A Spoken Language Dataset of Descriptions for Speech-Based Grounded Language Learning

Gaoussou Youssouf Kebe,[1]* Padraig Higgins,[1]* Patrick Jenkins,[1]* Kasra Darvish,[1] Rishabh Sachdeva,[1] Ryan Barron,[1] John Winder,[1,2] Don Engel,[1] Edward Raff,[1,3] Francis Ferraro,[1] Cynthia Matuszek[1]
[1]University of Maryland, Baltimore County
[2]Johns Hopkins Applied Physics Laboratory
[3]Booz Allen Hamilton

Abstract

Grounded language acquisition is a major area of research combining aspects of natural language processing, computer vision, and signal processing, compounded by domain issues requiring sample efficiency and other deployment constraints. In this work, we present a multimodal dataset of RGB+depth objects with spoken as well as textual descriptions. We analyze the differences between the two types of descriptive language and our experiments demonstrate that the different modalities affect learning. This will enable researchers studying the intersection of robotics, NLP, and HCI to better investigate how the multiple modalities of image, depth, text, speech, and transcription interact, as well as how differences in the vernacular of these modalities impact results.

1 Introduction

Grounded language acquisition is the process of learning language as it relates to the world—how concepts in language refer to objects, tasks, and environments [46]. *Embodied* language learning specifically is a significant field of research in NLP, machine learning, and robotics. While there is substantial current effort on learning grounded language for embodied agents [11, 28, 63], in this work we describe learning from multiple modalities, including text, transcribed speech, and speech audio.

Text is a common input domain when learning grounded language, yet many systems use speech once deployed [75]. In practice, embodied agents are likely to need to operate on imperfectly understood spoken language. Speech-based assistive devices have gained significant popularity in the last few years, representing perhaps the first widely deployed, communicative 'embodied agents' in human environments. Spoken language is critical for interactions in physical contexts, despite the inherent difficulties: spoken sentences tend to be less well framed than written text, with more disfluencies and grammatical flaws [56].

There are many ways in which robots learn grounded language [12, 16, 30, 43, 73, 76, 80], but they all require either multimodal data or natural language data—usually both. Current approaches to grounded language learning require data in both the perceptual ("grounded") and linguistic domains. While existing datasets have been used for this purpose [16, 31, 33, 51, 74], the language component is almost always derived from either textual input or manually transcribed speech [44, 73].

To that end, we present the **Gro**unded **L**anguage **D**ataset (GoLD), which contains images of common household objects and their description in multiple formats: text, speech (audio), and speech transcriptions (see fig. 1). The primary contributions of this paper are as follows:

*Equal contributions from first three authors

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

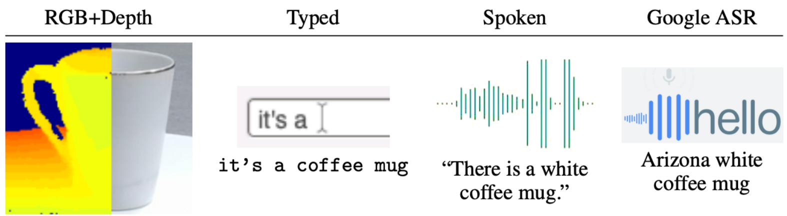


Figure 1: GOLD has RGB and depth point cloud images of 207 objects in 47 categories. It includes 16500 text and 16500 speech descriptions; all spoken descriptions include automatic transcriptions.

1. We provide a publicly available, multimodal, multi-labelled dataset of household objects, with image+depth data and textual and spoken descriptions with automated transcription.
2. We show that learning language groundings from transcribed or raw speech performs similarly to models trained on typed text, while allowing those descriptions to be provided in a more natural, convenient way.
3. We demonstrate that the dataset poses a number of interesting research challenges including identifying bias in speech processing from the unique perspective of language grounding.

2 Related Work

Language acquisition can be used for interactions with robots [2, 10, 43, 73]. On a robot, the grounded language acquisition task has a number of uses. Retrieving objects based on their descriptions [52] is a necessary component of caretaking and domestic robots. Grounding landmarks and instructions can aid robots in navigation of novel spaces [73, 79]. [71] surveyed the many machine learning methods used, possible applications, and the human-robotic interaction implications of grounded language learning on a robotic platform.

Grounded language acquisition is in the unique position of requiring a dataset that combines sensory perception with language. These combined datasets are frequently handcrafted for the specific task that the research seeks to accomplish [12, 59], often leading to narrower applications. For example, CLEVR [31] was designed as a benchmark for question answering tasks. Objects in CLEVR are limited to a small set of attributes which in turn limits the types of questions in both syntax and content. In comparison, GOLD contains more complex real-world objects and does not limit the scope of the annotations to a fixed set of characteristics.

We note that the image component of GOLD is heavily influenced by the University of Washington RGB-D dataset [35]. Both datasets contain large numbers of everyday objects from multiple angles. Our dataset is collected with a now state of the art sensor which enables us to capture smaller objects at a finer level of detail (such as an Allen key, the diameter of which pushes the limits of the depth sensor when laid on the table). Additionally, we select objects based on their potential utility for specific human-robot interaction scenarios, such as things a person might find in a medicine cabinet or first aid kit, enabling learning research relevant to eldercare and emergency situations [8].

Creating a dataset that includes speech has a high cost of collecting and transcribing audio. [59] presents a grounded language system that can generate descriptions for targets within a scene of colored rectangles. The visual data for this task is easily generated, but speech descriptions were recorded and transcribed over a long period of time. The manual audio transcription task can take four to ten hours per hour of audio [21, 85]. Such perfectly transcribed audio is also unrealistic for real-world usage, which must rely on automation. We acknowledge this challenge, and we evaluate automatically-produced transcriptions for their quality. We also include the automatically-produced transcriptions along with the raw audio.

Recent datasets that include speech such as Flickr Audio Captions [24], SpokenCOCO [29], SPEECH-COCO [26], Synthetically Spoken COCO, Synthetically Spoken STAIR get around this by generating spoken descriptions from the text captions provided by the Flickr8K, COCO [38], and STAIR [86] datasets. Speech COCO, Synthetically Spoken COCO [15], and Synthetically Spoken STAIR [27] generate their speech through text to speech systems while Flickr Audio Captions and SpokenCOCO use crowdsourced workers. Places Audio Captions [25] which uses the MIT Places 205 Database [87]

Table 1: Classes of objects in GOLD.

Topic	Classes of Objects
food	*potato, soda bottle, water bottle, apple, banana, bell pepper, food can, food jar, lemon, lime, onion*
home	*book, can opener, eye glasses, fork, shampoo, sponge, spoon, toothbrush, toothpaste, bowl, cap, cell phone, coffee mug, hand towel, tissue box, plate*
medical	*band aid, gauze, medicine bottle, pill cutter, prescription medicine bottle, syringe*
office	*mouse, pencil, picture frame, scissors, stapler, marker, notebook*
tool	*Allen wrench, hammer, measuring tape, pliers, screwdriver, lightbulb*

is the only other dataset in this area where the speech is collected directly from the spoken descriptions of crowd workers, however the descriptions are of all the salient objects in an image instead of a single object. All these datasets also only contain color images while GOLD extends this to include depth images and pointclouds.

In our work we adopt the manifold alignment model form [49] which is similar to [52]. The latter trained a grounded language model in order to retrieve objects with a robotic arm from natural language descriptions. The robot learned the functionality of objects through text data gathered from Wikipedia.

Embodied approaches [2, 72] are important for collecting multimodal data on robotic platforms. [77] created a robot that learned from both language and sensed traits including the visual, proprioceptive, and auditory characteristics of objects. However, the language was used only to identify named objects. [4] developed a robot that memorized which objects it had seen before by combining multimodal data about the object including visual, haptic, and researcher provided linguistic percepts.

Recently, manifold alignment has been used and outperformed traditional classification methods, particularly for grounded language tasks [9, 49, 52]. One particular benefit of manifold alignment is that it enables arbitrary embeddings to be used and aligned. In contrast to prior grounding approaches, these embeddings do not have to be restricted to individual words, and instead can be computed for an entire input (e.g., utterance). As a result, we use *grounded language manifold alignment* techniques to experimentally validate GOLD.

3 GOLD: The Grounded Language Dataset

GOLD is a collection of visual and English natural language data in five high-level groupings: *food*, *home*, *medical*, *office*, and *tools*. In these groups, 47 object classes (see table 1) contain 207 individual object instances. The dataset contains vision and depth images of each object from 450 different rotational views. From these, four representative 'keyframe' images were selected. These representative images were used to collect 16500 textual and 16500 spoken descriptions. The dataset contents are summarized in table 2.

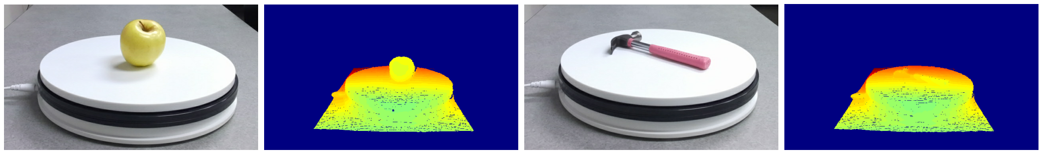
Table 2: Components of GOLD.

Categories (*e.g.*, medicine)	5	Images (vision + depth)	825
Classes (*e.g.*, apple)	47	Text descriptions	16500
Object instances (*e.g.*, `apple_3`)	207	Spoken descriptions	16500

Visual inputs were collected by rotating objects on a turntable in front of a commodity RGB-D (RGB + depth) video camera, as in [35]). For each object, four keyframes were manually selected to capture representative, diverse view angles of each object. Amazon Mechanical Turk workers were shown all four images and asked to provide either spoken or typed descriptions.

3.1 Vision + Depth Data Collection

Visual perception data were collected using a Microsoft Azure Kinect (i.e.., a Kinect 3), a low-cost, high-fidelity commodity sensor that is widely used in robotics. For each object instance (*i.e.*, for each of the four staplers in the dataset), this sensor was used to collect raw image and point cloud videos.



(a) Apple image frame. (b) Apple point cloud. (c) Hammer image frame. (d) Hammer point cloud.

Figure 2: Samples showing keyframes in GoLD, along with the aligned 3D point cloud with depth information. Only the RGB image was shown to labelers.

The resulting dataset contains 207 90-second depth videos, one per instance, showing the object performing one complete rotation on a turntable. To ensure that each object has diverse views, e.g., examples of a mug with the handle occluded and visible, we manually selected 825 pairs of image and depth point cloud from 207 objects as representative frames, which we refer to as keyframes (examples are shown in fig. 2).

Manually selecting keyframes avoids a known problem with many visual datasets: their tendency to show pictures of objects taken from a limited set of 'typical' angles [7]. For example, it is rare for a picture of a banana to be taken end-on. This aligns with our motivation of creating a dataset of household objects to support research on grounded language learning in an unstaged environment, as a robot looking at an object in a home may not see this typical view.

3.2 Text and Speech Description Collection

All descriptions were collected using Amazon Mechanical Turk (AMT).[2] Keyframes for randomly-chosen object instances were shown to the worker. They were asked to either type descriptions of objects in one or two short, complete sentences, or record descriptions using a microphone.

Collected speech was transcribed using Google's Speech to Text API, resulting in a spoken-language corpus of 16500 verbal descriptions. It should be noted that, although Mechanical Turk does not provide personally identifiable information about workers, it is possible that users may be identified by their voice or other side-channel information. For this reason, all collected language is limited to factual descriptions of simple household objects, and no value judgments, opinions, or emotional or potentially damaging subjects are discussed.

Quality	Value	Count
Perceived Gender	Men	271
	Women	274
	Undet.	7
Accent	Yes	279
	No	273
Creak	Yes	194
	No	358
Hoarseness	Yes	48
	No	504
Muffledness	1	393
	2	119
	3	40
Volume	1	10
	2	157
	3	331
	4	54
Background Noise	1	366
	2	143
	3	39
	4	4

Table 3: Number of workers labeled with each characteristic.

3.2.1 Speaker Voice Qualities

We collected spoken descriptions from 552 Amazon Mechanical Turk workers. We labeled each of these workers based on perceived gender (man, woman, or undetermined),[3] accent (whether the speaker has a non-mid-American accent), creak (whether the user has a raspy, low-register voice), hoarseness (whether the speaker has a strained, breathy voice), muffled-ness (the level of distortion of the user's microphone, 1 to 3), volume (1 to 4), and level of background noise (1 to 4). Section 3.2 shows the number of workers to whom each label has been attributed.

We intend for this data to be used as a test-bed for bias studies and other research into the performance of grounding models for different sub-populations. For example, a pilot study on this data has shown that accented users are particularly affected by the bias of speech-to-text models and that learning directly from raw speech can mitigate this bias.

[2] See Ethical Considerations section, appendix.

[3] Gender and sex are complex constructs. We asked annotators to choose the category that seemed to 'best describe' the speaker, but acknowledge the limitations of this approach.

3.2.2 Accuracy of Speech Transcriptions

Obtaining accurate transcriptions of speech in sometimes noisy environments is a significant obstacle to speech-based interfaces [37]. To understand the degree to which learning is affected by ASR errors, 250 randomly selected transcriptions were manually evaluated on a 4-point scale (see table 4). Of those, 80% are high quality ('perfect' or 'pretty good'), while only 11% are rated 'unusable.'

To get a more detailed understanding of transcription accuracy, we compare the ASR transcriptions and the human-provided transcriptions using the standard word error rate (WER) [55] and Bilingual Evaluation Understudy (BLEU) [54] scores. BLEU scores are widely used to measure the accuracy of language translations based on string similarity; we adopt this system to evaluate the goodness of transcriptions. BLEU is calculated by finding n-gram overlaps between machine translation and reference translations. We use tri-grams for our BLEU scores since some descriptions are shorter than four words such as "these are pliers", rendering a 4-gram BLEU score meaningless.

Table 4: Human ratings of 250 automatic transcriptions. These ratings are designed strictly to assess the accuracy of the transcription, not the correctness of the spoken description with respect to the described object.

Rating	Transcription Quality Guidelines	#
1	wrong or gibberish / unusable sound file	28
2	slightly wrong (missing keywords / concepts)	21
3	pretty good (main object correctly defined)	33
4	perfect (accurate transcription and no errors)	168

Figure 3 shows that many of the 250 manually transcribed descriptions were perfectly transcribed by automated speech-to-text. The marginal BLEU histogram shows more mistaken transcriptions (the second peak around 0) due to known problems with using BLEU to evaluate short sentences and tokens having mismatched capitalization or punctuation.

3.3 Comparative Analysis

Our initial hypothesis was that people would use more words when describing objects verbally than when typing, as it is lower effort to talk than to type. Accordingly, We find spoken descriptions to be slightly longer than their textual counterparts ($p \geq 13.71$ using a Welch's t-test) While speech has more average words per description, 11.7, compared to text at 10.46, when stop words are removed the averages are 6.1 and 5.89 respectively. The larger mean drop in the speech descriptions is likely due to the tendency of ASR systems to interpret noise or murmur utterances as filler words, the inclusion of which has been shown to detract from meaning [68]. Text descriptions are a more consistent length than speech, with a standard deviation of 6.7 words for text, versus 9.51 for transcribed speech. When we remove stop words, the standard deviation is 3.63 for text and 4.69 for speech.

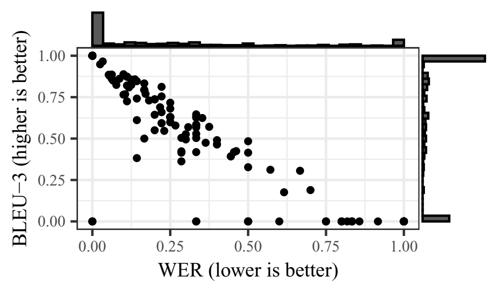


Figure 3: BLEU-3 and WER scores for 250 randomly selected speech transcriptions. A WER of 0 and a BLEU of 1 (top left corner) indicates perfect transcription. Marginal histograms show that some descriptions were perfectly transcribed.

Table 5 shows the top 20 most frequent words in both modalities. There is substantial overlap, as expected, since the same objects are being described, with colors dominating the lists. People use more filler words when describing the objects using speech; for example, the word 'like' appears 889 times in speech data whereas it was not significant in the text data.

Using the Stanford Part-of-Speech Tagger [78] to count the number of nouns, adjectives, and verbs between the two modalities yields no significant differences between the modalities. However, the word 'used' appears frequently (see table 5), typically to describe functionality. Developing grounded language models around functionality for the analysis of affordances in objects [52] is an important research avenue that our dataset enables, which is not possible in prior datasets that do not contain the requisite modalities.

3.4 Dataset Distribution

The data is publicly available as a GitHub repository[4]. The repository contains three high-level datatypes: perception and language. The perceptual data is split into RGB-D images and depth data in the form of point clouds [60]. Each of these sets of data is subdivided by object class (e.g., "apple") and then further by instance (e.g., "apple #4"). The language is subdivided similarly, and for each object instance contains multiple speech descriptions (as .wav files) along with ASR transcriptions of that speech. Each instance also has multiple associated typed descriptions, which are not related to the spoken descriptions—they were provided by different workers at a separate time.

Each description of an instance also includes associated meta-data describing the data collection process. This includes: (1) a numeric identifier for the worker who provided each description; (2) the amount of time each description took to provide; and (3) the ground-truth category and instance label for each object.

Table 5: Top 20 most frequently used words in text (left) and speech (right) by percentage of occurrence in descriptions.

Token	% Frequency
black	13.24
white	10.66
blue	9.97
bottle	9.50
red	9.45
yellow	9.02
object	7.99
small	6.44
green	5.82
pair	5.27
used	5.21
handle	4.58
plastic	4.40
silver	3.88
box	3.69
label	2.92
metal	2.79
pink	2.66
light	2.44
scissors	2.43

Token	% Frequency
black	13.92
white	12.85
blue	10.23
red	9.13
yellow	8.97
bottle	8.50
small	7.96
used	7.21
object	6.41
green	5.85
plastic	5.30
color	5.22
handle	4.85
like	4.62
looks	3.99
silver	3.66
turntable	3.33
pair	3.32
box	3.21
label	3.01

4 Experiments

GoLD is designed to enable multiple research directions. In our evaluation we will demonstrate initial baseline results for classification, retrieval, and speech recognition tasks that are enabled by GoLD. Each experiment will combine the RGB+depth images with one of the three language domains: text, transcribed speech, and speech audio. We also perform a fourth learning experiment on a combination of text and transcribed speech to test how the combination of the two might boost learning. In each case the goal is to learn how to ground the unconstrained natural language descriptions of objects with their associated visual percepts of color and depth. This allows research investigating the impact of information lost via reductions from raw speech, to text, to noisier ASR text. The textual inputs naturally lack the inflection and tonal characteristics that will be critical for user interaction with a robot, but not easily studied with current datasets. Since speech is a natural mode of communication for humans, and information such as inflection are lost after transcription, we would like to move in a direction where speech audio is the primary input into our models, forgoing transcription entirely.

Manifold Alignment. As noted in section 2, we use manifold alignment [3, 82, 83] with triplet loss [6, 49] to embed the visual percepts and language data from GoLD into a shared lower dimensional space. Within this space, a distance metric is applied to embedded feature vectors in order to tell how well a particular utterance describes an image, with shorter distances implying a better description. The manifold alignment model is shown in fig. 4.

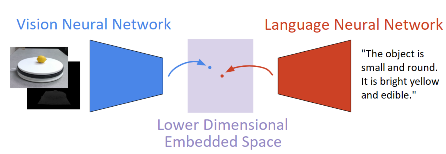


Figure 4: A high-level view of a manifold alignment model. Vision and language domains are embedded into a shared lower dimensional space. Pairs of vision and text are aligned to be closer to each other within the embedded space.

For example, a picture of a lemon and the description "The object is small and round. It is bright yellow and edible." should be closer together in the embedded space than the same picture of a lemon and the unrelated description "This tool is used to drive nails into wood," since the latter description was used to describe a hammer. Through this technique, even novel vision or language inputs should be aligned, meaning that a new or different description of a lemon should still be closely aligned in the embedded space. We would additionally expect other similar objects, such as an orange, to be described in a somewhat similar way, allowing for potential future learning of categorical information.

[4] https://github.com/iral-lab/gold

Vision. The vision feature vectors are created following the work of [20] and [58]. Color and depth images are passed through CNNs that have been pretrained on ImageNet [17] with the prediction layer removed so that the final layer is a learned feature vector. Depth images are "colorized" to enable classification re-using the same network [58]. The two vectors, one from color and one from depth, are concatenated into a 4096-dimensional visual feature vector.

Text and Speech Transcriptions. The language features of text and transcribed speech data are extracted using BERT [18]. Each natural language description is fed to a BERT pretrained model. We obtain the final embedding by concatenating the last four hidden layers of BERT. The resulting 3072-dimensional vector is taken as the description's language feature vector and associated to the visual feature vector of the frame it describes.

Speech. Self supervised pretrained models inspired by NLP methods have recently shown success in speech representation. We use wav2vec 2.0 [5], a self-supervised speech model that learns over continuous representations of raw speech through a BERT [18] inspired masked language modeling task. Similarly to the text featurization, features are then learned by performing average-pooling over the concatenation of the last four layers of the transformer.

To evaluate the benefit of using a pre-trained model, we also consider 40 dimensional Mel-frequency cepstral coefficient (MFCC) features [47] that are extracted from the raw audio with a 10 ms frame shift. Due to the lower-dimensional nature of MFCCs, the language network is modified to include a Long Short-Term Memory (LSTM) network. 64-dimensional outputs from the final 32 hidden states [13] are concatenated together to form a fixed length 2048-dimensional speech vector which are passed to a fully connected layer and output into the same embedded dimension as the visual network.

Triplet Loss. The triplet loss function [6, 62] uses one training example as an "anchor" and two more points, one of which is in the same class as the anchor (the positive), and one which is not (the negative). For example, while classifying tools the anchor might be a hammer, the positive would be a different hammer, and the negative would be a screwdriver. The loss function then encourages the network to align the anchor and positive in the embedded space while repelling the anchor and the negative. In order to align the networks to each other and keep each network internally consistent, the anchor, positive, and negative instances are chosen randomly (balanced across cases) from either the vision or language domains at training time.

For anchors (A), positive instances (P), and negative instances (N), we compute embeddings of these points, then compute triplet loss in the standard fashion with a default margin $\alpha = 0.4$ [62] where f is the relevant model for the domain of the input:

$$\mathcal{L} = \max(0, \|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + \alpha) \quad (1)$$

Training. Five models are trained from the data. Each combines the visual data with a different language domain out of text, transcribed speech, text + transcribed speech, and speech audio. Vision data are matched with language data by their instance names and approximately 80% is reserved for training, 10% for validation and 10% for testing for a total of 13,040 text and speech training examples.

The models are trained with the ADAM optimizer on a 20GB Quadro RTX 6000 GPU. Each model is trained for a different number of epochs to balance for the variation in size of the training sets. Text, transcribed speech and speech are trained for 300, and the combined text and transcribed speech model is trained for 150 epochs. Each model outputs into a 1024-dimensional embedded space.

5 Evaluation

A held out testing set containing at least one of each object class is used for evaluation. A given image can only appear in one of the training and testing sets. We have found that the manifold alignment approach does not perform well on unseen object classes. We evaluate the models in two ways. We calculate the precision, recall, and F1 metrics by classifying based on the proximity to a target embedded datum. This method is further explained in Section 5.1. Finally, we calculate the Mean Reciprocal Rank (MRR) of two mock object retrieval tasks.

5.1 Grounded Language as Classification

The manifold alignment models we employ from [49] do not output a binary yes/no classification. Instead, classification is based on the proximity within the embedded space. This raises the question of how to define when two embedded vectors are "close" enough to be classified as related. To test this, we normalize the distances within our validation set to be between 0 and 1 by dividing the cosine distance by 2. Given a distance threshold between 0 and 1, we then classify positive instances as being within the threshold distance and negative instances as being outside the threshold. We then calculate the precision, recall, and F1 measure on our testing data as a function of the threshold. The F1 score at different thresholds for the combined text and speech transcription model can be seen in section 5.1.

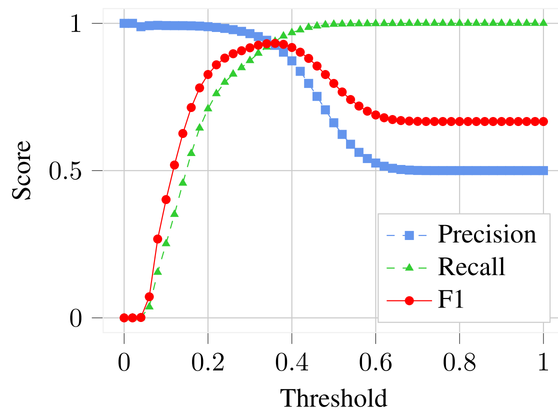


Figure 5: Precision, recall, and F1 on the validation set as a function of the threshold for classification for a combination of text and transcribed speech (peak F1=0.93). Graphs for pure text and pure speech show a very similar shape, reaching peak F1 of 0.91 and 0.94 respectively.

We see the best F1 results on the validation set with thresholds in the range $[0.35, 0.45]$. When those thresholds are applied to the testing set, the F1 for the text, transcribed speech, and combined models are .84, .94, and .92, respectively as shown in Table 6.

5.2 Grounded Language as Retrieval

Table 6: Mean Reciprocal Rank and F1 score on the testing set for models trained on Text and Speech descriptions over 5 runs. Triplet MRR is calculated from a query of the target and a positively and negatively associated test data point. Subset MRR is calculated from the target and a subset of four random test data points. The F1 score is calculated using the optimal threshold for each model.

Model	F1 score	Triplet MRR	Subset MRR
Text 300 epochs	0.84	0.85	0.89
Transcribed Speech 300 epochs	0.94	0.87	0.96
T + TS 150 epochs	0.92	0.87	0.94
(Test on T)	-	0.87	0.96
(Test on TS)	-	0.87	0.94
wav2vec 2.0 300 epochs	0.83	0.85	0.86
MFCC + LSTM 300 epochs	0.67	0.69	0.49
Random Baseline	-	0.61	0.46

The Mean Reciprocal Rank (MRR) is calculated by finding the distance of an embedded query vector to a list of possible embedded query response vectors, ordering them by cosine distance, and finding the rank of the target instance in the ordered list. The reciprocals of these ranks are summed over the testing set and then averaged by the number of testing examples. When the number of testing examples is very high, the MRR can quickly approach zero even when the rank of the instance near the top of query responses, rendering the metric difficult to interpret. To counteract this and to evaluate our model on a scenario that is more realistic to what it might be used for, such as object retrieval, instead of ranking the entire testing set we rank a select few instances. Our Triplet MRR metric is calculated from a triplet of the target, positive, and negative instances and the Subset MRR is calculated from a subset of the target and four other randomly selected instances.

The combined "T + TS" model is evaluated three separate times. First, it is tested individually on held-out sets where L is drawn first from text, then from speech. It is then evaluated on the combination of the two held-out sets. From our F1 evaluation, the transcribed speech model performs better than the other models, including the text model. These results seem to indicate that, despite potential errors in the transcription process, spoken input may be more linguistically meaningful than typed input. In all testing scenarios, there is little difference between the transcribed speech model and the combined text and transcription model.

All of our models perform better on the Subset MRR task than the Triplet MRR. This is likely due to the fact that the Subset MRR task does not intentionally contain a distracting positive instance. In a realistic environment, a robot could be faced with cluttered scenes with many distracting instances, both positive and negative, that it would need to distinguish between.

We train two models for grounding speech to images using manifold alignment. The first one uses the wav2vec 2.0 [5] representations as speech features and the second one uses MFCCs. We train both models for 300 epochs. The wav2vec 2.0 model achieves comparable performance to the model trained on transcribed speech on the Triplet MRR, showcasing that the speech data in our dataset is suitable for direct grounding of speech. However, the Subset MRR results show that there is a gap in performance between the two modalities.

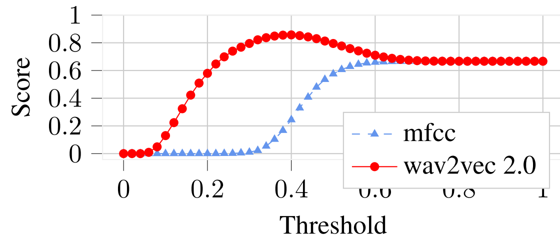


Figure 6: Threshold classification results for the speech models on the validation set. The wav2vec 2.0 model achieves a peak of 0.85 while the mfcc model stagnates at 0.66.

The MFCC model did not learn much. Figure 6 shows that the model achieves peak performance when the threshold is 1, classifying every pair as positive. The MRR results for the MFCC model in table 6 tell the same story with the model performing similarly to the random baseline. These results indicate that leveraging the semantic information learned by highly pretrained models such as wav2vec 2.0 significantly improves the quality of our grounding.

6 Conclusion

We introduced a new dataset that has four modalities of input (text, speech, RGB and depth) and allows us to tackle new challenges in grounded language learning such as learning directly from speech audio. Our investigation of the dataset establishes the quality of the data. Specifically we showed the feasibility of learning from typed text, transcriptions and raw speech. We also showed that the difference between the results of learning from typed or spoken descriptions is marginal. Our introduced baseline results show utility of the modalities and room for future methods to address issues not handled by current tools.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant Nos. 1813223, 1940931 and 1637937. Some experiments were conducted on the UMBC High-performance computing facility, funded by the National Science Foundation under Grant Nos. 1940931 and 2024878.

This material is also based on research that is in part supported by the Air Force Research Laboratory (AFRL), DARPA, for the KAIROS program under agreement number FA8750-19-2-1003. The U.S.Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of the Air Force Research Laboratory (AFRL), DARPA, or the U.S. Government.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-Embedding for Attribute-Based Classification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 819–826. IEEE, jun 2013.

[2] Muhannad Alomari, Paul Duckworth, David C Hogg, and Anthony G Cohn. Natural language acquisition and grounding for embodied robotic systems. In Thirty-Second AAAI Conference on Artificial Intelligence, 2017.

[3] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In International Conference on Machine Learning. PMLR, 2013.

[4] Takaya Araki, Tomoaki Nakamura, Takayuki Nagai, Kotaro Funakoshi, Mikio Nakano, and Naoto Iwahashi. Online object categorization using multimodal information autonomously acquired by a mobile robot. Advanced Robotics, 26(17):1995–2020, 2012.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Neural Information Processing Systems (NeurIPS), 2020.

[6] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In British Machine Vision Conference (BMVC), 2016.

[7] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In Advances in Neural Information Processing Systems, pages 9453–9463, 2019.

[8] Philipp Beckerle, Gionata Salvietti, Ramazan Unal, Domenico Prattichizzo, Simone Rossi, Claudio Castellini, Sandra Hirche, Satoshi Endo, Heni Ben Amor, Matei Ciocarlie, et al. A human–robot interaction perspective on assistive and rehabilitation robotics. Frontiers in Neurorobotics, 11(24):1, 2017.

[9] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, page 35–44, New York, NY, USA, 2018. Association for Computing Machinery.

[10] Joyce Y Chai, Maya Cakmak, and Candace Sidner. Teaching robots new tasks through natural interaction. Interactive Task Learning: Agents, Robots, and Humans Acquiring New Tasks through Natural Interactions, 2017.

[11] Joyce Y Chai, Rui Fang, Changsong Liu, and Lanbo She. Collaborative language grounding toward situated human-robot dialogue. AI Magazine, 37(4):32–45, 2016.

[12] David L Chen and Raymond J Mooney. Learning to sportscast: a test of grounded language acquisition. In International Conference on Machine Learning, pages 128–135. ACM, 2008.

[13] Guoguo Chen, Carolina Parada, and Tara N Sainath. Query-by-example keyword spotting using long short-term memory networks. In The International Conference on Acoustics, Speech, & Signal Processing (ICASSP), pages 5236–5240. IEEE, 2015.

[14] Alexandra Chouldechova and Max G'Sell. Fairer and more accurate, but for whom? In FAT ML Workshop, 2017.

[15] Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 613–622, 2017.

[16] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In Computer Vision and Pattern Recognition (CVPR) Workshops, pages 2054–2063, 2018.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In The Conference on Computer Vision and Pattern Recognition (CVPR), pages 248–255. Ieee, 2009.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT), 2019.

[19] William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Technical report, Northpointe, 2016.

[20] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin A. Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015.

[21] Jeanine C Evers. From the past into the future. how technological developments change our ways of data collection, transcription and analysis. In Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 2011.

[22] Bryce Goodman and Seth Flaxman. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". AI Magazine, 38(3):50, oct 2017.

[23] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. In Advances in Neural Information Processing Systems 29 (NIPS 2016), 2016.

[24] David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. In 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pages 237–244, 2015.

[25] David Harwath, Antonio Torralba, and James R. Glass. Unsupervised learning of spoken language with visual context. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, page 1866–1874, Red Hook, NY, USA, 2016. Curran Associates Inc.

[26] William Havard, Laurent Besacier, and Olivier Rosec. Speech-coco: 600k visually grounded spoken captions aligned to mscoco data set. GLU 2017 International Workshop on Grounding Language Understanding, Aug 2017.

[27] William N. Havard, Jean-Pierre Chevrot, and Laurent Besacier. Models of visually grounded speech signal pay attention to nouns: A bilingual experiment on english and japanese. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8618–8622, 2019.

[28] Julian Hough and David Schlangen. It's not what you do, it's how you do it: Grounding uncertainty for a simple robot. In 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages 274–282. IEEE, 2017.

[29] Wei-Ning Hsu, David Harwath, Tyler Miller, Christopher Song, and James Glass. Text-free image-to-speech synthesis using learned segmental units. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5284–5300, Online, August 2021. Association for Computational Linguistics.

[30] Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. Are you looking? grounding to multiple modalities in vision-and-language navigation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pages 6551–6557, 2019.

[31] Johanna E. Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1988–1997, 2016.

[32] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. Examples are not Enough, Learn to Criticize! Criticism for Interpretability. In D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett, editors, Advances in Neural Information Processing Systems 29. Curran Associates, Inc., 2016.

[33] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 123(1):32–73, 2017.

[34] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based explanations as feature importance measures. In International Conference on Machine Learning, 2020.

[35] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In 2011 IEEE international conference on robotics and automation, pages 1817–1824, 05 2011.

[36] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. 6th International Conference on Learning Representations, ICLR 2018, pages 1–14, 2018.

[37] Bo Li, Yu Tsao, and Khe Chai Sim. An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition. In Annual Conference of the International Speech Communication Association, 2013.

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European Conference on Computer Vision, pages 740–755. Springer, 2014.

[39] Zachary C Lipton. The Mythos of Model Interpretability. In ICML Workshop on Human Interpretability in Machine Learning, 2016.

[40] Zachary C. Lipton. The Doctor Just Won't Accept That! In Interpretable ML Symposium at NIPS, 2017.

[41] Nicholas Lourie, Ronan Le Bras, and Yejin Choi. Scruples: a corpus of community ethical judgments on 32,000 real-life anecdotes, 2021.

[42] D.G. Lowe. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, pages 1150–1157 vol.2. IEEE, 1999.

[43] Cynthia Matuszek. Grounded language learning: Where robotics and nlp meet. In Proceedings of the International Joint Conference on Artificial Intelligence, pages 5687–5691, 07 2018.

[44] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[45] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In European Conference on Computer Vision, pages 405–421, Cham, 2020. Springer International Publishing.

[46] Raymond J Mooney. Learning to connect language and perception. In Proceedings of the 23rd national conference on Artificial intelligence, 2008.

[47] Lindasalwa Muda, Begam KM, and I Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. Journal of Computing, 2(3):138–143, 2010.

[48] Arun Mukundan, Giorgos Tolias, Andrei Bursuc, Hervé Jégou, and Ondřej Chum. Understanding and Improving Kernel Local Descriptors. International Journal of Computer Vision, 127(11):1723–1737, 2019.

[49] Andre T. Nguyen, Luke E. Richards, Gaoussou Youssouf Kebe, Edward Raff, Kasra Darvish, Frank Ferraro, and Cynthia Matuszek. Practical cross-modal manifold alignment for robotic grounded language learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 1613–1622, June 2021.

[50] Anh Nguyen and Bac Le. 3d point cloud segmentation: A survey. In 2013 6th IEEE Conference on Robotics, Automation and Mechatronics (RAM), pages 225–230, 2013.

[51] Khanh Nguyen and Hal Daumé III. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 684–695, 2019.

[52] Thao Nguyen, Nakul Gopalan, Roma Patel, Matthew Corsaro, Ellie Pavlick, and Stefanie Tellex. Robot Object Retrieval with Contextual Natural Language Queries. In Proceedings of Robotics: Science and Systems, Corvalis, Oregon, USA, July 2020.

[53] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. Journal of Machine Learning Research, 22:1–64, 2021.

[54] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318, 2002.

[55] Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C Gates. An empirical analysis of word error rate and keyword error rate. In Ninth Annual Conference of the International Speech Communication Association, 2008.

[56] Gisela Redeker. On differences between spoken and written language. Discourse processes, 7(1):43–55, 1984.

[57] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.

[58] Luke E Richards, Kasra Darvish, and Cynthia Matuszek. Learning object attributes with category-free grounded language from deep featurization. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020.

[59] Deb K. Roy. Learning visually grounded words and syntax for a scene description task. Computer Speech & Language, 16:353–385, 2002.

[60] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, May 9-13 2011.

[61] Mert Bulent Sariyildiz and Ramazan Gokberk Cinbis. Gradient Matching Generative Networks for Zero-Shot Learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2168–2178, 2019.

[62] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 815–823, 06 2015.

[63] Pushkar Shukla, Carlos Elmadjian, Richika Sharan, Vivek Kulkarni, Matthew Turk, and William Yang Wang. What should i ask? using conversationally informative rewards for goal-oriented visual dialog. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6442–6451, 2019.

[64] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. Advances in Neural Information Processing Systems, 33, 2020.

[65] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical Networks for Few-shot Learning. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, Advances in Neural Information Processing Systems 30, pages 4080–4090. Curran Associates, Inc., 2017.

[66] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Y Ng. Zero-shot Learning Through Cross-modal Transfer. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1, NIPS'13, pages 935–943, USA, 2013. Curran Associates Inc.

[67] Megha Srivastava, Hoda Heidari, and Andreas Krause. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, pages 2459–2468, New York, NY, USA, 2019. ACM.

[68] Andreas Stolcke and Jasha Droppo. Comparing human and machine errors in conversational speech transcription. Annual Conference of the International Speech Communication Association, 2017.

[69] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3645–3650, 2019.

[70] Jared Sylvester and Edward Raff. Trimming the Thorns of AI Fairness Research. Data Engineering, 43(4):74–84, 2020.

[71] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. Annual Review of Control, Robotics, and Autonomous Systems, 3:25–55, 2020.

[72] Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. Asking for help using inverse semantics. In Proceedings of Robotics: Science and Systems. Robotics: Science and Systems Foundation, 2014.

[73] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In Proceedings of the AAAI Conference on Artificial Intelligence, 2011.

[74] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In Conference on Robot Learning, pages 394–406. PMLR, 2020.

[75] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond Mooney. Jointly improving parsing and perception for natural language commands through human-robot dialog. Journal of Artificial Intelligence Research, 67:327–374, 2020.

[76] Jesse Thomason, Jivko Sinapov, Raymond J Mooney, and Peter Stone. Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[77] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. Learning multi-modal grounded linguistic semantics by playing" i spy". In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pages 3477–3483, 2016.

[78] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 173–180. ACL, 2003.

[79] Emre Ünal, Ozan Arkan Can, and Yücel Yemez. Visually grounded language learning for robot navigation. In 1st International Workshop on Multimodal Understanding and Learning for Embodied Applications, pages 27–32, 2019.

[80] Andrea Vanzo, Danilo Croce, Emanuele Bastianelli, Roberto Basili, and Daniele Nardi. Grounded language interpretation of robotic commands through structured learning. Artificial Intelligence, 278:103181, 2020.

[81] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. Advances in Neural Information Processing Systems, pages 3637–3645, 2016.

[82] Chang Wang and Sridhar Mahadevan. Manifold alignment using procrustes analysis. In Proceedings of the 25th international conference on Machine learning, pages 1120–1127, 01 2008.

[83] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, IJCAI'11, page 1541–1546, 2011.

[84] Deng-Bao Wang, Li Li, and Min-Ling Zhang. Adaptive Graph Guided Disambiguation for Partial Label Learning. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, pages 83–91, New York, NY, USA, 2019. ACM.

[85] Ayelet Weizs. How long it really takes to transcribe (accurate) audio, Jul 2019.

[86] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 417–421, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[87] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1, NIPS'14, page 487–495, Cambridge, MA, USA, 2014. MIT Press.

Appendix

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] The limitations are discussed in Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] The potential negative impacts are discussed in section 7
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] There is code, sample data, and instructions are included in the supplemental materials.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The training details can be found in the Section 4
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] The computing resources that were used are detailed in Section 4
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] The code used for manifold alignment was an existing assets, pretrained BERT and ImageNet models were also and Google's speech to text API was used and all were cited.
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] How consent was obtained is discussed in Section 7
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] The data in this work does potentially contains personally identifiable information and this is discussed in Section 7
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] The screenshots were added to the supplementary materials.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] Potential risk were discussed in Section 7
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] The intent is to pay 15 USD per hour, the per hit wage was set after a pilot study determined that the rate was sufficient to reach a 15 USD/hour wage, see Section 9 for the exact payment amounts.

7 Ethics Statement

Data deletion. It is critical that workers understand the possible uses of their data and consent to all such, and we are committed to ensuring that understanding. Due to an error in the URL pointing to the consent form during initial data collection, we were unable to be sure that the text of that consent was available to every worker in the course of completing certain tasks. Accordingly, we discarded descriptions obtained under those circumstances and re-collected approximately 12,000 spoken and 8000 textual descriptions.

Possible societal harm. We hope this dataset will ultimately support a wide range of machine learning thrusts (see section 8). While there is never any guarantee that research will do no harm, the set of possible approaches we envisage is primarily benign, used to support basic research rather than specific domain targets. As an example, the goals of our work are focused on providing tools to support understanding physically situated human language, rather than on any specific application. Given the benign nature of the topics in the dataset itself, there is no more specific risk of the technology being developed than is typical for research into natural language or language grounding.

Possible worker harm. The primary ethical concerns that may be raised by this work relate to the privacy of the individuals who provide descriptions of objects. Mechanical Turk does not provide personally identifiable information (P.I.I.) about workers, and the data we have access to—such as the Amazon Worker ID—is not contained in the dataset. Because there are possible ways of re-determining a contributor's identity, including standard de-anonymization techniques, all collected language is limited to factual descriptions of simple household objects, and no data is collected that might obviously harm a participant if revealed. No value judgments, opinions, emotional topics, or discussions of personal situation or standing are included. This work was judged by our institution's IRB to be no more than minimal risk.

Worker confidentiality. Beyond that, we identify two primary ways in which worker confidentiality might be breached. First, background noise during may be audible in voice recordings, and may potentially leak information the worker would prefer to keep private. We attempt to mitigate this risk by requiring workers to replay the recording themselves before submitting, and by keeping individual recordings very short. Second, workers may be identified as having participated by their voice, or potentially by something non-obvious in their descriptions, such as an unusual turn of phrase; we warn workers of this possibility, as well as describing the intended use of the data. Worker participation is always voluntary.

8 Other Machine Learning Research GOLD can Support

While GOLD was designed for studying issues in grounded language learning that are not easily done with prior datasets, we note that the large number of modalities provided allows studying many different AI/ML tasks using GOLD. This includes more real-life data for representing point clouds as regressive geometries [64], and related active areas like NeRF for reconstructing novel views from the point cloud data [45]. Recognizing the same object from novel views using image (or 3D) descriptors [42, 48] is also possible due to GOLD's multiple views and relates to enabling robots to understand object permanence.

Many current active research directions can be expanded in new directions using GOLD. For example, a user may want their robot to explain its action when teaching it or in frustration after an errant behavior. But despite rich and growing literature on the topic of explainable AI we are not aware of any methodologies for explanations when multiple modalities are apart of the decision process [22, 32, 34, 39, 40, 57]. There are also unique perspectives around fairness in object recognition when we consider assistive robotics, where it may be highly desirable to alter the system due to an individuals unique capabilities. We are not aware of any work exploring these kinds of fairness concerns that address different persons' abilities to use a system (e.g., stutter) and specific needs (e.g., fall risk) that would make a single system well-intentioned but sub-optimal, and that many different customizable biases are preferred [14, 19, 23, 41, 67, 70]. Normalizing flows [53] between manifolds defined by different modalities due to changes in the contraction of spaces between domains (e.g., the tokens "orange" and "apple" are easy to separate linguistically, but harder to separate visually). Zero-shot learning in particular is predicated on having some form of side information that infers or describes the new class [1, 61, 65, 66, 81], and our multiple modalities provides another avenue for exploring this in a domain that requires few-shot learning for practical use.

Beyond these possibilities from the raw data we provide, more options are also available given augmentation. Point cloud segmentation [50] allows extracting the individual objects and imposing them in new scenes with other geometries, so one can generate more complex training and evaluation scenes or mix our data with other datasets. Partial label learning [84] by inserting label noise based on visual or linguistic similarities to study the difficulty of determining the correct label when users erroneously over-specify an object. Enabling broader use of robotic technology with Machine translation using side information of the visual modalities is also possible. Machine translation has been done without parallel corpora by exploiting the similarity in manifolds produced by (sufficiently linguistically similar) word embeddings for different languages [36]. Given professionally translated transcriptions, or collecting additional descriptions of the objects, one could use our data to study augmented translation given a forcefully shared manifold of the described objects visual properties (RGB+depth).

9 Datasheets for Datasets

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
This dataset was created to aid in the development of grounded language acquisition models. Existing datasets for this purpose focus on text descriptions either written or transcribed, while GoLDcontains both text, spoken speech, and speech transcriptions, allowing for grounded language learning to be performed directly on spoken language.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
This dataset was created by the Interactive Robotics and Language lab (IRAL) at the University of Maryland, Baltimore County

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
This dataset is based in part upon work supported by the National Science Foundation under Grant Nos. 1940931 and 1637937 and is also based on research that is in part supported by the Air Force Research Laboratory (AFRL), DARPA, for the KAIROS program under agreement number FA8750-19-2-1003.

Any other comments?
NA

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
GoLD consists of color images, depth images and pointclouds of 207 objects, and written and spoken descriptions of them.

How many instances are there in total (of each type, if appropriate)?
There were 207 object instances in 47 classes with 1 to 5 instances per class, broadly fitting into 5 high level groups as seen in Table 1.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
The dataset is a nonrepresentative sample of objects that may be found in household environments. The greater goal that this dataset is intended to support is the development of language grounding models for use with assistive robots. To that end the objects chosen to be in this dataset were common household objects, office supplies, hand tools, food, and medical supplies.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each instance consists of color images, depth images and pointclouds of each object instance from 4 different views as seen in Fig. 1 for each of the 207 instances. There are also 16500 spoken descriptions, and 16500 text descriptions.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance has a set of corresponding spoken, transcribed, and written descriptions.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
There is not any information missing from any of the instances.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
Yes, all instances that are related are explicitly named, e.g. apple_1, apple_2, apple_3.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
There are no training, development, validation, or testing splits.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
For each object there are multiple descriptions both written and verbal, and there is the expected noise in the depth images, and audio and transcribed descriptions.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
The dataset does not contain any confidential data.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
The dataset does not contain any data that may be offensive, insulting, threatening, or might otherwise cause anxiety.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
The dataset does relate to people, as it contains peoples written and spoken descriptions of objects.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
This dataset does not explicitly identify subpopulations, but as it does contain voice recordings it is possible to manually identify subpopulations through characteristics of their voice.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
It may be possible to identify individual subjects through their voice.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
The dataset does not contain any sensitive data.

Any other comments?
NA

COLLECTION

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
For each instance the images and pointclouds of the objects were directly observed, and the descriptions were all reported by the subjects. The spoken and written descriptions were obtained directly from the subjects while the transcribed spoken descriptions were obtained from Google's speech to text API. The data collected from Amazon Mechanical Turk was manually evaluated to detect any bad actors, whose responses were removed from the dataset.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.
The dataset was collected between November of 2019.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
All the image data was collected using a Microsoft Azure Kinect and captured using the Robot Operating System (ROS). The descriptions were collected using Amazon Mechanical Turk, and Google's speech to text API was used to obtain transcriptions of the spoken descriptions. As mentioned previously the descriptions obtained from Amazon Mechanical Turk were manually curated to remove responses from bad actors.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell *et al.*[69] for approaches in this area.)
Unknown

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
NA

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
The images of the objects were collected by the authors, and the descriptions were collected from Amazon Mechanical Turk crowdworkers and were compensated \$0.13 per hit for text descriptions(five object descriptions), and \$0.08 for spoken descriptions (one object description).

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or

other access point to any supporting documentation.
There was an ethical review process conducted, through the UMBC's Institutional Review Board and was approved. The consent form for gathering the speech description can be seen at `http://tiny.cc/spoken-hit-consent` and for the text descriptions at `http://tiny.cc/text-hit-consent`.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
The descriptions of the objects were collected using Amazon Mechanical Turk.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
The data collection was conducted on Amazon Mechanical Turk, using the title "Give a short description of everyday objects", and description "Give a short description of everyday objects."

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
Yes, individuals were provided a copy of a consent form, and were given the option to return the hit if they did not consent. The consent form for gathering the speech description can be seen at `http://tiny.cc/spoken-hit-consent` and for the text descriptions at `http://tiny.cc/text-hit-consent`

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)
No.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
No.

Any other comments?
NA

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
The raw data that was collected were videos of each object instance on a turntable going through a full rotation captured at 5 frames per second, from this four representative keyframes were manually selected to capture diverse view angles of the object.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw"

data.
No.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
No.

Any other comments?
NA

USES

Has the dataset been used for any tasks already? If so, please provide a description.
The GOLDdataset was evaluated by performing classification, retrieval, and speech recognition tasks. Each experiment will combine the RGB+depth images with one of the three language domains: text, transcribed speech, speech audio and a combination of text and transcribed speech.

Visual features were extracted by passing the color and colorized depth images through CNNs that have been pretrained on ImageNet [17], with the last prediction layer removed, leaving the final layer as a learned feature vector [20, 58]. The language features of text and transcribed speech were extracted using a pretrained BERT [18] model with the last four hidden layers concatenated into a 3072-dimensional feature vector. wav2vec 2.0 [5], a self-supervised speech model that learns over continuous representations of raw speech through a BERT [18] inspired masked language modeling task. Similarly to the text featurization, features are then learned by performing average-pooling over the concatenation of the last four layers of the transformer. To evaluate the benefit of using a pre-trained model, we also consider 40 dimensional Mel-frequency cepstral coefficient (MFCC) features [47] that are extracted from the raw audio with a 10 ms frame shift. Due to the lower-dimensional nature of MFCCs, the language network is modified to include a Long Short-Term Memory (LSTM) network. 64-dimensional outputs from the final 32 hidden states [13] are concatenated together to form a fixed length 2048-dimensional speech vector which are passed to a fully connected layer and output into the same embedded dimension as the visual network. Manifold alignment [3, 82, 83] with triplet loss [6, 49] is used to embed the visual percepts and language data from GOLDinto a shared lower space.

Four models were trained, each combining the visual data with a different language domain from text, transcribed speech, text + transcribed speech, and speech audio. Vision data are matched with language data by their instance names and approximately 80% is reserved for training, 10% for validation and 10% for testing.

The manifold alignment models employed from [49] do not output a binary yes/no classification, the classification is instead based on the proximity in the embedded space. The optimal threshold was found to be in the range of threshold in the range $[0.35, 0.45]$. When these thresholds are applied to the test set, the F1 for the text, transcribed speech, and combined models was found to be .84, .94, and .92, respectively.

In the retrieval task was evaluated using Triplet and Subset Mean Reciprocal Rank (MRR). As when the number of testing examples is high MMR can rapidly approach zero we rank a select few instances. The Triplet MRR metric was calculated from a triplet of the target, positive, and negative instances and the Subset MRR was calculated from a subset of the target and four other randomly selected instances.

The combined text and transcribed speech model were evaluated three times. First, it is tested individually on held out sets where L is drawn first from text, then from speech. It is then evaluated on the combination of the two held-out sets.

The results can be seen in 6. The combined "T + TS" model is evaluated three separate times. First, it is tested individually on held-out sets where L is drawn first from text, then from speech. It is then evaluated on the combination of the two held-out sets. From our F1 evaluation, the transcribed speech model performs better than the other models, including the text model. These results seem to indicate that, despite the potential errors in the transcription process, spoken input might lead to more meaningful language utterances than typed input. In all testing scenarios, there is little difference between the transcribed speech model and the combined text and transcription model.

The speech model achieves comparable performance to the model trained on transcribed speech on the Triplet MRR, showcasing that the speech data in our dataset is suitable for direct grounding of speech. However, the Subset MRR results show that there is a gap in performance between the two modalities.

The MFCC model did not learn much. Figure 6 shows that the model achieves peak performance when the threshold is 1, classifying every pair as positive. The MRR results for the MFCC model in table 6 tell the same story with the model performing similarly to the random baseline. These results prove that leveraging the semantic information learned by highly pretrained models such as wav2vec 2.0 significantly improves the quality of our grounding.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
No.

What (other) tasks could the dataset be used for?
Since this dataset contains 3D information it is possible to perform data augmentation, building more complex scenes using the pointclouds. Since this dataset includes speech and perceived characteristic of the speaker can be annotated, it is possible investigate methods that can avoid bias in learned language models.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
The transcribed speech was directly taken from Google's speech to text API and would need to be fully evaluated for accuracy before being used as a source of ground truth.

Are there tasks for which the dataset should not be used? If so, please provide a description.
Unknown

Any other comments?
NA

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
Yes, the dataset will be publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset is available at: `https://github.com/iral-lab/gold`

When will the dataset be distributed?
Unknown

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

No.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No.

Any other comments?
NA

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
The dataset will be hosted on GitHub and will be maintained by Gaoussou Youssouf Kebe and Cynthia Matuszek.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The maintainers can be contacted at gaoussou1@umbc.edu or cmat@umbc.edu.

Is there an erratum? If so, please provide a link or other access point.
No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
The dataset will be updated with new object descriptions, and they will be communicated via GitHub.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
Yes, they will be available through GitHub.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
No.

Any other comments?
NA