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