Language Grounding with 3D Objects

Anonymous Author(s)
Affiliation
Address
email

Abstract:
Seemingly simple natural language requests to a robot are generally underspecified, for example *Can you bring me the wireless mouse?* When viewing mice on the shelf, the number of buttons or presence of a wire may not be visible from certain angles or positions. Flat images of candidate mice may not provide the discriminative information needed for *wireless*. The world, and objects in it, are not flat images but complex 3D shapes. If a human requests an object based on any of its basic properties, such as color, shape, or texture, robots should perform the necessary exploration to accomplish the task. In particular, while substantial effort and progress has been made on understanding explicitly visual attributes like color and category, comparatively little progress has been made on understanding language about shapes and contours. In this work, we introduce a novel reasoning task that targets both visual and non-visual language about 3D objects. Our new benchmark **S**hape**N**et **A**nnotated with **R**eferring **E**xpressions (SNARE) requires a model to choose which of two objects is being referenced by a natural language description. We introduce several CLIP-based [1] models for distinguishing objects and demonstrate that while recent advances in jointly modeling vision and language are useful for robotic language understanding, it is still the case that these models are weaker at understanding the 3D nature of objects – properties which play a key role in manipulation. In particular, we find that adding view estimation to language grounding models improves accuracy on both SNARE and when identifying objects referred to in language on a robot platform.

Keywords: Benchmark, Language Grounding, Vision, 3D

1 Introduction

Joint language and vision models are often trained on image captions which have a bias towards canonically "visual" attributes of objects, such as color, rather than functional ones like shape. Image captions omit properties that require understanding objects as 3D, rather than flat, concepts. In this work, we show how the effect of this disconnect is that while robotics research has benefited greatly from advances in computer vision, vision-and-language models cannot always be directly applied to robotics. People use geometric and physical properties of objects when describing them. For example, a robot carrying out a request to *Bring me the mug with the wide handle* needs to be able to spot the *wide handle*, even if the mug rests on a countertop with the handle oriented out of view, but a robot can rotate such an object to investigate in further detail. Generally, identifying objects based on natural language is more challenging from non-canonical viewpoints, which are common for robots in home, office, and industrial environments.

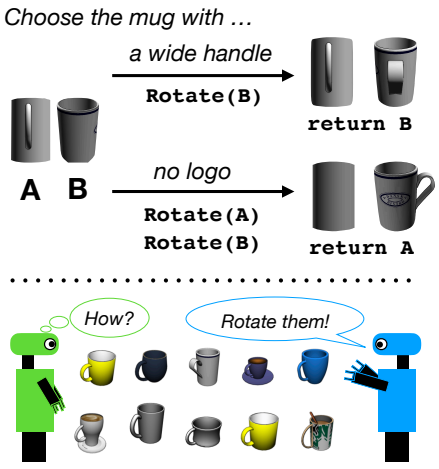


Figure 1: Deciding between 3D objects described in natural language can require rotating objects to see multiple faces.

Submitted to the 5th Conference on Robot Learning (CoRL 2021). Do not distribute.

The codification and commodification of ResNet [2] architectures pretrained on ImageNet [3] has yielded immense progress in off-the-shelf computer vision techniques. Such representations provide a strong foundation for visual reasoning. With the introduction of pretrained Transformer [4] architectures like BERT [5], natural language processing has similar off-the-shelf tools for representing language. Modern visual language grounding is done through massive, internet-scale pretraining to combine these strengths: aligning images with their captions and content descriptions in natural language using Transformer models [6, 1, 7, 8].

However, internet images suffer from a type of *reporting bias*. These images are captured by sighted humans, from human-centric vantage points [9]. There is a domain shift between such canonical depictions of objects in images and objects as seen through a robot camera [10, 11], as studied in computer vision between such images and those captured by the blind [12]. For example, images of *mugs* online almost exclusively feature a full or three-fourths view of the handle, while a mug in the wild can easily be oriented such that the handle is occluded from the camera by the mug body.

A person requesting a mug with a wide handle has a *mental model* of the referent object that includes its full 3D spatial information; humans do not imagine objects only in 2 dimensions. Further, someone trying to retrieve said mug will differentiate it from other mugs on the countertop by viewing the mugs from different angles to inspect the handles.

Our goal is to similarly imbue robots with three dimensional notions of language grounding, by encouraging them to *rotate* an object when selecting the referent of a language description. We introduce the **S**hape**N**et **A**nnotated with **R**eferring **E**xpressions (SNARE) dataset, which provides discriminative natural language descriptions of 3D ShapeNet [13] object models. SNARE includes both *visual* descriptions that focus on colors and object categories and *blindfolded* descriptions that focus on shapes and parts (Section 3). We train agents to select the correct object given its description, and find that estimating the current view and performing rotation on objects to obtain an additional vantage point improves accuracy (Figure 1).

In particular, we introduce **La**nguage **G**rounding through **O**bject **R**otation (LAGOR) which performs view estimation as an auxiliary loss to encourage 3D object understanding (Section 4) during the SNARE task. We find that using multiple views to select the correct object improves over single-view object selection, using the large-scale vision-and-language CLIP [1] model as a backbone to score how well an image and language MATCH. We show that enabling agents to rotate objects to a new view, while performing auxiliary view estimation, can result in higher accuracy at selecting referent objects than consuming panoramic views of objects while simultaneously being more realistic for a physical robot platform.

Our key contributions are:

- SNARE, a benchmark dataset for identifying 3D object models given natural language referring expressions including *visual* and *blindfolded* descriptions;
- Baseline models for SNARE demonstrating both zero-shot and fine-tuned performance of state-of-the-art vision-and-language models; and
- LAGOR, an initial SNARE model that estimates current view, performs rotations, then makes a referent prediction, which we demonstrate on a physical robot platform.

2 Related Work

Using natural language to work with robot partners is a long-standing goal in robotics [14]. We argue that internet-scale, pretrained vision-and-language models offer a powerful starting point for human-robot collaboration. Unlike 2D internet images, physical objects can be picked up and moved by robot agents. Agents can perform information-seeking behaviors on objects to improve their world [15, 16, 17] and language understanding [18, 19]. Such world interaction is inextricable from language grounding [20], motivating language annotations for higher-fidelity referents than static images. We introduce SNARE, comprised of language referring expressions for 3D objects, and the LAGOR model to combine language and multiple object views to achieve better 3D understanding. SNARE extends lines of work tying language to static images, 3D object models, and even physical objects. We demonstrate that LAGOR generalizes to physical objects manipulated by a tabletop robot §5.2. LAGOR is inspired by models that perform information-seeking actions informed by learned world models to better achieve goals.

Image-based Object Identification Object classification [3, 2] is the first step towards image captioning and visual question answering [21]. Particular object instances can be found with region segmentation models such as MaskRCNN [22], enabling object referent tasks such as GuessWhat!? [23]. Combining visual recognition with Transformer-based language understanding to jointly attend to language and visual tokens leads to improved downstream performance on many vision-and-language tasks [6, 24, 8]. Joint embedding approaches that learn a shared subspace for representing language and vision tokens [25] also achieve state-of-the-art performance when trained at scale [1]. We show that these large-scale, pretrained models fall short of human performance on the SNARE task, and that synthesizing 2D views from multiple vantage points improves object identification performance.

Language Grounding in 3D Prior works have associated single-word attributes with 3D object models based on latent representations of 3D meshes [26]. To learn spatial language, vision-and-language navigation (VLN) [27] models infer navigation actions from instructions and visual observations in 3D simulated worlds. Such tasks can be extended to include simulated world interactions such as picking and placing objects [28] and using appliances and tools [29]. Models for these tasks extend Transformer representations to include action taking [30, 31] and invoke object classification methods to create a semantic understanding of the world [32]. SNARE poses a complementary challenge, providing data for selecting referent objects in the presence of distractors by taking into account the multiple views possible of objects in 3D space.

The work most similar to ours is ShapeGlot [33], a collection of referring expressions for ShapeNet objects to discriminate between two distractors. While ShapeGlot is similar in spirit, it contains training data only for `chair` objects, and tests on four additional categories. By contrast, SNARE spans 262 distinct object categories (Table 1). ShapeGlot focuses on a model's abilities to learn particular parts of objects, such as chair arms and legs, while SNARE aims to ground language more generally to 3D features of objects spanning color, shape, category, and parts. Further, ShapeGlot models take in a pointcloud representation of objects along with visual features, where our LAGOR model does not assume access to the 3D model of an object and instead operates solely on different 2D views achieved through object rotation.

Data	Fold	# Cats	# Objs	# Ref Exps
ShapeGlot	Chairs	1	4,511	78,789
	Other	4	200	400
	Total	5	4,711	79,189
SNARE	Train	207	6,153	39,104
	Val	7	371	2,304
	Test	48	1,357	8,751
	Total	262	7,881	50,159

Table 1: Fold summaries in SNARE, with the ShapeGlot (SG) benchmark size for contrast.

Physical Object Identification Modeling the connections between language to physical objects enables robots to identify objects for manipulation by color, shape, and category [34, 35]. The physical forces and sounds objects make during manipulation actions can also be associated with words such as *rattling* and *heavy* for multimodal understanding beyond vision [18, 36]. Prior work has gathered language annotations for the YCB Benchmark object set [37] to explore how language descriptions provide priors on object affordances [38]. In our tabletop robot experiments, we use camera views of novel objects to evaluate zero shot transfer of LAGOR to the real world with minimal object rotations to achieve language-aligned camera views to select the correct referent object.

Information-Seeking Actions Embodiment, whether in simulation or the physical world, affords agents the opportunity to seek out new information to help with a given task. By predicting the new information that can be gotten from different actions, agents can *prospect* over potential futures for planning [39, 40, 41]. In the aforementioned VLN task, predicting *what* will be seen along different potential routes facilitates more efficient navigation [42]. Our LAGOR model estimates the current object view before performing a rotation to obtain a new view, for example rotating mugs to view handles head-on when charged to find a *wide handle* (Figure 1).

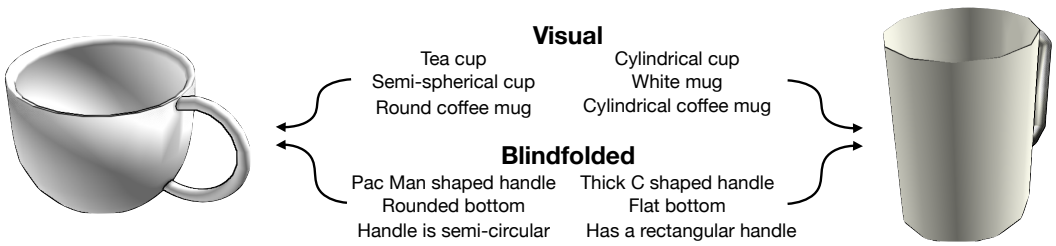


Figure 2: Example object pair and object referring expressions in SNARE. Given a referring expression and one (or more) views of two contrasting objects, a model must decide which object is being referenced by the language. Language annotations are collected in two forms: **Visual** and **Blindfolded**. The latter expressions tasked annotators with describing objects tactilely.

3 ShapeNet Annotated with Referring Expressions (SNARE)

We introduce SNARE, as a new benchmark for grounding natural language referring expressions to distinguish 3D objects. The annotations are collected to complement the ACRONYM[1] [43] grasping dataset and include language that targets both visual and tactile attributes of objects. Our goal is to enable future research to ground both from multi-view vision, as in this work, and directly from robotic grasp contact point data (Section 6).

To construct SNARE, we select a subset of 7,897 ACRONYM [43] object models from ShapeNet-Sem [44, 13]. We obtain over 50 thousand natural language referring expressions in English for these object models via Amazon Mechanical Turk (AMT).[2] Figure 2 gives an example of referring expressions collected for two `mug` objects used to distinguish between the two.

To elicit referring expressions with high specificity, we frame the annotation as a discriminative task. AMT workers were presented with two object models side-by-side from the same ShapeNet category, for example two different `OutdoorTable` object meshes, and asked to complete sentences like

- The way to tell Object A from Object B is that Object A looks like a(n) ___
- Blindfolded, the way to tell Object A from Object B is that Object B is a(n) ___

with referring expressions. We displayed the object models as GIFs rendered to give a 360 rotating view of each object. The resulting SNARE task asks models to take in one such referring expression and decide whether it applies to Object A or Object B.

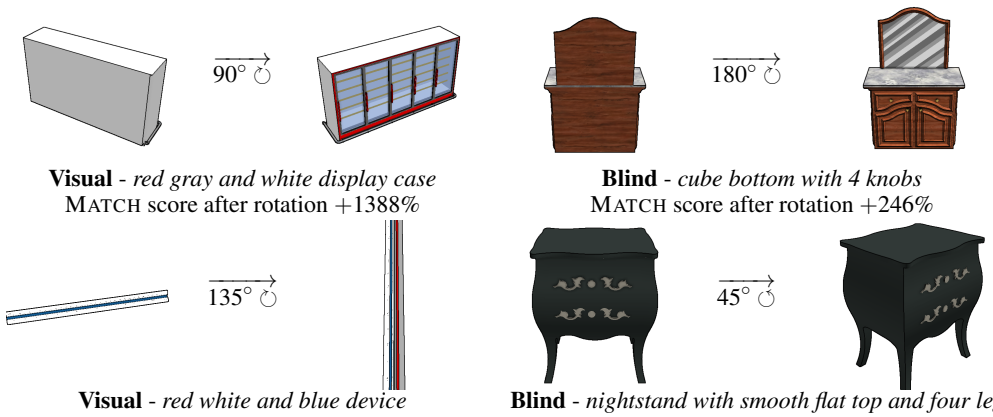
Priming in the annotation prompts invokes visual features (*... looks like...*) and non-visual features (*Blindfolded, ...*) in the referring expressions. By pairing objects from the same ShapeNet category, referring expressions must go beyond categorical information like *brown dresser* to differentiate one object from the other with higher specificity, similar to ShapeGlot's choice to use distractor objects from a single training category [33]. We collected six referring expressions per object—three primed to be visual and three primed to be blindfolded. Each referring expression was vetted through a secondary task on AMT where, given the language expression, workers had to correctly select the referent object. Every referring expression in SNARE was correctly associated to its referent object by a majority of such annotators.

SNARE referring expressions average 4.27 words, with blindfolded expressions longer (4.95) than visual expressions (3.63). Visual expressions focus more on color and object category words, such as *table*, than blindfolded expressions. By contrast, blindfolded expressions use more shape and part words, such as *rectangle* and *legs*. We estimated that blindfolded expressions use more shape words (14% vs 5% of all words) while visual use more color words (22% vs 1% of all words), by traversing the WordNet [45] hierarchy for each word and noting whether it is a hyponym of *color* or *shape*.

Each SNARE instance is a tuple of (referring expression, Object A, Object B), and models must select which of Object A or B the referent of the natural language expression. We split these data instances into train, validation, and test folds by ShapeNet category. We ensure that closely related categories such as `2Shelves` and `3Shelves` or `DiningTable` and `AccentTable` are within

[1] Yes, that dataset title is "ACRONYM" [43].

[2] Section 7.1 contains additional details about the AMT study.



Visual - *red gray and white display case*
MATCH score after rotation +1388%

Blind - *cube bottom with 4 knobs*
MATCH score after rotation +246%

Visual - *red white and blue device*
MATCH score after rotation +131%

Blind - *nightstand with smooth flat top and four legs*
MATCH score after rotation +205%

Figure 3: Example MATCH score increases after performing object rotation. Rotations involve exposing colors (top left, bottom left), canonical faces (top right), and parts (bottom right). A robot encountering objects in the wild will find initial object orientations that do not line up with natural language referring expressions for those objects, as reflected in SNARE.

the same fold. For example, all shelf-related and bed-related categories are sorted into the train and test folds, respectively, so that inter-category information does not leak across folds. More details can be found Section 7.3. Table 1 summarizes the overall data statistics.

4 Methods

There has been a recent proliferation of off-the-shelf, increasingly large-scale pretrained vision and language alignment models applicable for language grounding in robotics [1, 6, 7, 8, 46]. We set out to answer two questions about such models, using SNARE as a testbed:

- Can existing language and vision models ground language in the SNARE benchmark?
- Do SNARE models generalize to a robot object selection task better than off-the-shelf models?

To answer the first question, we first train a MATCH module that learns a predictive head on top of a CLIP [1] backbone, and evaluate this module on single and multiple views of 3D objects. To answer the second question, we introduce the LAGOR model, which uses view estimation as an auxiliary loss while predicting language expression referent objects for SNARE, and evaluate on both SNARE §5.1 and a physical robot platform §5.2. LAGOR examines an initial 3D object view and an additional, post-rotation view, striking a balance between single-view models and those that attempt to capture each object's entire 3D structure.

4.1 Language-View Match (MATCH) Module

The MATCH module takes in a language expression L and single view of an object, V_i, and produces a match score $s(L, V_i)$. To decide whether Object A or Object B is the referent of a language expression on SNARE, we interpret $\text{argmax}_{O \in \{A,B\}} s(L, V_{i,O})$ as the model's prediction.

CLIP [1] serves as a backbone on which we add additional, learnable layers for the SNARE task. We use CLIP's transformer-based sentence encoder to extract language features $\mathbf{l} : \mathbb{R}^{1\times512}$. We use CLIP ViT-B/32 to extract visual features for image $V_{i,A}$, $\mathbf{v}_{\text{i,A}} : \mathbb{R}^{1\times512}$, and image $V_{i,B}$, $\mathbf{v}_{\text{i,B}} : \mathbb{R}^{1\times512}$. These modality-specific feature vectors are concatenated: $[\mathbf{v}_{\text{i,A}};\mathbf{l}] : \mathbb{R}^{1\times1024}$ and $[\mathbf{v}_{\text{i,B}};\mathbf{l}] : \mathbb{R}^{1\times1024}$ and independently run through a learnable, multi-layer perceptron that gradually reduces the dimensionality from 1024 to 512, then 256, and finally to a single-dimensional score s. This final score comparison training mirrors the multiple-choice formulation used in existing unimodal [5] and multimodal transformers [7]. We keep the CLIP encoders frozen during training.

ShapeNet objects are 3D meshes, incompatible with the 2D input expected by off-the-shelf vision and language models. We sample eight rendered viewpoints around each object at 45 degree increments,

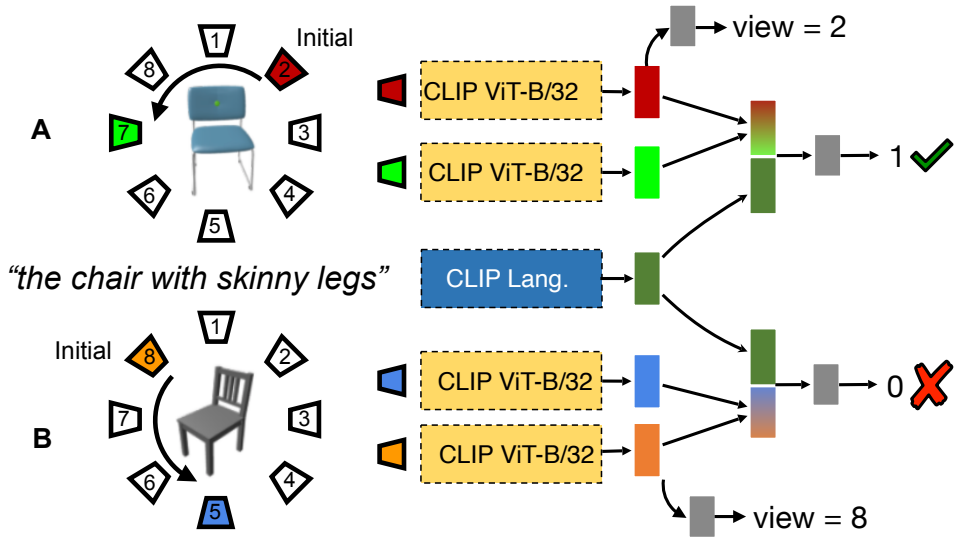


Figure 4: **LAGOR model** relies on the pretrained CLIP [1] architecture as a backbone encoder across multiple views compared to the encoded referring expression. In addition, the model sees substantial improvements from an auxiliary loss that predicts the orientation of the initial view.

and compare off-the-shelf CLIP and MATCH module accuracies when considering single or multiple of these views. To aggregate multiple views, we `maxpool` over view embedding vectors $\mathbf{v}_{\text{i,A}} \ldots \mathbf{v}_{\text{j,A}}$.

MATCH is trained with cross-entropy-loss, $\mathscr{L}_s$, to predict a binary label of whether the referring expression matches the object represented in the view image. MATCH is trained for 50 epochs on SNARE, with object views chosen at random during each step.

Baselines. We compare our fine-tuned MATCH against a zero-shot CLIP classifier. Instead of fine-tuning CLIP, we use the cosine distance between visual and language features to pick the referred object. That is, we select $\text{argmax}_{O \in \{A,B\}} \mathbf{l} \cdot \mathbf{v}_{\text{i,O}}$. We also evaluate against a trained ViLBERT baseline that consumes multiple object views at once, but find that it does not perform as well as the MATCH module; more details about the ViLBERT baseline can be found in Section 7.4.

4.2 Language Grounding through Object Rotation (LAGOR)

A robot tasked with retrieving an object given a natural language expression should not have to gather a 360 degree view of each candidate referent object to make a decision. We find that taking two random views of an object as input, as opposed to just one, nearly closes the gap with MATCH module performance on a 360 view of objects in SNARE (Section 5.1). Further, we have the intuition that estimating the *current* viewing angle can enable a model to develop a global reference frame for language grounding against 3D objects (Figure 3). Thus, we propose LAGOR, a model that takes in two views of each candidate object before selecting the language referent, and performs *view estimation* on candidate objects as an auxiliary prediction (Figure 4).

In addition to predicting the object referent using a pre-trained MATCH module that considers two views, LAGOR predicts the current view of each object using a cross-entropy loss, $\mathscr{L}_v$, against a 1-hot vector representing a discrete set of 8 views at 45 degree offsets. LAGOR learns a multi-layer perceptron that takes in a visual embedding V_i and reduces dimensionality from 512 to 256 to 128 to 64 to a vector of 8 logits representing the 8 discrete object views. The MATCH loss, $\mathscr{L}_s$, is combined with the view estimation loss for a final loss function $\mathscr{L} = \mathscr{L}_v + 0.2 * \mathscr{L}_s$. LAGOR is trained for 50 epochs on SNARE, with object views chosen at random during each step. An interesting direction for future work is predicting the *optimal*, rather than random, next view to rotate to when disambiguating a query, though that could introduce a cyclic dependency with the MATCH module.

4.3 Robot Demonstration

We evaluate LAGOR trained on SNARE on a robot platform, taking two random views of objects and evaluating the robot's ability to select which is the correct referent of a language expression

		Validation			Test		
Model	Views	Visual⊂	Blind⊂	All	Visual⊂	Blind⊂	All
ViLBERT	360	89.5	76.6	83.1	80.2	**73.0**	76.6
CLIP	360	84.5	66.1	75.3	80.0	61.4	70.9
MATCH	360	**90.6**	**79.3**	**85.0**	**85.9**	71.3	**78.7**
CLIP	Single	79.5	65.2	72.3	73.9	60.4	67.3
MATCH	Single	**89.4**	**75.6**	**82.5**	**84.1**	**69.6**	**77.0**
CLIP	Two	81.7	65.5	73.6	76.2	61.0	68.8
MATCH	Two	91.2	75.1	83.2	85.8	70.9	78.5
LAGOR	Two	**91.5**	**81.2**	**86.3**	**86.6**	**72.0**	**79.4**
Human (U)	360	94.0	90.6	92.3	93.4	88.9	91.2
Human (M)	360	100.0	100.0	100.0	100.0	100.0	100.0

Table 2: **Accuracy on the SNARE benchmark.** LAGOR outperforms off-the-shelf vision and language models, often even compared to those considering a 360 representation of the object, by adding a level of 3D object understanding in the form of a view estimation auxiliary loss. For multi-view models, we maxpool over visual embedding vectors from multiple views. CLIP rows are zero-shot performance without fine-tuning on SNARE, while MATCH rows use SNARE training data. Human accuracy is conservatively calculated as the number of SNARE instances for which the correct referent object was identified by *every* voting annotator (**U**nanimous); for all SNARE instances a majority of voting annotators correctly selected the referent (**M**ajority).

(Section 5.2). We compare LAGOR to the off-the-shelf CLIP model that considers a single view of the object only, before performing any view-gathering rotations.

In particular, we set two objects on the workspace of a Franka Emika Panda[3] with a wrist-mounted Intel Realsense D414.[4] We capture an initial view of the scene, then segment the objects from the workspace using the Unseen Object Clustering [47] algorithm. We feed the segmented objects as the initial views to LAGOR and CLIP. We move the arm to a second vantage point above the workspace and capture another image of the objects, repeat segmentation, and treat these segmentations as the second views for LAGOR.

These experiments introduce a number of new sources of noise not present in the simulated setting, for example changes in camera angle, shifting lighting during rotation, and automatic segmentation to account for cluttered scenes. In Section 5.2, we include a pair of examples where it is clear how each of these variables affects the final results. For example, segmentation leads to the creation of voids inside of objects not present in SNARE.

5 Results

While CLIP [1] and ViLBERT [6] are competitive baselines for many language and vision tasks, in this section we show that they select correct 3D object referents in SNARE substantially less often than the MATCH and LAGOR models. Then, we deploy CLIP and LAGOR on a robot tasked with selecting objects conditioned on natural language referring expressions, and find that view estimation and gathering additional object viewpoints improves referent identification.

5.1 LAGOR Performance on SNARE

Table 2 gives LAGOR accuracy compared to other two-view alternatives, as well as single view and 360 view models. We begin with comparisons to existing models trained (ViLBERT) and zero-shot (CLIP) as advertised and recommended in their respective papers. There are three key takeaways. First, LAGOR outperforms other two-view models, and the performance gap against MATCH with two views is solely due to the view estimation auxiliary loss. Thus, estimating initial

[3] https://www.franka.de/

[4] https://www.intelrealsense.com/depth-camera-d415/

object view appears to imbue LAGOR with some level of 3D object understanding that aids in the interpretation of how language corresponds to 3D object models. Second, LAGOR outperforms all 360 view models on nearly every metric, indicating that gathering *more* object views is not as helpful as understanding the 3D *relationship* between views. Third, and most importantly, while no models achieve human level accuracy, the performance difference is particularly striking on the blindfolded referring expression subset of data, which lags 5-15% behind visual referring expressions across models. That gap supports our intuition that blindfolded referring expressions capture a complementary and challenging linguistic space currently understudied for vision-language models but key to everyday manipulation. Thereby, SNARE opens interesting avenues for future work exploring the shape and grasp-points of objects.

5.2 Robot Results

We run single view, zero-shot CLIP against LAGOR on 8 pairs of objects with language descriptions. CLIP is able to select the correct object 3/8 times (38% accuracy), while LAGOR is correct 4/8 times (50% accuracy). Figure 5 highlights a success and a failure case for LAGOR. LAGOR succeeds over CLIP on shape and part descriptions, such as *thin metal handle* and *the one with a lid*. A task at which CLIP succeeds but LAGOR fails is optical character recognition (OCR); CLIP is known to "read" labels, enabling it to identify *the juice carton*, labeled with "juice", while LAGOR's training on SNARE caused this ability to be forgotten. Additional examples and discussions are given in Section 7.5.

Figure 5: (Top) Success: LAGOR correctly identifies the thin handle of the red mug. (Bottom) Failure: LAGOR fails to build a good representation of the more abstract concept "open".

6 Conclusions

We introduce **S**hape**N**et **A**nnotated with **R**eferring **E**xpressions (SNARE), a challenge task to ground language to 3D object models. We show that fine-tuned, massively pretrained vision and language models fall short of human performance at identifying object referents of natural language expressions by a wide margin (Table 2). While models improve as more 2D views of 3D objects are available, a robot agent can only take in one at a time. We introduce **La**nguage **G**rounding through **O**bject **R**otation (LAGOR), a model that performs a rotation operation on a candidate referent object to achieve a more language-aligned view (Figure 1). We find that LAGOR is able to identify referent objects using only an initial and final view more accurately than models utilizing 360 views. LAGOR achieves this result by performing view estimation as an auxiliary loss, enabling learning latent 3D structure. Finally, we transfer the trained LAGOR model to a physical robot and demonstrate its ability to select language-aligned object views and better select referent objects after performing informed rotations.

In the future, a rotation *policy* could be learned to examine up to N views across M candidate objects subject to a referring expression, rather than performing a random rotation to a novel view. By assigning rotation actions expected discriminative values using a trained rotation module, and penalties based on the time it takes a particular hardware to achieve an object rotation, one could learn a POMDP policy aggregating object views seen so far to decide whether to next obtain a new view or make a guess about the referent object. Obtaining different object views could be done either by picking up and rotating the object or by moving the robot base to see the object from another angle (as in this paper). Further, since 3D meshes are available for each object, it may be possible to extract shape-level information using a PointNet [48], similar to experiments attempted by ShapeGlot [33]. Because we target physical robot applications, we may be able to tie language to sets of *graspable points* encoding gripper orientation and position [43] for each object.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *arXiv*, 2021.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Computer Vision and Pattern Recognition (CVPR)*, 2009.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Neural Information Processing Systems (NeurIPS)*, 2017.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[6] J. Lu, D. Batra, D. Parikh, and S. Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neural Information Processing Systems (NeurIPS)*, pages 13–23, 2019.

[7] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*, 2020.

[8] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv*, 2019.

[9] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[10] T. Kanade and M. Hebert. First-person vision. *Proceedings of the IEEE*, 100(8):2442–2453, 2012.

[11] 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 (CoRL)*, 2020.

[12] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[13] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[14] S. Tellex, N. Gopalan, H. Kress-Gazit, and C. Matuszek. Robots that use language. *The Annual Review of Control, Robotics, and Autonomous Systems*, 15, 2020.

[15] K. Takahashi, T. Ogata, J. Nakanishi, G. Cheng, and S. Sugano. Dynamic motion learning for multi-dof flexible-joint robots using active–passive motor babbling through deep learning. *Advanced Robotics*, 31(18):1002–1015, 2017.

[16] A. Mousavian, C. Eppner, and D. Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *International Conference on Computer Vision (ICCV)*, pages 2901–2910, 2019.

[17] T. Nagarajan and K. Grauman. Learning affordance landscapes for interaction exploration in 3D environments. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[18] J. Thomason, A. Padmakumar, J. Sinapov, N. Walker, Y. Jiang, H. Yedidsion, J. Hart, P. Stone, and R. J. Mooney. Jointly improving parsing and perception for natural language commands through human-robot dialog. *The Journal of Artificial Intelligence Research (JAIR)*, 67, 2020.

[19] C. Lynch and P. Sermanet. Grounding language in play. In *arXiv*, 2020.

[20] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020.

[21] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[22] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. *International Conference on Computer Vision (ICCV)*, 2017.

[23] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. Guess-what?! visual object discovery through multi-modal dialogue. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee. 12-in-1: Multi-task vision and language representation learning. In *Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[25] C. W. Leong and R. Mihalcea. Going beyond text: A hybrid image-text approach for measuring word relatedness. In *International Joint Conference on Natural Language Processing (IJCNLP)*, 2011.

[26] V. Cohen, B. Burchfiel, T. Nguyen, N. Gopalan, S. Tellex, and G. Konidaris. Grounding language attributes to objects using bayesian eigenobjects. In *Intelligent Robots and Systems (IROS)*, 2019.

[27] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[28] J. Abramson, A. Ahuja, A. Brussee, F. Carnevale, M. Cassin, S. Clark, A. Dudzik, P. Georgiev, A. Guy, T. Harley, F. Hill, A. Hung, Z. Kenton, J. Landon, T. Lillicrap, K. Mathewson, A. Muldal, A. Santoro, N. Savinov, V. Varma, G. Wayne, N. Wong, C. Yan, and R. Zhu. Imitating interactive intelligence. *arXiv*, 2020.

[29] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[30] A. Majumdar, A. Shrivastava, S. Lee, P. Anderson, D. Parikh, and D. Batra. Improving vision-and-language navigation with image-text pairs from the web. In *European Conference on Computer Vision (ECCV)*, 2020.

[31] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould. A recurrent vision-and-language BERT for navigation. *arXiv*, 2020.

[32] K. P. Singh, S. Bhambri, B. Kim, R. Mottaghi, and J. Choi. MOCA: A modular object-centric approach for interactive instruction following. *arXiv*, 2020.

[33] P. Achlioptas, J. Fan, R. X. Hawkins, N. D. Goodman, and L. J. Guibas. ShapeGlot: Learning language for shape differentiation. In *International Conference on Computer Vision (ICCV)*, 2019.

[34] M. Shridhar, D. Mittal, and D. Hsu. INGRESS: Interactive visual grounding of referring expressions. *The International Journal of Robotics Research (IJRR)*, 39(2-3):217–232, 2020.

[35] T. Kollar, S. Tellex, M. R. Walter, A. S. Huang, A. Bachrach, S. Hemachandra, E. Brunskill, A. Banerjee, D. Roy, S. Teller, and N. Roy. Generalized grounding graphs: A probabilistic framework for understanding grounded commands. *arXiv*, 2017.

[36] J. Thomason, J. Sinapov, M. Svetlik, P. Stone, and R. J. Mooney. Learning multi-modal grounded linguistic semantics by playing "I spy". In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[37] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar. Yale-cmu-berkeley dataset for robotic manipulation research. *The International Journal of Robotics Research (IJRR)*, 36(3):261–268, 2017.

[38] R. Scalise, J. Thomason, Y. Bisk, and S. Srinivasa. Improving robot success detection using static object data. In *Intelligent Robots and Systems (IROS)*, 2019.

[39] X. Zhang, J. Sinapov, and S. Zhang. Planning multimodal exploratory actions for online robot attribute learning. In *Robotics: Science and Systems (RSS)*, 2021.

[40] S. Nair, S. Savarese, and C. Finn. Goal-aware prediction: Learning to model what matters. In *International Conference on Machine Learning (ICML)*, 2020.

[41] C. Paxton, Y. Bisk, J. Thomason, A. Byravan, and D. Fox. Prospection: Interpretable plans from language by predicting the future. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[42] J. Y. Koh, H. Lee, Y. Yang, J. Baldridge, and P. Anderson. Pathdreamer: A world model for indoor navigation. *arXiv*, 2021.

[43] C. Eppner, A. Mousavian, and D. Fox. ACRONYM: A large-scale grasp dataset based on simulation. In *2021 IEEE Int. Conf. on Robotics and Automation, ICRA*, 2020.

[44] A. X. Chang, M. Savva, and P. Hanrahan. Semantically-enriched 3d models for common-sense knowledge. In *FPIC Workshop, Computer Vision and Pattern Recognition (CVPR)*, 2015.

[45] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

[46] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. VinVL: Making visual representations matter in vision-language models. *Computer Vision and Pattern Recognition (CVPR)*, 2021.

[47] Y. Xiang, C. Xie, A. Mousavian, and D. Fox. Learning rgb-d feature embeddings for unseen object instance segmentation. In *Conference on Robot Learning (CoRL)*, 2020.

[48] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Neural Information Processing Systems (NeurIPS)*, 2017.

[49] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[50] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Association for Computational Linguistics (ACL)*, 2018.

[51] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.