). We did not apply such touch-ups to our 3D models and left their imperfections intact, thereby preserving the automaticity and consistency in our dataset's production.

B. Overall Retrieval of Each Object Class

Let us define "image classification" in the context of this paper to mean the inclusion of an object class's name in the prompt generated for that image. That is, an image of an apple

Fig. 5. Example of can_opener_3 from 3 perspectives.

Fig. 6. The shampoo_4 model was particularly malformed.

Object Name	Percent	Object Name	Percent
APPLE_1	73.6%	ONION_3	42.6%
APPLE_2	89.8%	ONION_4	56.0%
APPLE_3	80.2%	ONION_5	43.3%
DAPPLE_4	92.9%	POTATO_1	65.0%
APPLE_5	92.1%	POTATO_2	87.9%
BANANA_1	99.0%	POTATO_3	9.8%
BANANA_2	99.5%	POTATO_4	3.1%
BANANA_3	98.6%	POTATO_5	44.3%
CAN_OPENER_2	5.7%	SHAMPOO_1	0.2%
CAN_OPENER_2_BOX	31.0%	SHAMPOO_2	11.7%
CAN_OPENER_3	6.2%	SHAMPOO_3	0.0%
CAN_OPENER_4	1.2%	SHAMPOO_4	1.2%
CAN_OPENER_5	3.1%	SHAMPOO_5	10.0%
GAUZE_1	40.0%	TOOTHPASTE_1	71.4%
LEMON	74.8%	TOOTHPASTE_2	78.3%
LIME	18.6%	TOOTHPASTE_3	64.3%
ONION_1	54.5%	TOOTHPASTE_4	79.5%
ONION_2	45.0%	TOOTHPASTE_5	78.6%

TABLE I
PERCENT OF IMAGES FOR EACH OBJECT INSTANCE WHERE THE OBJECT CLASS NAME APPEARS IN THE GENERATED PROMPT.

which includes the word "apple" anywhere in its prompt is considered correctly classified. Notably, this definition does not preclude an image from being classified into multiple classes at once (i.e., a prompt might say both "lemon" and "lime"); we address multi-classification in the next section.

Before exploring multiple classifications or the dependence of generated prompts on viewing angle, we ask a simpler question: to what extent are the names of object classes included in the images of each instance of that object class, across all viewing angles? This question is motivated by the fact that some objects classes pose challenges with generating quality models using photogrammetry, as discussed in Section IV-A. Another overarching issue is that the CLIP neural network may include more accurate representations of some object classes than others.

As shown in Table I, there is variation both by class and by instance. Bananas are the most consistently classified correctly (98.6% of images for the worst banana to 99.5% of images for the best banana), with apples close behind (89.3% to 92.1%, except for the only yellow apple, apple_1, at 73.6%). Although images of toothpaste proved harder to classify, there is limited variation as a function of which toothpaste instance (i.e., 3D model) is being considered.

Some other classes have significant differences within that class. The potato class is the most significant example of this, with 88% of the images of potato_2 correctly classified, while potato_4 is correctly classified only 3% of the time. This is perhaps explained by the relatively spherical, less oblong shape of potato_4, and a human would perhaps have similar difficulty distinguishing a nearly-round potato on a black background, without context, from being a rock or a moon.

All object instances of shampoos and can openers are quite low in their image classification scores, which is perhaps explained by relatively poor model quality due to these object classes being more specular than the other object classes.

C. Incorrect Classifications and Multiple Classifications

As discussed in Section IV-B, an image will be classified into multiple classes of object if the names of multiple object classes occur within its generated prompt. Of the 15,120 images generated from the 36 object instances, 41% images gave CLIP Interrogator output with no class names; 51% included exactly one class name; 7% included exactly two class names; and 1% included exactly three class names. No labels included more than three class names.

Actual Class	apple	banana	can opener	gauze	lemon	lime	onion	potato	shampoo	toothpaste
apple	86%	0%	0%	0%	6%	19%	2%	0%	0%	0%
banana	0%	99%	0%	0%	2%	0%	0%	0%	0%	0%
can opener	1%	2%	2%	0%	0%	0%	0%	0%	0%	2%
gauze	0%	0%	0%	40%	0%	0%	0%	0%	0%	30%
lemon	21%	10%	0%	0%	75%	5%	0%	0%	0%	0%
lime	70%	0%	0%	0%	16%	19%	2%	0%	0%	0%
onion	21%	0%	0%	0%	0%	2%	48%	1%	0%	0%
potato	9%	0%	0%	0%	0%	1%	0%	42%	0%	0%
shampoo	0%	0%	0%	0%	0%	1%	0%	0%	5%	37%
toothpaste	0%	0%	0%	0%	0%	0%	0%	0%	0%	74%

TABLE II
PERCENT OF IMAGES (PER OBJECT CLASS) THAT HAVE EACH CLASS LABEL. BECAUSE IMAGES CAN BE CLASSIFIED AS ZERO, ONE, OR MULTIPLE OBJECT CLASSES, ROWS CAN SUM TO LESS THAN 100% OR MORE THAN 100%.

As shown in Table II, some classes are more likely to be confused with each other. 70% of the rendered images of lime object instances included the word "apple" in their generated prompt; only 19% of the lime images were categorized as "lime," followed by 16% of the lime images being categorized as "lemon."

At the other end of the spectrum, banana objects proved easier to categorize correctly, with 99% of rendered banana images categorized as "banana," followed distantly by 2% being categorized as "lemon." This label makes sense when viewing a banana along its axis, with the bottom tip of the banana occluding most of the rest of the banana from view.

D. Effect of Viewing Angle on Object Classification

Having considered the likelihood of categorizing an image correctly, as well as the possibility of how it may be categorized incorrectly, we can now turn to the major research question posed herein: how does the viewing angle affect an image and language model's ability to correctly identify an object?

In this section, we will explore those object classes for which we have at least three instances (i.e., distinct 3D models) of that class. As discussed in earlier sections, apples and bananas proved easiest to classify. Interestingly, there is a clear angular dependence on the ease of classification.

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

Fig. 7. Top: The five apple instances as viewed from $(0°, 0°)$; Bottom: Heatmap showing the number of prompts generated by Interrogate CLIP which included the word "apple"

All five apple instances (Figure 7) are correctly categorized 100% of the time when viewed from the northern hemisphere of the viewing sphere. When viewed from even slightly below the equator, accuracy begins to suffer. It is likely that a visual cue, embedded in the CLIP model, is lost as the top of the apple begins to become occluded. As the stem and the indented area around the stem are removed from view, the word "apple" is much less likely to appear in the generated text prompt. It seems likely that a human would face similar difficulty. As discussed earlier, this is particularly true for the yellow apple, likely due to its atypical coloration.

In Section IV-C, we noted that images of instances of the banana class are the most easily classified. As shown in Figure 8, the orientation at which bananas become the most difficult for CLIP to recognize is when they are laying flat

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

Fig. 8. Top: The three banana instances as viewed from $(0°, 0°)$; Bottom: Heatmap showing the number of prompts generated by Interrogate CLIP which included the word "banana"

(so their signature curved shape is hidden) and pointed along the line of sight - i.e., $(0°, \pm 90°)$. At this orientation, the observer sees only a yellow roundish shape, which is likely why Interrogate CLIP includes "lemon" in banana images under such circumstances.

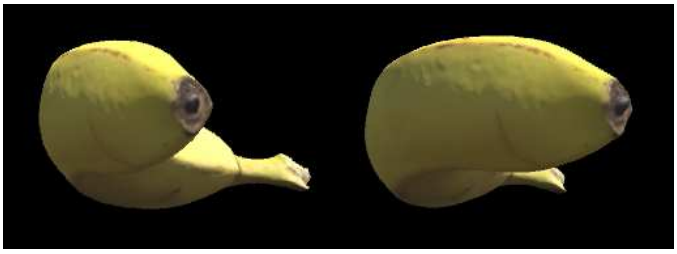


Fig. 9. banana_1 viewed from $(-9°, 90°)$ and $(-9°, 108°)$

For example, when banana_1 is viewed from $(-9°, 90°)$ (Figure 9, left), it is classified as both "banana" and "lemon":

> a close up of a **banana** on a black background, **lemon** wearing sunglasses, anorld render, a bot in the game super mario 64, spherical body, 3d game object, videogame asset, slimy unreal engine, hatched ear, seperated game asset, game asset, pear, povray, yellow beak, low quality 3d model

From $(-9°, 108°)$ (Figure 9, right), banana_1 is only classified as "lemon":

> a close up of a **lemon** on a black background, anorld render, modeled in 3 d, 3 d render of jerma 9 8 5, lemon wearing sunglasses, an angry lemon, masterpiece. rendered in blender, moonray render,

low quality 3d model, 3d rendered, 3 d raytraced masterpiece, 3 d rendered, 3 d model rip

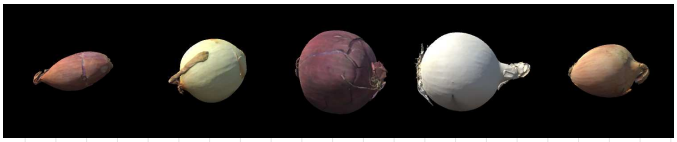
	-180	-162	-144	-126	-108	-90	-72	-54	-36	-18	0	18	36	54	72	90	108	126	144	162	180
90	0	0	0	1	1	2	2	1	1	1	1	2	2	1	0	2	1	0	1	1	0
81	0	0	0	0	0	2	1	0	0	0	0	1	0	0	2	0	2	0	0	0	0
72	0	0	1	0	0	2	1	0	1	0	0	2	0	1	1	1	0	0	0	0	0
63	1	0	0	0	0	0	0	0	0	1	0	2	1	0	0	1	0	0	0	0	1
54	1	0	0	0	1	0	0	0	0	1	1	1	1	0	1	2	0	0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

Fig. 10. Top: The five can opener instances as viewed from $(0°, 0°)$; Bottom: Heatmap showing the number of prompts generated by Interrogate CLIP which included the phrase "can opener"

Can openers (Figure 10) were second only to shampoo bottles in being the least likely to be correctly classified. As the heatmap shows, there appears to be a very strong dependence on perspective, with images classified correctly only when oriented around $(0°, 0°)$, which is the orientation depicted in the figure. However, some of this may be attributed to the inclusion of the can opener box in this image set, as it has the words "Can Opener" on its packaging and is a significant contributor to the non-zero values in the heatmap, as shown in Table I.

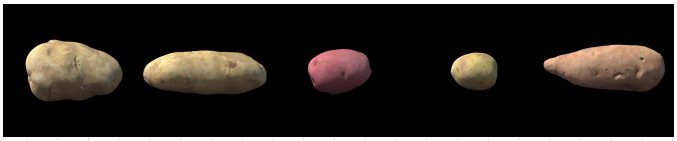
The onion heatmap (Figure 11) shows a viewing angle dependency similar in nature to that of the apple, discussed above. When the onions are rotated such that their stems are occluded - which happens around $(0°, -90°)$ - they become much more difficult to identify as onions (i.e., Interrogate CLIP is much less likely to include the word "onion" in the generated prompt). There is significantly more sensitivity to orientation than we saw with the apples, perhaps because apples (other than the yellow one) are more easily identified by their round shapes and distinctive coloring, while the onion instances have more varied colors and shapes.

The potato heatmap (Figure 12) shows a hotspot around $(0°, 0°)$, with all five potatoes correctly classified across six contiguous rotation steps. This is the orientation depicted in the figure, which is a viewpoint perpendicular to the major axis of the potatoes. As discussed in Section IV-B, potato_4 performs worst with classification and is also the most spherical. The second most spherical potato, potato_3, fares nearly as badly. This tells the same story as the heatmap; a potato image is most likely to be missing the word "potato" in the CLIP Interrogator output when the potato appears round, rather



	-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
81	4	4	4	4	4	4	4	4	4	4	3	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
63	3	3	3	3	3	3	4	4	3	3	3	4	4	4	5	5	5	4	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
45	3	2	3	2	0	0	1	0	3	4	4	4	5	5	5	5	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
27	2	0	1	1	1	0	1	2	1	5	4	5	5	3	3	2	3	3	3	3	2
18	2	1	0	0	0	1	0	1	0	1	3	5	4	5	3	3	2	3	4	1	2
9	1	0	0	0	0	0	0	2	0	1	3	5	5	4	2	2	1	1	4	0	1
0	0	0	0	0	0	0	0	0	0	2	4	4	5	3	2	2	1	1	3	0	0
-9	0	1	0	0	0	0	0	0	1	0	2	3	4	3	2	2	2	1	2	1	0
-18	1	1	1	1	0	1	0	0	1	3	3	4	4	3	3	1	2	1	2	1	1
-27	2	1	2	1	1	0	0	0	1	1	3	5	4	3	2	1	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
-45	3	1	2	3	2	2	1	1	2	2	2	2	4	3	3	2	2	3	1	2	3
-54	2	1	2	2	3	2	2	3	2	3	2	1	4	2	3	2	2	3	2	1	2
-63	2	2	2	2	2	2	3	2	2	3	2	2	1	2	2	3	4	3	3	1	2
-72	3	2	2	2	2	2	3	2	1	2	2	2	2	2	2	0	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
-90	1	1	1	1	3	2	3	2	1	2	0	1	2	2	3	1	3	3	2	2	1

Fig. 11. Top: The five onion instances as viewed from $(0°, 0°)$; Bottom: Heatmap showing the number of prompts generated by Interrogate CLIP which included the word "onion"



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

Fig. 12. Top: The five potato instances as viewed from $(0°, 0°)$; Bottom: Heatmap showing the number of prompts generated by Interrogate CLIP which included the word "potato"

than oblong. This is true regardless of whether the apparent roundness is due to a particular viewing perspective or due to the potato being nearly spherical in 3D.

While in Section III-A it was noted that shampoo bottles may be difficult to classify in part due to relatively poor model quality, we nonetheless see a very strong dependence on viewing angle in Figure 13. All five bottle instances have cross-sections that are quite oblong, with labels printed on the wide surfaces of the bottles. It is when the viewpoint is perpendicular to these wide surfaces, around $(0°, 0°)$ and $(0°, 180°)$, that the rendered images are most likely to include the correct "shampoo" categorization.

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

Fig. 13. Top: The five shampoo instances as viewed from $(0°, 0°)$; Bottom: Heatmap showing the number of prompts generated by Interrogate CLIP which included the word "shampoo"

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

Fig. 14. Top: The five toothpaste instances as viewed from $(0°, 0°)$; Bottom: Heatmap showing the number of prompts generated by Interrogate CLIP which included the word "toothpaste"

The toothpaste instances (Figure 14) have a strong dependency on viewpoint that is related to both the shampoo bottles and the bananas. When the flatter, larger faces are perpendicular to the line of sight, they are more likely to be correctly classified. This is particularly clear from the cold spots in the heatmap, at $(0°, \pm 90°)$, where the viewpoint is directly along the length of the toothpaste object, resulting in an image which would only show the tip or end of the toothpaste tube (or box), occluding the signature shape, as well as the labelling.

E. Conclusions

Despite significant differences between object classes in how easily they can be classified, there is nonetheless a consistently strong dependence on viewing perspective in determining the likelihood of correctly classifying a given object instance. In some ways, it is hardly surprising that objects are harder or easier to identify depending on viewpoint, regardless of whether that viewpoint is held by a human or by a vision-and-language neural network model. However, in the context of models, 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. One would rarely think to take a picture of a banana with its length oriented along the line of sight, or of the bottom of an apple. For certain applications of multimodal large models, such as robots embedded in a real 3D environment, objects may be less likely to be oriented in those "photogenic" perspectives, from the robot's perspective, as it navigates the world.

On the other hand, our study shows that the weakest viewpoints for CLIP with each object mirror the weakest viewpoints of humans for those same objects. To some degree, it is plausible that objects become fundamentally harder to identify when oriented in ways which mask their most distinctive traits (e.g., curvature of a banana or stem of an apple).

In either case, these results show that large models do have viewpoint dependencies, which can be taken into account in the future, either by anticipating these dependencies and planning accordingly, or by building models which are trained on images captured from a wider range of visual perspectives.

V. Future Work

This work is a component of a longer-term project aimed at using virtual reality to study grounded language acquisition. The GoLD dataset which inspired this project was a set of object descriptions (text and audio) collected using 2D images and Mechanical Turk, with the aim of providing a dataset for grounded language acquisition research that will advance human-robot interaction. The GoLD images were captured from viewpoints at a fixed height around a turntable, yet robots are immersed in 3D world. We intend to use the 3D models and knowledge described herein to conduct a study similar to GoLD, but which would instead record humans describing objects they are seeing in virtual reality. We are interested in seeing how these descriptions compare to those from the original GoLD dataset and perhaps to those which could be gathered by in-person viewing of the real physical objects.

VI. 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.

References

[1] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1, 1990.

[2] K. M. Hermann, F. Hill, S. Green, F. Wang, R. Faulkner, H. Soyer, D. Szepesvari, W. M. Czarnecki, M. Jaderberg, D. Teplyashin, M. Wainwright, C. Apps, D. Hassabis, and P. Blunsom, "Grounded language learning in a simulated 3d world," *CoRR*, vol. abs/1706.06551, 2017.

[3] P. Anderson, A. Shrivastava, J. Truong, A. Majumdar, D. Parikh, D. Batra, and S. Lee, "Sim-to-real transfer for vision-and-language navigation," in *Conference on Robot Learning*. Proceedings of Machine Learning Research, 2021, pp. 671–681.

[4] B. Ichter, A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian, D. Kalashnikov, S. Levine, Y. Lu, C. Parada, K. Rao, P. Sermanet, A. T. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, M. Yan, N. Brown, M. Ahn, O. Cortes, N. Sievers, C. Tan, S. Xu, D. Reyes, J. Rettinghouse, J. Quiambao, P. Pastor, L. Luu, K.-H. Lee, Y. Kuang, S. Jesmonth, N. J. Joshi, K. Jeffrey, R. J. Ruano, J. Hsu, K. Gopalakrishnan, B. David, A. Zeng, and C. K. Fu, "Do as I can, not as I say: Grounding language in robotic affordances," in *Proceedings of The 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. Proceedings of Machine Learning Research, 14–18 Dec 2023, pp. 287–318.

[5] G. Y. Kebe, P. Higgins, P. Jenkins, K. Darvish, R. Sachdeva, R. Barron, J. Winder, D. Engel, E. Raff, F. Ferraro *et al.*, "A spoken language dataset of descriptions for speech-based grounded language learning," in *Thirty-fifth Conference on Neural Information Processing Systems*, 2021.

[6] G. Y. Kebe, L. E. Richards, E. Raff, F. Ferraro, and C. Matuszek, "Bridging the gap: Using deep acoustic representations to learn grounded language from percepts and raw speech," in *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, February 2022.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[8] pharmapsychotic, "GitHub - pharmapsychotic/clip-interrogator: Image to prompt with BLIP and CLIP," https://github.com/pharmapsychotic/clip-interrogator, [Accessed 29-07-2023].

[9] AUTOMATIC1111, "GitHub - AUTOMATIC1111/stable-diffusion-webui: Stable Diffusion web UI," https://github.com/AUTOMATIC1111/stable-diffusion-webui, 2023, [Accessed 29-07-2023].

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.

[11] Z. Lian, A. Godil, and X. Sun, "Visual similarity based 3d shape retrieval using bag-of-features," in *2010 Shape Modeling International Conference*, 2010, pp. 25–36.

[12] L. Schaerf, P. Ballesteros, V. Bernasconi, I. Neri, and D. N. del Castillo, "AI art curation: Re-imagining the city of Helsinki in occasion of its biennial," September 2023, [arXiv preprint http://arxiv.org/abs/2306.03753].

[13] J. Ji, W. Ren, and U. Naseem, "Identifying creative harmful memes via prompt based approach," in *Proceedings of the ACM Web Conference 2023*, ser. WWW '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 3868–3872.

[14] X. Shen, Y. Qu, M. Backes, and Y. Zhang, "Prompt stealing attacks against text-to-image generation models," February 2023, [arXiv preprint http://arxiv.org/abs/2302.09923].

[15] H. Udo and T. Koshinaka, "Image captioners sometimes tell more than images they see," May 2023, [arXiv preprint http://arxiv.org/abs/2305.02932].

[16] G. C. Burdea, "Invited review: the synergy between virtual reality and robotics," *IEEE Transactions on Robotics and Automation*, vol. 15, no. 3, pp. 400–410, 1999.

[17] J. J. Roldán, E. Peña-Tapia, D. Garzón-Ramos, J. de León, M. Garzón, J. del Cerro, and A. Barrientos, "Multi-robot systems, virtual reality and ros: developing a new generation of operator interfaces," *Robot Operating System (ROS) The Complete Reference (Volume 3)*, pp. 29–64, 2019.

[18] H. Shi, G. Liu, K. Zhang, Z. Zhou, and J. Wang, "Marl sim2real transfer: Merging physical reality with digital virtuality in metaverse," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 4, pp. 2107–2117, 2022.

[19] M. Wozniak, C. T. Chang, M. B. Luebbers, B. Ikeda, M. Walker, E. Rosen, and T. R. Groechel, "Virtual, augmented, and mixed reality for human-robot interaction (vam-hri)," in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 938–940.

[20] L. Wijnen, S. Lemaignan, and P. Bremner, "Towards using virtual reality for replicating hri studies," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 514–516.

[21] A. Martín-Barrio, J. J. Roldán, S. Terrile, J. Del Cerro, and A. Barrientos, "Application of immersive technologies and natural language to hyper-redundant robot teleoperation," *Virtual Reality*, vol. 24, pp. 541–555, 2020.

[22] M. Chen, P. Zhang, Z. Wu, and X. Chen, "A multichannel human-swarm robot interaction system in augmented reality," *Virtual Reality & Intelligent Hardware*, vol. 2, no. 6, pp. 518–533, 2020.

[23] T. Williams, D. Szafir, T. Chakraborti, and H. Ben Amor, "Virtual, augmented, and mixed reality for human-robot interaction," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 403–404.

[24] A. Mathys, J. Brecko, D. Van den Spiegel, and P. Semal, "3d and challenging materials," in *2015 Digital Heritage*, vol. 1, 2015, pp. 19–26.