Assessing VLM-Driven Semantic-Affordance Inference for Non-Humanoid Robot Morphologies

Jess Jones
Bristol Robotics Laboratory, University of Bristol
Bristol, United Kingdom
jess.jones@bristol.ac.uk

Sabine Hauert*
Bristol Robotics Laboratory, University of Bristol
Bristol, United Kingdom
sabine.hauert@bristol.ac.uk

Raul Santos-Rodriguez*
University of Bristol
Bristol, United Kingdom
enrsr@bristol.ac.uk

ABSTRACT

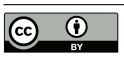
Vision-language models (VLMs) have demonstrated remarkable capabilities in understanding human-object interactions, but their application to robotic systems with non-humanoid morphologies remains largely unexplored. This work investigates whether VLMs can effectively infer affordances for robots with fundamentally different embodiments than humans, addressing a critical gap in the deployment of these models for diverse robotic applications. We introduce a novel hybrid dataset that combines annotated real-world robotic affordance-object relations with VLM-generated synthetic scenarios, and perform an empirical analysis of VLM performance across multiple object categories and robot morphologies, revealing significant variations in affordance inference capabilities. Our experiments demonstrate that while VLMs show promising generalisation to non-humanoid robot forms, their performance is notably inconsistent across different object domains. Critically, we identify a consistent pattern of low false positive rates but high false negative rates across all morphologies and object categories, indicating that VLMs tend toward conservative affordance predictions. Our analysis reveals that this pattern is particularly pronounced for novel tool use scenarios and unconventional object manipulations, suggesting that effective integration of VLMs in robotic systems requires complementary approaches to mitigate over-conservative behaviour while preserving the inherent safety benefits of low false positive rates.

KEYWORDS

Robotics; Multi-Robot Systems; Non-Humanoid Robots; Affordance; VLMs

ACM Reference Format:
Jess Jones, Sabine Hauert, and Raul Santos-Rodriguez. 2026. Assessing VLM-Driven Semantic-Affordance Inference for Non-Humanoid Robot Morphologies. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/WTKR8312

*These authors contributed equally to this work.



Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/WTKR8312

1 INTRODUCTION

Bringing robotic systems out of controlled laboratory settings into dynamic, real-world environments remains a fundamental challenge on the path toward a future where smart machines are ubiquitous within human society. Achieving robust and adaptive autonomy requires robots to not only perceive their surroundings but also intuitively understand how they can interact with objects and other robots. This is particularly true for future embodied AI, likely heterogeneous in both form and function, with diverse robots developed for specific tasks, often by different commercial entities [7]. Enabling adaptive, collaborative behaviours within such heterogeneous multi-robot systems necessitates a foundational approach to grounding robot interactions.

We hypothesise that grounding these interactions through the concept of affordances [12] can substantially improve their capacity to learn and generalise collaboration strategies. Affordances, defined as the actionable possibilities an environment offers an agent, intrinsically link perception to potential actions. For a robot, understanding affordances means recognising not just what an object is but how it can be interacted with given the robot's unique physical capabilities and the current task. This shifts the paradigm from basic object classification and localisation to dynamic, context-aware interaction, paving the way for more flexible and intelligent robotic systems.

Our broader research proposes to leverage the semantic encoding power of VLMs to enhance the scene-representation of each robot with semantic-affordance descriptions of the objects it encounters, linking its unique capabilities to its environment, and providing a framework from which collaborative interactions can be derived in multi-robot settings in the future. While VLMs show promise, their application to robotics presents a critical challenge, rooted in the fact that their training data is predominantly human-centric [20]. This raises a fundamental question, which this paper aims to address through empirical analysis: Can VLMs effectively infer affordances for robots with fundamentally different embodiments than humans? We use a novel hybrid dataset, curated for non-humanoid robotic affordance inference, to investigate the performance of VLMs in supporting the zero-shot affordance characterisation of objects across a breadth of diverse contexts, including household items, food products, environmental clean-up, and construction materials. Performance is evaluated on multiple non-humanoid robot morphologies to understand the generalisability and limitations of these models. Our research reveals variations in affordance inference capabilities, identifying biases where VLMs tend toward conservative predictions, characterised by low false positive rates but notably high false negative rates. This conservative bias, while minimising

erroneous actions, means robots may underutilise their true capabilities, missing valid interaction and collaboration opportunities, particularly for novel tool use or unconventional manipulations.

Our findings provide critical insights into the practical deployment of VLMs in real-world multi-robot systems. By quantifying the performance variations and identifying the nature of VLM biases in affordance understanding for non-humanoid forms, this work lays the groundwork for developing more robust and reliable embodied AI. The enriched scene-representations, which can be derived from the affordance characterisations of this approach, capture not only what and where objects are, but how the local environment can be interacted with given an arbitrary robot morphology. This will be leveraged in future work to inform action priors that ground learning task decomposition and decision-making strategies in collaborative multi-robot settings, integrating both intrinsic robot abilities and situational demands.

2 RELATED WORK

The ability of smart machines to perceive and interact with their environment is fundamental to achieving robust autonomous behavior. A cornerstone concept in this regard is that of affordances [12]. Affordances refer to the actionable possibilities that the environment offers to an individual, which Gibson argues, is not exclusive to humans, but extends to all animals based on a fundamental animal-environment interaction system [11]. The framing of affordances provides a model which intrinsically links perception and morphology to action. For robotic systems, understanding affordances is crucial for tasks ranging from object manipulation and navigation, to collaborative multi-robot, and human-robot interaction [2]. Rather than merely identifying objects, robots must infer how they can interact with those objects given their own physical capabilities and the task at hand.

Traditionally, affordance perception in robotics has relied on hand-engineered features, geometric reasoning, or supervised learning on curated datasets specific to particular objects, robots, and tasks [6, 14]. While effective in constrained environments, these approaches often struggle with generalisation to novel objects, unseen environments, or diverse robot morphologies. The high cost of data annotation and the limited scalability of these methods present significant barriers to deploying robots in unstructured, dynamic real-world settings.

The recent emergence of large-scale VLMs has impacted various fields within artificial intelligence, including computer vision and natural language processing [10]. Trained on vast amounts of internet-scale image-text data, VLMs have shown capabilities in zero-shot generalisation, common-sense reasoning, and cross-modal understanding [21, 23]. Their ability to connect visual input with high-level semantic concepts, often expressed in natural language, presents a compelling opportunity to address the limitations of traditional affordance perception methods in robotics.

Initial research has begun to explore the application of VLMs to robotic affordance prediction, primarily by leveraging their inherent understanding of human-object interactions [3, 8, 17, 18, 22, 24]. These works often frame affordance as a segmentation, localisation or trajectory modelling problem, identifying regions of objects that afford certain human actions (e.g. "grasp," "cut," "pour"). However, a critical gap remains in understanding how these human-centric affordance concepts translate to non-humanoid robotic embodiments. Robots possess diverse manipulators, grippers, and end-effectors, with physical capabilities that may differ significantly from human hands. This disparity raises fundamental questions about the direct applicability and generalisation capabilities of VLMs across the broader spectrum of existing and future robotic morphologies, and their unique interaction possibilities. This work represents an initial exploration of this unexplored territory, investigating the effectiveness of VLMs in inferring affordances for robots with fundamentally different embodiments than humans, and through empirical analysis, understanding the systematic biases that may arise from their human-centric training data.

Building upon advancements in foundation models and their applications, a body of work has focused on the creation and utilization of rich semantic scene representations. Rather than simply identifying individual objects or their spatial location, these representations aim to capture a more holistic understanding of the environment, including the relationships between objects, their attributes, and even the higher-level context of a scene. Recent work in this area, leveraging open-vocabulary concepts [5, 9, 16] and multimodal inputs [1, 4, 13], has significantly enhanced robots' abilities to navigate and plan in novel environments. Our research explores introducing an additional dimension to scene representation by investigating how different robotic agents can meaningfully interact with those environments, specifically by inferring actionable affordances using VLMs. This bridges the gap between static scene understanding and dynamic, embodied interaction.

3 AFFORDANCE INFERENCE PIPELINE

This section details our experimental pipeline and demonstrates a potential application for VLM-driven affordance inference within multi-robot systems.

VLM Semantic-Affordance Inference. We begin by defining natural language prompts for the VLM with a description of the robot's physical properties. This enables the VLM to ground its understanding of affordances within the specific context of the robot's embodiment. We are interested in exploring the generalisable inference capabilities of these models, and therefore restrict these prompts to concise descriptions which could be applicable to any task domain. For instance, in our experiments, we provide descriptions such as:

"I am a cylindrical robot with dimensions 0.3m diameter and 0.5m height. I can move around the world in two dimensions. I have a mechanism which allows me to scoop objects from the ground if the object weighs less than 1 kilogram and is no wider than 0.3m."

For our experiments with non-humanoid robot morphologies, we consider a heterogenous set of six robots with distinct capabilities: Scoop, Push, Lift, Cut, Pick, and Collect. With a view towards future works where this pipeline will be deployed on real-world multi-robot collaboration problems, we take the DOTS [15] robots as a base, and envisage modifications to facilitate these capabilities. A complete set of robot prompts is provided in the associated datasets linked in Section 8.

As the robot navigates its environment, the on-board imaging system captures visual data. Frames are sampled at rate n to balance computational load with the need for fresh environmental

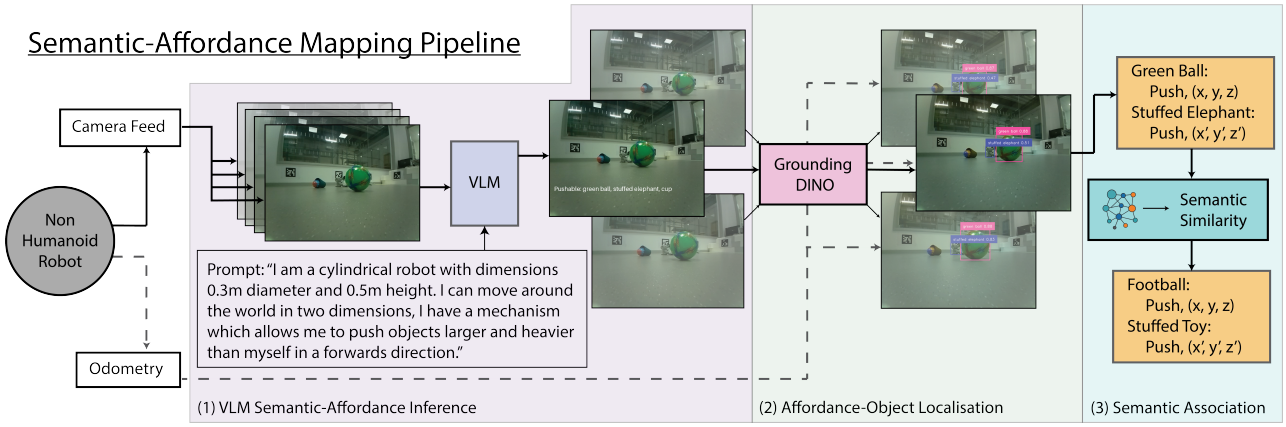


Figure 1: An illustration of our Semantic-Affordance Mapping pipeline applied to a non-humanoid robot (1) The camera feed of the robot is sampled every n frames and provided as input to the VLM alongside a natural language description of the physical properties of the robot. The VLM is tasked to identify objects in the frame and the affordances available to the robot, given its physical description. A structured output of affordance-object relations is returned. (2) Identified objects are passed to GroundingDINO with frame n and two frames either side $(n-k, n+k)$ to establish bounding boxes in the image plane. These bounding boxes are used with the recorded robot odometry to triangulate object positions in world space, and form the semantic-affordance mapping tuple: ⟨Object, Affordance, Location⟩. (3) Semantic similarity measures are performed against existing objects in the scene graph within distance d to consolidate variance in the VLM object labels.

information. We choose $n = 24$ to capture fresh information every second. Each sampled image along with the robot's specific physical description is then transmitted to the VLM via an API call. The VLM outputs explicitly identify detected objects within the sampled image and provide corresponding affordance characterisations relevant to the robot's described capabilities. Outputs are formatted as dictionaries of individual affordances, and a corresponding list of identified objects which afford this capability.

Localisation. To anchor the VLM's 2D image-based object-affordance inferences into a unified, global 3D reference frame, we employ a multi-view localisation strategy. For each VLM-identified object and its associated affordances, we leverage nearby frames captured by the on-board camera, specifically, given a detection event in frame n_i we sample frames $n_i \pm k$. We set $k = 0.5*$frame-rate. The semantic object descriptions generated by the VLM (e.g. "cup," "plastic pipe") are passed to GroundingDINO [19], a zero-shot object detection model, to generate 2D bounding boxes for these objects across the sequence of frames. By combining these 2D detections with the robot's odometry and camera calibration parameters, we triangulate the global 3D position of each object.

Semantic-Association. The inherent stochastic nature of generative model outputs can lead to inconsistencies where the same physical object might be identified with slightly varying labels (e.g. "green ball" vs. "football") or conflicting affordance predictions across different observations (e.g. "grasp", "grip-able"). To mitigate this and ensure the construction of a robust, consistent overall scene representation, we implement a semantic-association mechanism. This mechanism continuously evaluates the semantic similarity of newly detected objects and their affordances against previously labelled entities within a specified spatial proximity d of the current detection's global position. In our experiments, we assign $d <$ L2 norm threshold of 0.1.

Specifically, when a new object-affordance pair is localised, sentence transformer embeddings are generated for both object and affordance labels, and compared via cosine similarity to those of existing objects in the scene graph that fall within distance d. If semantic similarity exceeds a predefined threshold ($\tau = 0.45$) and the global positions of the objects are sufficiently close, the new detection is associated with the existing object. This fusion process resolves conflicting labels by, for instance, merging "cup" and "mug" detections into a single entity or by consolidating "pickable" and "graspable" affordances for the same object. This ensures that the scene representation is coherent, reduces noise from transient VLM outputs, and facilitates long-term mapping and consistent interaction planning.

Source	Unique Objects	Instances
Real	14	246
Synthetic	86	528
Total	100	774

Table 1: Composition of our novel dataset. The table presents a breakdown of the 774 labelled object instances, categorized by their real and synthetic origins. This combined dataset forms the basis for all experimental evaluations.

4 EVALUATION METHODOLOGY

With this work, we provide a preliminary novel hybrid dataset specifically designed for affordance characterisation in the context of non-humanoid robot morphologies. The dataset is presented not as a comprehensive collection but as an initial contribution to

Humanoid Robot		
Affordance	**Unique Objects**	**Instances**
Lift	93	736
Pick	92	735
Push	93	736
Non-Humanoid Robots		
Affordance	**Unique Objects**	**Instances**
Scoop	76	499
Lift	1	11
Collect	76	507
Push	92	735
Pick	78	533
Cut	39	253

Table 2: Distribution of affordance-object pairs in the dataset. The table itemizes both unique pairs and total instances, which sum to 4,745 instances overall. For analytical clarity, data for the baseline humanoid robot is presented separately from the six non-humanoid robots.

the study of affordance inference for non-humanoid robots. This collection comprises both synthetic and real-world video sequences, captured from the on-board perspectives of non-humanoid robots in environments consisting of everyday objects relevant to diverse real-world scenarios. The dataset is curated to reflect the challenging environments we explore: household settings, construction sites, logistics warehouses, and environmental clean-up. One hundred unique objects have been selected for diversity and include items such as: Book, Metal Rod, Table Tennis Bat, Stuffed Toy, Tissue, and Pebble. A summary of the unique object counts, and labelled object instances within the dataset is presented in Table 1, and an overview of the affordance-object pairs in the dataset is presented in Table 2. For both synthetic and real-world datasets, affordance ground truth is established through human annotation. Links to the complete dataset are provided in section 8.

Synthetic Data. Our synthetic material comprises a collection of videos generated using Google's Veo3 model. These videos feature a wide range of objects across a diverse set of environments from the on-board perspective of a robot with non-humanoid morphology. We also include a collection of isolated synthetic images of individual objects from these contexts, designed for direct VLM analysis of affordance capabilities independent of complex scene clutter. All synthetic materials are manually labelled to mitigate some of the risks of inherent bias, and we observe a difference of ±0.05 in mean object-affordance inference scores between the real and synthetic datasets, indicating that any bias introduced by the generative models is negligible in the scope of this work.

Real-World Data. To assess the practical applicability and generalisation of our findings, we provide real-world videos captured using the DOTS robots. The material is captured in a controlled but cluttered indoor arena environment, designed to mimic elements of the target scenarios.

Our evaluation across these datasets demonstrates the intrinsic capabilities of the VLM-driven affordance characterisation pipeline across a broad spectrum of controlled conditions and diverse robot morphologies. The real-world materials serve to validate that these capabilities translate effectively to tangible robotic systems within a physical environment, and with limited imaging and processing capabilities.

We consider three VLMs in our assessment using the providers' default hyper-parameters and conduct five independent trials to identify standard deviation in the results. The models selected are noted by their respective API keys, and represent the latest image-capable public releases as of September 2025:

- GPT: gpt-5
- Gemini: gemini-2.5-pro
- Claude: claude-opus-4-1-20250805

We examine affordances for six independent non-humanoid robot morphologies, each with unique action capabilities: Scoop, Push, Pick, Lift, Cut, Collect. The VLMs are prompted with a simple description of a given robot and its unique physiological properties, for example, the 'Lifting' robot:

"I am a cylindrical robot with dimensions 0.3m diameter and 0.5m height. I can move around the world in two dimensions. I have a mechanism which allows me to lift objects vertically upwards as long as I can get underneath them."

In addition, we consider a humanoid robot as a baseline for comparison. To provide a direct comparison for this baseline, we constrain our evaluation labels for the humanoid to three capabilities (Push, Pick, Lift) which reflect the capabilities of three of our six non-humanoid robots. We provide a minimal prompt to the VLMs describing this robot:

"I am a humanoid robot, I have two arms and two legs, and have the same capabilities as a human."

All experiments were conducted offline on an Apple M3 Pro with 18GB memory. A full package list with versions is provided with the code repository linked in Section 8. The DOTS robot [15] hardware was used for the real data collection.

4.1 Metrics

Our evaluation methodology assesses the performance of the VLMs in accurately inferring affordance labels and corresponding objects. For each sampled frame $f \in \mathcal{F}$, where $\mathcal{F}$ is the set of all evaluated frames, VLM annotations are compared against human-annotated ground truth labels.

Annotation Representation. Let an individual annotation, whether ground truth or VLM-generated, be represented as a tuple (a, O), where:

- a is one of the affordance labels associated with a given robot morphology.
- $O = \{o_1, o_2, \ldots, o_m\}$ is a set of m object labels associated with the affordance a.

For a given frame f, let $\mathcal{GT}_f = \{(a_{GT,i}, O_{GT,i})\}_{i=1}^{N_{GT}}$ be the set of N_{GT} ground truth annotations representing unique affordances and corresponding object relations, and $\mathcal{VLM}_f = \{(a_{VLM,j}, O_{VLM,j})\}_{j=1}^{N_{VLM}}$ be the set of N_{VLM} VLM-predicted annotations.

Semantic Similarity and Pairwise Validation. For each sampled frame $f \in \mathcal{F}$: Let $\text{sim}(l_1, l_2)$ be the semantic similarity function

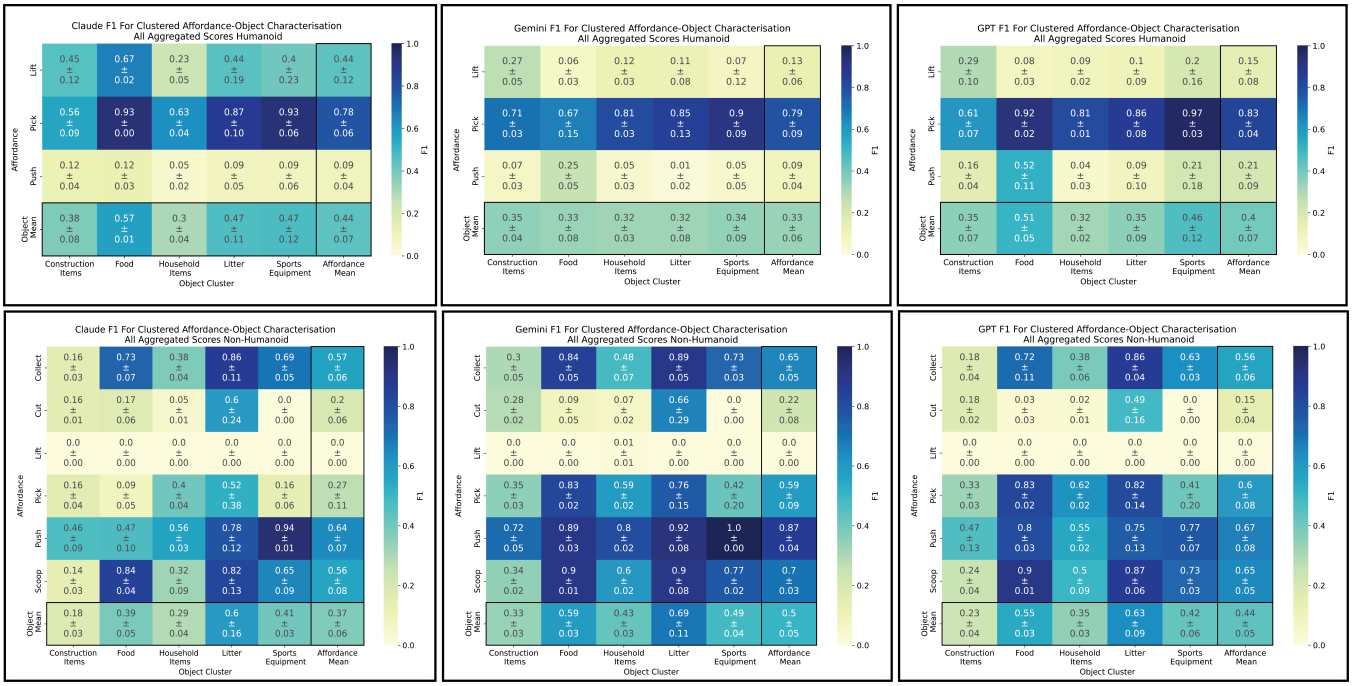


Figure 2: Affordance-object inference F1 scores and standard deviation over five independent trials. Objects have been clustered into five distinct groups. The top row describes the results produced by the baseline humanoid robot, and the bottom row the non-humanoid robots. Of the three affordances we examine, the humanoid baseline captures the 'Pick' affordance, synonymous with 'Grasp', but tends to default to out-of-scope human-like affordances in place of 'Lift' in the case of Gemini and GPT and 'Push' across all models, leading to a high-false negative rate and lower overall performance. Improvements are observed in non-humanoid robots for the 'Push' affordance, indicating that constrained descriptions of the robots capabilities are beneficial for novel tool use. Notably, Claude's inference capabilities of the 'Pick' affordance dramatically diminishes for non-humanoid robots, suggesting a lesser capacity for embodiment generalisation. The non-humanoid robot which affords 'Lift' underperforms due to a criterion of objects being above the robot chassis, exposing the weaknesses in spatial awareness and generalised context awareness.

between two labels l_1 and l_2 (e.g. cosine similarity of word embeddings), yielding a value in $[0, 1]$. We use the SBERT 'all-MiniLM-L6-v2' sentence transformer to generate our word embeddings, and by sampling a selection of typical VLM affordance assignments we find a semantic similarity threshold of $\tau = 0.45$ strikes a suitable balance between precision and recall.

Evaluation Process. Evaluation for each VLM prompt response corresponds to a specific robot morphology within the frame.

- Let $\mathcal{A}_{GT}^{(f,r)}$ be the list of ground truth affordance labels relevant to a specific robot morphology r in frame f.
- Let $O_{GT}^{(f,a)}$ be the list of ground-truth object labels relevant to each affordance $a \in \mathcal{A}_{GT}^{(f,r)}$.
- Let $\mathcal{A}_{VLM}^{(f,r)}$ be the list of affordance labels predicted by the VLM, relevant to a specific robot morphology r in frame f.
- Let $O_{VLM}^{(f,a)}$ be the list of object labels associated by the VLM, relevant to each affordance $a \in \mathcal{A}_{VLM}^{(f,r)}$.

A triangular similarity matrix $S_a^{(f)}$ is computed, where each element $S_a^{(f)}[i, j]$ represents the semantic similarity between the i-th ground truth affordance in $\mathcal{A}_{GT}^{(f,r)}$ and the j-th VLM-predicted affordance in $\mathcal{A}_{VLM}^{(f,r)}$: $S_a^{(f,r)}[i, j] = \text{sim}(\mathcal{A}_{GT}^{(f,r)}[i], \mathcal{A}_{VLM}^{(f,r)}[j])$.

A set of matched affordance pairs $\mathcal{M}_a^{(f,r)}$ is identified. Each pair $(i, j) \in \mathcal{M}_a^{(f,r)}$ indicates that the i-th ground truth affordance in $\mathcal{A}_{GT}^{(f,r)}$ has a semantically similar counterpart in the j-th VLM-predicted affordance in $\mathcal{A}_{VLM}^{(f,r)}$: $\mathcal{M}_a^{(f,r)} = \{(i, j) \mid S_a^{(f,r)}[i, j] > \tau\}$ For each such match, the corresponding VLM-predicted affordance is denoted as $\hat{a}_{VLM}^{(f)}$.

Object matching is then computed, conditional on affordance matching: For each matched affordance $\hat{a}_{VLM}^{(f)}$: A triangular similarity matrix $S_o^{(f,\hat{a})}$ is computed. This matrix compares the ground truth objects associated with this affordance $O_{GT}^{(f,\hat{a})}$ against the objects predicted by the VLM. Each element $S_o^{(f,\hat{a})}[k, l]$ represents the semantic similarity between the k-th ground truth object in $O_{GT}^{(f,\hat{a})}$ and the l-th VLM-predicted object in $O_{VLM}^{(f,\hat{a})}$: $S_o^{(f,\hat{a})}[k, l] = \text{sim}(O_{GT}^{(f,\hat{a})}[k], O_{VLM}^{(f,\hat{a})}[l])$.

We are interested in understanding the performance of the VLMs in correctly identifying when an object does and does not afford the capabilities of a given robot morphology. In our analysis we

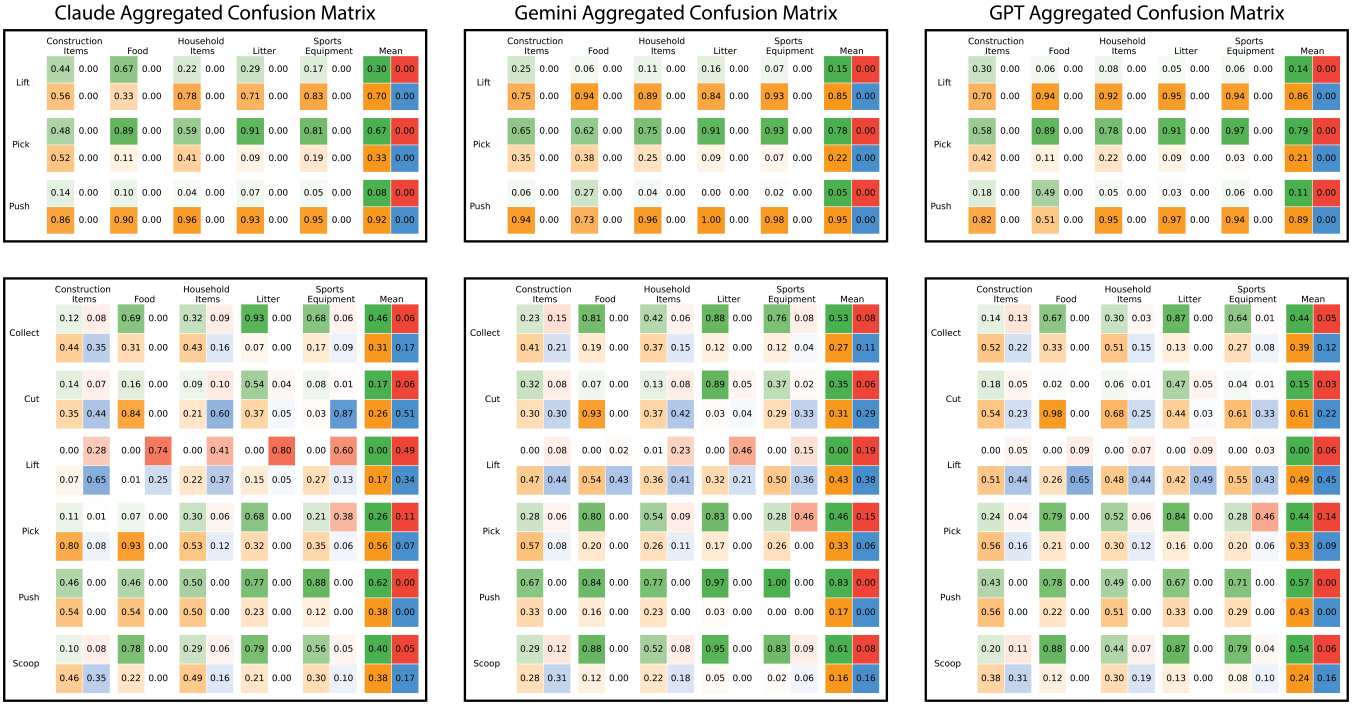


Figure 3: Confusion matrices across the three VLMs for aggregated performance of True-Positive (green), False-Positive (red), True-Negative (blue), False-Negative (orange) across the clustered object classes. Opacity in columns (excluding the mean) reflect the weight of the respective score. As with Figure 2, the humanoid robot baseline is presented on the top row, and the non-humanoid robots on the bottom. Note that results are predominantly weighted towards true-positive (correct affordance characterisation), or false negative (conservative bias) for most affordance-object relations with notable outliers being the 'Lift' and 'Cut' affordances for the non-humanoid robots.

consider the full confusion matrix of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) across all frames $f \in \mathcal{F}$. Understanding the balance of precision and recall is important to this assessment, therefore we select the F1 score as our evaluation metric, computed as:

$$F_1 = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

5 RESULTS AND DISCUSSION

In this section, we analyse the affordance inference performance of the three selected VLMs. Our analysis covers the baseline humanoid robot and the six non-humanoid robot morphologies. Aggregated F1 scores and detailed confusion matrices are presented in Figure 2 and Figure 3, respectively, with data clustered into five distinct object categories for clarity. Individual affordance-object scores are available in the data linked in Section 8. In Figure 4 we present frames from the synthetic and real world video data to visualise the performance classifications in our experiments. For the humanoid baseline, the models achieved moderate overall performance, with Claude (0.53) and GPT (0.51) showing better mean affordance-object scores vs. Gemini (0.36). There was a clear performance disparity across the affordances, where all models were generally successful at inferring the 'Pick' action but struggled significantly with 'Push'. For the non-humanoid robots, affordance characterisation was more varied, with Gemini showing the strongest overall results with a mean F1 score of 0.5. The non-humanoid results highlight a clear trend where affordances explicitly linked to the robot's physical form, such as 'Scoop' and 'Push', were identified far more successfully than actions like 'Cut' and 'Lift', which were defined with more complex spatial and material constraints.

Humanoid Baseline. The VLMs showed the strongest performance when inferring the 'Pick' affordance, which is synonymous with the common human action of grasping. This was particularly evident for object classes like 'Food' and 'Sports Equipment', which are frequently associated with human interaction in the models training data.

However, an important observation was made across all affordances. Instead of identifying the labelled capabilities, the models frequently defaulted to generating a wide range of plausible, but out-of-scope, human-like actions such as:

- "Throw", ["Tennis Ball", "Rugby Ball"]
- "Squeeze", ["Gray Stuffed Animal"]

Detailed example response logs are provided in the data linked in Section 8. This tendency to default to general human-centric world knowledge resulted in a high rate of false negatives across all models for the examined affordances (Figure 3, top row). We

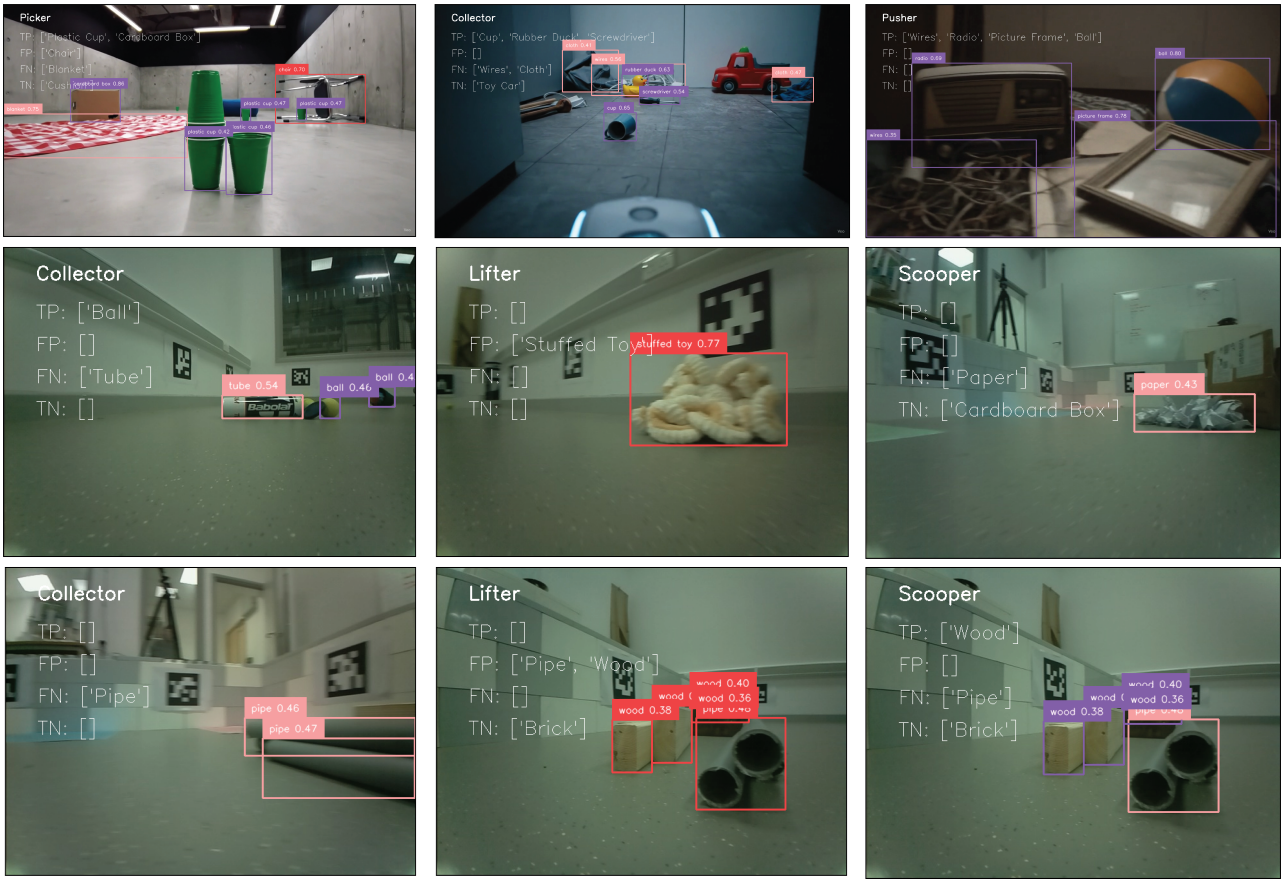


Figure 4: Examples of VLM semantic-affordance inference mapped to bounding boxes with GroundingDINO. The top row illustrates examples from the synthetic video dataset, and the middle and bottom rows from the real world data; household items, and construction materials respectively. Affordance and performance classifications have been overlayed to provide evaluation insights in these cases. True-positives are bounded by purple boxes, false-positives by red boxes, and true-negatives by pink boxes.

suggest this reveals an inherent bias: the models will omit a correct, task-relevant affordance if a more common but out-of-scope action seems applicable from their training. This observation requires further investigation and is discussed in more detail in Section 7.

Non-Humanoid Robots. In contrast, the non-humanoid robots showed a marked improvement on the equivalent 'Push' affordance (Figure 2, bottom row). This affordance is clearly inferred from the physical description of the robot: "*I have a mechanism which allows me to push objects larger and heavier than myself in a forwards direction.*", but crucially, the models generated significantly fewer out-of-scope affordances. The case is similar for robots with tools to 'Scoop' and 'Collect'. This indicates that providing a concise physical description of the robot successfully constrains inference, grounding its reasoning in the specific capabilities of the robot. This grounding helps overcome the tendency to revert to general human actions, leading to more accurate and relevant affordance characterization.

However, there is also notable trend of false-negatives across the non-humanoid robots (Figure 3, bottom row). Given the lack of out-of-scope affordance characterisations, this indicates that models tend towards a conservative bias, where they are more likely to assign no affordance than an incorrect affordance. This tendency preserves a low false positive rate, which is beneficial for safety, but raises an important consideration for the effective application of VLMs in scenarios with diverse robot morphologies and unconventional applications.

Despite the benefits of physical grounding, our results also highlight key limitations in the VLMs' reasoning, particularly when faced with unconventional scenarios and novel tool use.

A significant weakness in spatial reasoning was revealed by the models attribution of the 'Lift' affordance. Even though this action was explicitly constrained to objects above the non-humanoid robot's chassis, Claude and GPT incorrectly identified items on the ground as 'liftable' leading to high rates of false-positives (Figure 3, bottom row). This inability to apply a general concept to a novel physical context demonstrates a fundamental lack of spatial and contextual understanding. Interestingly, this is less evident in the

Gemini results, where the model either generated a false-negative (conservative bias) or a true-negative (correct interpretation).

Inspection of the individual affordance-object scores revealed that models also showed weaknesses in constraining inference when considering the 'Cut' affordance. This robot description constrained the valid cut-able materials to paper, plastic or wood. However, there was a tendency across the board to assign cut to other materials types, including 'Golf Ball', 'Screwdriver', 'Paint Pot'. This is perhaps indicative of a lack of materials understanding, and warrants further enquiry.

We also observe notable performance inconsistencies across different object classes (Figure 2, bottom row). F1 scores were generally higher for 'Food', 'Litter', and 'Sports Equipment', but considerably lower for categories like 'Construction Items'. This further supports the hypothesis that this is a result of the models human-centric training data, where the vast corpus of images depicting humans interacting with food and sports equipment allows for better generalization. Conversely, the models limited exposure to robotic manipulation in domains like construction hinders their ability to infer affordances for less common object interactions, revealing a clear boundary in their inference capabilities.

6 CONCLUSION

In summary, for a humanoid robot, the VLMs performed well on common actions like 'Pick'. However, they frequently suggest plausible human actions (e.g. "Throw" a ball). This resulted in a high rate of false negatives, where the correct action was overlooked in favour of a more common, but out-of-scope, human one. This insight highlights the value of considering alternative evaluation techniques, where VLM outputs are validated by a human annotator, with the caveat that consideration must be given to the significant human resourcing required to effectively do this.

In contrast, for non-humanoid robots, providing a clear physical description (e.g. "... I have a mechanism to push objects ...") successfully grounded the models' reasoning, leading to fewer out-of-scope suggestions. This improvement came with a trade-off: the models adopted a conservative bias, often returning no affordances rather than inferring an incorrect one, which also led to false negatives. The research also highlighted several critical weaknesses in the underlying inference capabilities of the VLMs:

Poor Spatial Reasoning. Models incorrectly identified objects on the ground as 'liftable' for a robot that could only lift objects above itself, demonstrating a failure to apply spatial constraints.

Lack of Material Understanding. Models suggested that a robot could 'Cut' inappropriate materials like a golf ball or screwdriver, ignoring specified materials constraints.

Training Data Bias. Performance was consistently better with familiar object categories like 'Food' and 'Sports Equipment' and worse with 'Construction Items', indicating that the models' abilities are limited by their human-focused training data.

7 FUTURE WORKS

Our work proposes using VLMs to support semantic-affordance inference and enhance robot scene representation. In future work these representations can be used to ground multi-robot collaboration, enabling teams of diverse robots to collaborate to solve problems by reasoning about their shared environment and their respective contribution capacity. While we have shown the potential of this approach, our findings also highlight key limitations of using VLMs for affordance inference, presenting clear opportunities for additional streams of research.

Task Constraints. We have observed that VLMs often suggest actions that are plausible in a general sense but irrelevant to the task at hand. Since affordances are inherently task-dependent, a crucial next step is to explore task-conditioned prompting. By supplementing the robot's physical description with a high-level task description (e.g. "tidy the room"), we can investigate whether VLMs can better constrain their inference to output only the most relevant affordances. A preliminary probe, involving a very simple task description on our real data set, shows an immediate uplift in average object-affordance mean of between $0.03 - 0.1$ for the non-humanoid robots, indicating scope to expand this line of enquiry further.

Physical Embodiments. Development of the heterogenous capabilities as an extension to the base DOTS platform is noted as important future work for forthcoming research and real-world execution prototyping.

Validation Approach. We also note that there is value in exploring alternative methods to evaluate the outputs of generative AI models. Instead of providing a set of 'gold-standard' label outputs, which we want the model to generate, we suggest each generated output is validated by a human for the correctness of the affordance characterisation. This would provide insights to understand the extent of noise in the labels and the potential for more complex, 'unexpected' inference capabilities. Naturally, annotation resources and consistency are a potentially limiting factor of this evaluation approach, as significant human validation efforts would be required.

Inference Speed. A significant bottleneck for real-world deployment is slow inference speed, with API calls in our experiments taking three to ten seconds, prohibiting real-time updates. Future work should focus on accelerating inference, with promising avenues including local, fine-tuned models and model distillation.

Dataset Expansion. Finally, our study relied on a dataset created specifically for this investigation, which could be further improved and expanded. To accelerate progress, the field requires a large-scale, public benchmark dataset for non-humanoid robot affordance inference. A well-annotated community benchmark would enable rigorous, standardised evaluation of different models and propel the development of more generalizable and robust systems.

8 CODE & DATA REPOSITORIES

The code and data repository links are made available here: https://bitbucket.org/hauertlab/vlm_driven_non_human_sem_aff

ACKNOWLEDGMENTS

JJ is support by FARSCOPE EPSRC CDT. RSR is funded by the UKRI Turing AI Fellowship [grant number EP/V024817/1]. SH is supported by an EPSRC Open Plus Fellowship.

REFERENCES

[1] Michael Ahn, Debidatta Dwibedi, Chelsea Finn, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Karol Hausman, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, et al. 2024. Autort: Embodied foundation models for large scale orchestration of robotic agents. *arXiv preprint arXiv:2401.12963* (2024).

[2] Paola Ardón, Èric Pairet, Katrin S Lohan, Subramanian Ramamoorthy, and Ronald Petrick. 2020. Affordances in robotic tasks–a survey. *arXiv preprint arXiv:2004.07400* (2020).

[3] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. 2023. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13778–13790.

[4] Matthew Chang, Theophile Gervet, Mukul Khanna, Sriram Yenamandra, Dhruv Shah, So Yeon Min, Kavit Shah, Chris Paxton, Saurabh Gupta, Dhruv Batra, Roozbeh Mottaghi, Jitendra Malik, and Devendra Singh Chaplot. 2023. GOAT: GO to Any Thing. arXiv:2311.06430 [cs.RO] https://arxiv.org/abs/2311.06430

[5] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S Ryoo, Austin Stone, and Daniel Kappler. 2023. Open-vocabulary queryable scene representations for real world planning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 11509–11522.

[6] Dongpan Chen, Dehui Kong, Jinghua Li, Shaofan Wang, and Baocai Yin. 2023. A survey of visual affordance recognition based on deep learning. *IEEE Transactions on Big Data* 9, 6 (2023), 1458–1476.

[7] Ophelia Deroy, Davide Bacciu, Bahador Bahrami, Cosimo Della Santina, and Sabine Hauert. 2024. Shared Awareness Across Domain-Specific Artificial Intelligence: An Alternative to Domain-General Intelligence and Artificial Consciousness. *Advanced Intelligent Systems* 6, 10 (2024), 2300740.

[8] Thanh-Toan Do, Anh Nguyen, and Ian Reid. 2018. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 5882–5889.

[9] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. 2024. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401* (2024).

[10] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214* (2024).

[11] James J Gibson. 2014. *The ecological approach to visual perception: classic edition*. Psychology press.

[12] James J Gibson. 2014. The theory of affordances:(1979). In *The people, place, and space reader*. Routledge, 56–60.

[13] Jiayuan Gu, Sean Kirmani, Paul Wohlhart, Yao Lu, Montserrat Gonzalez Arenas, Kanishka Rao, Wenhao Yu, Chuyuan Fu, Keerthana Gopalakrishnan, Zhuo Xu, et al. 2023. Rt-trajectory: Robotic task generalization via hindsight trajectory sketches. *arXiv preprint arXiv:2311.01977* (2023).

[14] Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and José Santos-Victor. 2016. Affordances in psychology, neuroscience, and robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems* 10, 1 (2016), 4–25.

[15] Simon Jones, Emma Milner, Mahesh Sooriyabandara, and Sabine Hauert. 2022. DOTS: An open testbed for industrial swarm robotic solutions. *arXiv preprint arXiv:2203.13809* (2022).

[16] Christina Kassab, Matias Mattamala, Lintong Zhang, and Maurice Fallon. 2024. Language-extended indoor slam (lexis): A versatile system for real-time visual scene understanding. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 15988–15994.

[17] Olivia Y Lee, Annie Xie, Kuan Fang, Karl Pertsch, and Chelsea Finn. 2024. Affordance-guided reinforcement learning via visual prompting. *arXiv preprint arXiv:2407.10341* (2024).

[18] Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. 2024. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3086–3096.

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*. Springer, 38–55.

[20] Raphaël Millière and Charles Rathkopf. 2024. Anthropocentric bias and the possibility of artificial cognition. In *ICML 2024 Workshop on LLMs and Cognition*.

[21] Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. 2023. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5607–5612.

[22] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. 2024. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7587–7597.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[24] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. 2024. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721* (2024).