GRAFFORD: A Benchmark Dataset for Testing the Knowledge of Object Affordances of Language and Vision Models

Anonymous ACL submission

Abstract

We investigate the knowledge of object affordances in pre-trained language models (LMs) and pre-trained Vision-Language models (VLMs). Transformers-based large pre-trained language models (PTLM) learn contextual representation from massive amounts of unlabeled text and are shown to perform impressively in downstream NLU tasks. In parallel, a growing body of literature shows that PTLMs fail inconsistently and non-intuitively, showing a lack of reasoning and *grounding*. To take a first step toward quantifying the effect of grounding (or lack thereof), we curate a novel and comprehensive dataset of object affordances – GRAFFORD, characterized by 15 affordance classes. Unlike affordance datasets collected in vision and language domains, we annotate *in-the-wild* sentences with objects and affordances. Experimental results reveal that PTLMs exhibit limited reasoning abilities when it comes to uncommon object affordances. We also observe that pre-trained VLMs do not necessarily capture object affordances effectively. Through few-shot fine-tuning, we demonstrate improvement in affordance knowledge in PTLMs and VLMs. Our research contributes a novel dataset for language grounding tasks, and presents insights into LM capabilities, advancing the understanding of object affordances.

1 Introduction

The task of object affordances has been well studied in Computer Vision and Robotics (Zhu et al., 2014; Kjellström et al., 2011; Gupta et al., 2009). However, infusing such knowledge while learning textual representation is hard; as in NLP, we lack corresponding images (or videos) which may provide necessary shape, property information to predict affordances. The state-of-the-art transformers-based large pre-trained language models (PTLMs) learn textual representation from massive amounts of unlabeled text and are shown to perform impressively in downstream NLU tasks. In parallel, a growing body of literature shows that PTLMs fail inconsistently and non-intuitively, showing a lack of reasoning and *grounding*. In NLP, these results prompted authors in (Bender and Koller, 2020) to reiterate the gap between *form* and *meaning*.

The authors argue that language models which are exposed to only text (surface form) may never truly understand *meaning*, as PTLMs are unaware of true *groundings* of the surface text. Our goal is to first precisely *quantify* the gap, capturing one aspect of grounding, aka *affordability*. To this end, we take inspiration from vision and robotics research and explore affordance properties of entities or objects in text, which a model with proper grounding should be able to estimate. We propose a novel text-to-affordance dataset, and explore to what extent textual representations learnt by PTLMs and VLMs can enable reasoning with *affordance* of entities mentioned in text.

As an example, for a sentence "an apple in the tree", we should infer that the "apple" can be thrown, and is rollable. However we cannot roll an "apple logo". In computer vision and robotics efforts, an accompanying image (or video) often provides necessary information about shape and physical properties of an entity, which can be used to predict affordances (Zhu et al., 2014). However such information is absent in NLP tasks. To capture this nuance, we annotate crowdsourced text intended for other tasks (such as NLI) with the objects and affordances. We use 15 affordance classes from Zhu et al. (2014). Through extensive pilot studies, we train a set of annotators using the `toloka.ai` platform. We choose 25 highly-skilled annotators who annotates a total of 2368 sentence-object pairs with 15 affordance classes, on a 0-3 Likert-like scale. We name this novel dataset GRAFFORD.We use the dataset for zero-shot evaluations of small LMs, two open-source LLMs and also some VLMs. We evaluate the effect

of few-shot fine-tuning on few PTLMs and VLM. Our contributions can be summarized as follows.

- We curate a novel large scale crowdsource based text to affordance dataset – GRAFFORD, consisting of 2368 sentence-object pairs and 15 affordance classes for each pair.
- We perform zero-shot evaluation of many state-of-the art PTLMs along with a few VLMs in different settings to identify the extent to which they gain the knowledge of affordance using our dataset. We further ensemble the VLM and the PTLM predictions to examine whether pre-training with images can enrich affordance prediction from text.
- We also fine-tune few PTLMs on a small subset of our data as well as on some commonsense reasoning tasks to understand how quickly the affordance knowledge get scaled up and how far the affordances are related to common sense knowledge. In addition, we examine the in-context learning (ICL) ability of few of the SOTA generative LLMs and VLMs in affordance prediction task.
- We provide with a comprehensive analysis of the model prediction and explore how the linguistic ambiguity of the objects can affect the models' ability of affordance prediction.

2 Related work

Natural language grounding. In vision and robotics, grounding has traditionally referred to as locating and identifying text expressions within images (or videos). Yu et al. (2015) introduce referring expression grounding which grounds referring expressions within given images via jointly learning the region visual feature and the semantics embedded in referring expressions. Chen et al. (2017) present phrase grounding which aims to locate referred targets by corresponding phrases in natural language queries. Shridhar and Hsu (2018) employ expressions generated by a captioning model (Johnson et al., 2016), gestures, and a dialog system to ground targets. Another line of work (Bastianelli et al., 2016; Alomari et al., 2017) explored non-dialog based methods to ground text queries.

Reasoning about object affordances. Sun et al. (2014) learnt object affordances through human demonstrations & Kim and Sukhatme (2014) deduced affordance through extracted geometric features from point cloud segments. Zhu et al. (2014) reasoned affordance through querying the visual attributes, physical attributes, and categorical characteristics of objects by employing a pre-built knowledge base. Myers et al. (2015) perceive affordance from local shape and geometry primitives of objects. Recent methods employed deep learning approaches to detect object affordance. Roy and Todorovic (2016) used a multi-scale CNN to extract mid-level visual features and combine them to segment affordances from RGB images. Sawatzky et al. (2017) regard affordance perception as semantic image segmentation and adopt a CNN based architecture to segment affordances from weakly labeled images. Nguyen et al. (2016); Mi et al. (2019) utilized features from a pre-trained CNN model to predict object affordances. Nguyen et al. (2017) applied an object detector, CNN and dense conditional random fields to detect object affordance from RGB images.

Probing methods. Talmor et al. (2020) utilized probing and employed Multi-choice MLM (Masked Language Modelling) and Multi-choice QA (Question Answering) setup to capture reasoning capabilities of pre-trained Language Models. Yang et al. (2022) examine zero-shot prediction performances on different tasks by LLM through novel visual imagination. Aroca-Ouellette et al. (2021) highlighted the shortcomings of state-of-the-art pre-trained models in physical reasoning, with a further performance decline observed when incorporating option shuffling and superlatives in reasoning questions. Liu et al. (2022) proposed a novel spatial common sense probing framework to investigate object scales and positional relationship knowledge in text-based pre-trained models and models with visual signals.

3 GRAFFORD dataset construction

Preprocessing. We select 20000 sentences from a crowdsourced English dataset (XNLI English) (Conneau et al., 2018)[1] and extract the noun phrases using the Stanford CoreNLP tool. As we restrict to the affordances that humans can directly perform, we filter the phrases which do not represent a tangible object (using ConceptNet). We manually filter out objects that cannot be acted upon directly by humans (such as school, building). After this preprocessing, we obtain a set of sentence-object pairs ($\langle x_i, o_i \rangle$), where the sentence acts as the context for the corresponding object.

[1]We choose XNLI as a source to facilitate multilingual extensions of our dataset.

Each sentence on average has 2-3 such objects. We use the 15 predefined affordance classes from Zhu et al. (2014) to label each sentence-object pair for annotation.

We further expand our dataset with the labeled dataset provided by Zhu et al. (2014). Authors present 62 common objects and their corresponding 15 affordance labels. Given that our task is *context-based affordance* prediction, we require to have sentence-object pairs for labelling. To generate diverse context for this dataset, we utilize the ChatGPT UI[2][3] model to generate synthetic sentences for each of the objects, followed by careful manual correction.

Pilot studies & annotator training. We annotate the dataset using the Toloka platform[4]. We design an interface on this platform, which contained clear instructions and examples for annotating the data. We conduct two rounds of pilot studies to analyze the subjective understanding of the annotators and, thereby, filter out the high quality, serious annotators. For the first pilot study, we present the annotators with the smaller 62 sentence-object pairs and ask them to label the instance with each affordance class on a scale of 0 to 3, indicating whether or not the affordance can be performed on the object. Here, 0-1 indicates that the affordance cannot be performed (high-low) and 2-3 indicates that the affordance can be performed low-high). We will further use these 62 synthetic sentence-object pairs for few-shot training. For quality control, we select the top 90% of the available annotators in the platform, who are proficient in English, and use computers to complete the tasks[5]. A total of 15 annotators labelled the data, and all of them were incentivized uniformly. After the first pilot, we find that there is an extremely poor agreement among the annotators, and the overall precision is around 28%. Therefore, we moved on to a second pilot study. Here, we use all the 62 sentence-object pairs from the previous study, along with 32 randomly selected sentence-object pairs from the XNLI data. We use the top 30% of the annotators (based on the quality determined by the platform) available on the platform, while other criteria remained the same. We annotate 32 sentence-object pairs ourselves, and use all the labelled examples as *control* data points to guide the annotators while labelling. A total of 114 annotators participated in this version of the pilot study. We assign a specific skill to the annotators who attained more than 30% precision and 30% recall. In total, 48 annotators passed this criteria. Through initial pilot studies, we learnt that without grounded images, the task appears quite subjective to annotators. The main goal of the pilot studies have been to understand the annotators' quality, their comprehension of the task, and their preferences for incentives per task. We have also conducted two additional AMA (Ask Me Anything) sessions with interested annotators to further clarify the task.

# of sentence-object pairs annotated	2368
# of affordance class	15
# of instances annotated	106560 (2368 $\times$ 15 $\times$ 3)
Avg # of objects / class	333
Most prominent class	Lift (851 objects)
Least prominent class	WriteWith (3 objects)
Total skilled annotators used	25
Avg agreement (Kripendorff's α)	0.68

Table 1: GRAFFORD dataset statistics.

Final annotation. In the final phase, we conduct the annotation on a larger set of sentence-object pairs, carefully selecting a total of 2,368 pairs. To ensure diverse perspectives and minimize bias, we engage 25 skilled annotators in this phase. Three annotators independently annotated each of the sentence-object pairs. Each annotator meticulously evaluated the affordance classes for every pair, contributing to a comprehensive annotation of the dataset. We perform the annotations in phases and complete the full task over 10 phases. We measure class-wise agreement and average agreement across all classes after each annotation phase to ensure the quality of the annotations. The overall statistics for this *currently* constructed dataset – GRAFFORD is in Tab. 1. Throughout the data processing pipeline, we put meticulous attention to the quality control, including the use of pilot studies, iterative annotation refinement, and manual filtering. These measures ensure that the dataset is comprehensive, accurate, and aligned with the objectives of the study. Overall, our GRAFFORD dataset consists of 2368 sentence-object pairs having $\sim 100k$ annotations (2368 $\times$ 15 $\times$ 3).

GRAFFORD data exploration. We observe that classes such as 'Grasp', 'Lift', 'Throw', 'Push', and 'Watch' are the most common affordances

[2] https://chat.openai.com

[3] Prompt used: Can you make realistic sentences with the following objects? Followed by the list of object names.

[4] https://toloka.ai/

[5] We exclude mobile-users as we believe the instructions may not appear clearly on mobile devices.

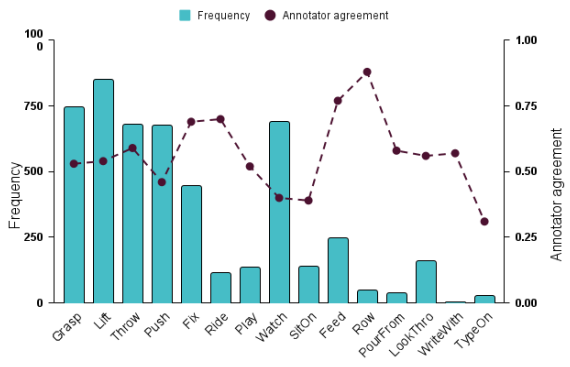


Figure 1: Classwise distribution of the number of objects and the annotator agreement.

for the objects present in the dataset (see Figure 1). Most frequent objects and their corresponding agreement scores are shown in Appendix A.7 Fig. 8. We observe, agreement scores are fairly uniform (0.5-0.6) for frequent objects, with high agreement for some frequent objects (0.8 for "the movie"). In Figure 9 (see Appendix A.6), we also see that 'Grasp', 'Lift', and 'Throw' are highly correlated classes. There is similar positive correlation between the class 'SitOn' and 'Ride', and some correlation between 'Watch' and 'LookThrough'. In Table 2, we list down the affordance classes based on the annotator agreement score, and divide it into *three* categories to understand which of the affordance classes pose the most and least difficulties for the human annotators. We observe that the classes - 'Watch', 'SitOn', and 'TypeOn' are the most difficult to disambiguate.

4 Experiments

We explore various state-of-the-art baselines using pre-trained language models (RoBERTa-large, BART-large), instruction-fine-tuned large language models (e.g., FLAN-T5, Falcon), pre-trained multimodal vision and language architectures (CLIP-ViT, ViLT, InstructBLIP, IDEFICS, LLaVA). We observe whether these models gain the knowledge of affordances through their pre-training, fine-tuning on commonsense tasks (NLI, PIQA), or few-shot fine-tuning scenarios.

4.1 Zero-shot affordance prediction

4.1.1 Pre-trained language models

Zero-shot prediction task is framed in different ways.

Masked language modelling (MLM) based approach. Here, we pose the zero-shot task as masked word prediction problem. Talmor et al. (2020) performed zero-shot prediction by employing Multi-Choice Masked Language Modelling approach, where the pre-trained model are required to predict the masked word from a set of given keywords. We choose BERT-large-uncased, RoBERTa-large (Zhuang et al., 2021), and BART-large (Lewis et al., 2020) models for the experiment. We pass the sentence and prompt separated by a `[SEP]` token as an input to the model. We use the prompt "`<Object> can be used for <MASK_TOKEN> by human`" and obtain the probability of each affordance label using the logit corresponding to the `<MASK_TOKEN>`.

Predictions from generative LLMs. Petroni et al. (2019); Schick and Schütze (2021) treats NLU tasks as a cloze test using prompts. We apply autoregressive language models such as FLAN-T5 (Chung et al., 2022) (large, xl, and xxl), Falcon (Almazrouei et al., 2023) (7b and 40b), ChatGPT to get the predictions. We provide with a 'YES/NO' question-answer based prompt to the LLMs to predict whether a particular affordance can be performed on the given object. Based on rigorous prompt engineering we choose specific prompts for the different models as shown in the Appendix Table 7. We map 'YES/NO' predictions to 1/0 labels respectively.

4.1.2 Common sense reasoning tasks

To understand whether the injection of the common sense knowledge in the pre-trained models can enhance the performance of the affordance prediction, we first fine-tune the pre-trained models on common sense reasoning dataset such as PiQA (Bisk et al., 2019). Then we run the fine-tuned models on our dataset using the MLM setup. We use BERT-base, BERT-large, RoBERTa-large, and BART-large in this setup.

Natural language inference (NLI) based approach. The NLI task considers a premise and a hypothesis as input pair $\langle p, h \rangle$, and the models are trained to predict the probability whether the hypothesis is entailed by, contradicts or neutral with respect to the premise. Here we use the entailment probability from the models: $p_{La}(h|p) = p(l = \text{“}ENTAILMENT\text{”}|(p, h))$. This approach requires language models to be fine-tuned on premise-hypothesis pairs with the corresponding labels. Here we use RoBERTa-large and BART-large fine-tuned on the Multi-genre NLI (MNLI) corpus (Williams et al., 2018) consisting of 433k sentence

Agreement category	Affordance classes	Objects	Object-affordance pair
High agreement (>0.6)	Row, Feed, Ride, Fix	the horse, striped white shirts, a brown paper sack, Chinese lanterns, Adrin's sword, The movie	breakfast-Feed, a horse-Watch, crops-Fix, sports-Grasp, sports-Lift, sports-Push, the phone-Feed, football-Ride
Medium agreement (>0.45 & <0.6)	Throw, PourFrom, WriteWith, LookThrough, Lift, Grasp, Play, Push	A red flag, An arrow, Art, Automatic weapons, Babies, Black-and-white TV	computers-WriteWith, cats-Feed, football-Play, book-WriteWith, the door-Push
Low agreement (<0.45)	Watch, SitOn, TypeOn	Brandy from Spain, stone circles, iron, batteries, his fist, historical artifact, gift, olive oil, outdoor tables, bumper sticker on a car	weapon-Push, The table-Lift, boat-Fix, paintings-LookThrough, cats-Throw

Table 2: Annotator agreement based on the difficulty in disambiguating different affordance classes and objects.

pairs. For each sentence-object pair in our dataset, we use the corresponding context sentence in the MNLI dataset as the premise. We use the hypothesis as "*<object> can be used for <affordance> by human*" for each object present in the sentence and 15 affordance classes. Using the NLI setting, we predict the entailment score for each affordance class for the given sentence-object pair. We use these scores for ranking the affordance classes and report mAP scores as well as accuracy.

4.1.3 Multi-modal models

We explore both unimodal and multi-modal task setup for pre-trained vision and language models.

Unimodal text-only MLM setup

VLMs are pre-trained on large datasets having both image and text. The main goal of their pre-training is to capture some visual knowledge corresponding to the text while pre-training on multi-modal dataset such as image-caption pairs. To examine this, we first use the vision-language model CLIP, by providing only text prompt as the input and predict the affordance in an MLM setup.

Multimodal task setup

Images contain necessary information about shape, texture, and size (visual attributes) of objects that can be utilized to effectively predict an object affordance (such as the handle of the bucket can be used to grasp and lift). Hence, we also convert the problem into a multi-modal task by retrieving (or generating) a corresponding image from the context sentence, and predict the affordance of an object (mentioned in the sentence) based on the input.

Retrieving images. In this setup, we use two different techniques to retrieve *semantically close* images to corresponding context sentences using 1) retrieval and 2) generation. We further use top five images for both, to get an accurate estimation.

Retrieval based: Yang et al. (2022) used Bing Image Search[6] to retrieve images based on object. However, this process is costly as it requires paid subscription. So, we employ Visualgenome (Krishna et al., 2017) dataset, consisting of 108,077 images and 3.8 million object instances as the image database. We first encode the images using multi-modal CLIP (Radford et al., 2021) based sentence-transformers architecture, and index those image embeddings using Approximate Nearest Neighbour search (ANN)[7], for making the search efficient. Now, for each sentence, we search for top five images from the database to be used further.

Generation based: Recently, the multi-modal generative models (Ramesh et al., 2022; Saharia et al., 2022) have shown incredibly good performance for text based image generation tasks. We adopt the recent *StableDiffusion* (Rombach et al., 2022) model to generate top five images based on the sentence as a text prompt.

We use the top five retrieved images by using retrieval and generation methods each. We use *CLIP* (Radford et al., 2021) and *ViLT* (Kim et al., 2021) as our vision-text models. CLIP is pre-trained on 400M image-caption pairs with the contrastive learning strategy. CLIP has a text encoder f_T and a visual encoder f_V, which can project text and image into the shared latent space. We aggregate the k (=5) corresponding images and use CLIP to compute the relevance score of (x, y): $Score_{VI}(x, y) = \frac{1}{k}\sum_{i=1}^{K} \cos\left(f_T(x), f_v(I_y^k)\right)$, where I_y^k is the k^{th} image for the input text y. ViLT uses patch projection (Dosovitskiy et al., 2021) to encode images, and uses image-text matching, MLM, and image-patch alignment tasks as objectives. For the zero-shot prediction, we provide the text prompt along with the representative images as input to the ViLT model to predict the

[6] https://learn.microsoft.com/en-us/azure/cognitive-services/bing-image-search/overview

[7] https://pypi.org/project/annoy/

masked token. We use the same prompt as the previous MLM task (i.e., "*<Object> can be used for <MASK_TOKEN> by human.*") and get the probability of each affordance class as the logit corresponding to the <MASK_TOKEN>.

Text generation based. Similar to section 4.1.1, we utilize state-of-the-art VLMs to make predictions regarding object affordances. We provide with a 'Yes/No' question answering based text prompt along with the aligned images as input to the VLMs, and the model should generate an answer whether a particular affordance can be performed on the given object. We use state-of-the-art VLMs such as IDEFICS (Laurençon et al., 2023), LLaVA (Liu et al., 2023), InstructBLIP (Dai et al., 2023) for this task. The text prompt used for the models can be found in Table 7.

4.2 Ensemble language and vision prediction

Following Yang et al. (2022), we use the weighted sum as the late fusion over the final output probabilities of each affordance class from the language and multi-modal models. Before late fusion, we normalize the output probability scores from different models. We calculate the score as: $P_{ens}(y|x) = (1-w)p_{L_a}(y|x) + wp_{V_I}(y|x)$ where w is the relative size of the vision-text model and the language model (following Yang et al. (2022)): $w = Sigmoid\left(\frac{\rho_{V_I}}{\rho_{L_a}}\right)$. Here ρ_{V_I} and ρ_{L_a} denote the number of parameters of the multi-modal and language models respectively.

4.3 Few-shot prediction

We conduct few-shot experiments by 1) fine-tuning the encoder based models, 2) randomly selecting 5 demonstration examples for the generative models to perform few-shot in-context learning (ICL). We consider the 62 annotated objects and corresponding 15 affordance classes by (Zhu et al., 2014) for the few-shot based experiments.

Training data To create few-shot training examples for fine-tuning encoder based PTLMs, we take all the 62 objects, and for each object we randomly select exactly 1 positive affordance class (i.e., the class label annotated as 1) and 1 negative affordance class (i.e., the class label annotated as 0) for generating the training prompt. As this dataset does not contain any context sentences for a corresponding object, we use ChatGPT UI to generate the sentences for the corresponding objects and manually verify the sentences, so that it does not contain any invalid information. Finally, we have 62 sentence-object pairs and 2 classes (one positive and one negative) per pair, which we use to generate training examples. Each training example consists of a prompt and a label. They constitute 124 training examples (62 sentence-object pairs and 2 selected classes for each) for the few-shot experiment.

Selecting examples for in-context learning: We randomly sample five sentence-object-affordance triples from the above training data as the in-context demonstration examples in such a way that there should be k positive affordance classes. We vary the number of positive affordance classes $k \in \{1, 2, 3\}$ and report the average accuracy.

Experimental setup. We fine-tune the encoder based language models using the training data, and for the generative LLMs and the VLMs, we utilize the training data to select in-context demonstration examples.

Fine-tuning PTLM: We fine-tune the PTLMs in two different setups - NLI based and prompt based. For the NLI based setup we have the context sentence as premise and use same prompt (i.e., "<object> can be used for <affordance> by human") which we use in the zero-shot settings as hypothesis. We use label as 1 for the positive affordance and label as 0 for the negative affordance. We use BERT-large-uncased, RoBERTa-large and BART-large for fine-tuning in this setup. We reuse these fine-tuned models for few-shot predictions in MLM setup. We use Adam optimizer with a learning rate of 2×10^{-5}. We fine-tune the model for 5 epochs for each case.

In-context learning for generative models: We employ the same generative LLMs as well as VLMs to perform affordance prediction using *five* demonstration examples from the training data. We use the same text prompt as zero-shot setting and concatenate the five demonstration examples along with corresponding label (i.e., 'Yes' for positive class, and 'No' for the negative class) to the prompt and ask the LLMs and VLMs to predict the affordance. In case of the VLMs, we do not provide any additional image example here.

5 Result

Evaluation metric. To assess the performance of the zero-shot affordance prediction, we calculate accuracy in the following way. Each affordance class is treated as a binary classification problem,

where a value of 1 represents a positive class indicating that the affordance can be performed on the object, and a value of 0 represents a negative class indicating that the affordance cannot be performed.

For each positive class $\in \{P_1, P_2, ..P_n\}$, we compare the predicted scores of that affordance class with the predicted scores of the negative classes $\in \{N_1, N_2, ..N_m\}$. If the predicted score of the positive class is higher than the predicted score of all the negative classes (i.e., $p(P_i) > p(N_j)_{\forall j}$), we increment the correct count by 1[8]. Conversely, if the predicted score of the negative class is higher, we increment the wrong count by 1. The final accuracy is calculated by dividing the total number of correct counts by the total number of the instances. To rank the affordance classes based on the predicted score, we also report the Mean Average Precision (mAP@K, where K is the number of affordance classes).

Encoder based						
NLI based (Normal)						
	Actual		LM + VI (CLIP)		LM + VI (ViLT)	
Model	Acc	mAP	Acc	mAP	Acc	mAP
RoBERTa-large-mnli	0.64	0.43	**0.79**	0.52	**0.79**	**0.54**
BART-large-mnli	**0.65**	0.38	0.62	0.4	0.64	0.43
MLM based						
Model	Acc	mAP	Acc	mAP	Acc	mAP
BERT-large-uncased	0.46	0.26	0.55	0.38	0.53	0.37
RoBERTa-large	**0.55**	0.36	0.61	0.41	0.62	0.43
BART-large	0.47	0.28	0.56	0.35	0.52	0.34
Multi-modal models						
Model	Acc	mAP	Acc	mAP	Acc	mAP
CLIP-VIT (text-only)	0.47	0.34	-	-	-	-
CLIP-VIT (retrieval)	0.56	0.35	-	-	-	-
CLIP-VIT (generation)	**0.61**	0.4	-	-	-	-
ViLT (retrieval)	0.41	0.31	-	-	-	-
ViLT (generation)	0.44	0.32	-	-	-	-

Table 3: Zero-shot performance for affordance prediction using encoder based models. Acc: Accuracy, LM: Language model, VI: Vision. Only LMs are ensembled with VI. The best results are marked in bold.

5.1 Zero-shot performance

Table 3 shows the results of the zero-shot affordance predictions from the mentioned models. The second column (i.e., Actual) indicates the values from the original LM and multi-modal models. The third and fourth columns (i.e., LM + VI) indicate the performances of ensembling language models with two of the multi-modal models we used. We observe that, the PTLMs have some knowledge about object affordances, but they still lack the comprehensive reasoning ability about these affordances, which is reflected in the low mAP values. Further, the performances vary across different settings. In case of NLI based setup, the fine-tuned RoBERTa and BART models show improvement in the performance, which indicates that *during fine-tuning on MNLI dataset, those models gain some reasoning ability*. In Table 4 we show the generation based results in a zero-shot setting. In case of FLAN-T5-large model, where we use it to predict a binary label (Yes/No) for an affordance class, the performance drops significantly (the accuracy is less than 7%). This shows that there are still some challenges for the text-to-text models in general reasoning ability about the object affordances. In addition, we find that, the multi-modal models do not perform well in text-only settings, despite being pretrained on text and image data. The performances of the language models get boosted when ensembling with the multi-modal models, which indicates that the prediction of object affordance from sentence is a difficult task, and can be enhanced in presence of images.

[8]During calculation we discard the cases when there is no positive class for a sentence-object pair in the ground truth. We do not find any instance where no negative class is present.

Generation based				
Predictions from generative LLM				
Model	Acc (zero-shot)		Acc (ICL)	
FLAN-T5-large	0.06		0.13±0.04	
FLAN-T5-xl	0.07		0.21±0.03	
FLAN-T5-xxl	0.33		0.39±0.04	
Falcon-7b-instruct	0.19		0.24±0.03	
Falcon-40b-instruct	**0.43**		**0.47**±0.06	
ChatGPT (GPT-3.5 turbo)	0.41		0.44±0.05	
Multi-modal models				
Model	Acc (zero-shot)		Acc (ICL)	
	IR based	IG based	IR based	IG based
Idefics-9b-instruct	0.26	0.25	0.36±0.02	0.37±0.03
llava-1.5-7b	0.32	0.34	0.36±0.03	0.40±0.04
instructblip-vicuna-13b	0.37	0.39	0.43±0.03	0.45±0.03
instructblip-flan-t5-xl	0.12	0.16	0.15±0.02	0.18±0.02
instructblip-flan-t5-xxl	0.39	**0.45**	0.48±0.04	**0.53**±0.05

Table 4: Zero-shot and in-context learning (ICL) performance for affordance prediction using generative models. IR: Image Retrieval; IG: Image Generation. Number of demonstration examples used for ICL = 5. We also mention the variance over different selections of examples. The best results are marked in bold.

5.2 Effect of finetuning on commonsense datasets

We observe that the fine-tuned model on common sense reasoning task (Table 5) show improved performance for the affordance prediction task. This indicates that the pre-trained models lack the reasoning of object affordance. We find that the smallest BERT-base model fine-tuned on PiQA, per-

MLM based		
Model	Accuracy	mAP
BERT-base-uncased-finetuned-piqa	0.45	0.26
BERT-large-uncased-finetuned-piqa	0.56	0.29
RoBERTa-large-finetuned-piqa	**0.64**	**0.45**
BART-large-finetuned-piqa	0.59	0.35

Table 5: Affordance prediction using models trained on common sense data. Best results are marked in bold.

NLI based (Normal)		
Model	Accuracy	mAP
RoBERTa-large	**0.72**	**0.49**
BART-large	0.69	0.48
MLM based		
Model	Accuracy	mAP
BERT-large-uncased	0.58	0.33
RoBERTa-large	**0.77**	**0.39**
BART-large	0.65	0.38

Table 6: Few-shot fine-tuning performances of the PTLMs. Number of training data points used: 124.

forms almost similar to that of the BERT-large or BART-large models (see Table 3).

5.3 Few-shot performance

We find that, in presence of few examples from our affordance dataset, the reasoning capability about object affordances can be enhanced for the PTLMs. The results with 124 shots (62 pairs as discussed earlier) are noted in Table 6. In Table 4, we note the results for the in-context learning performance of the generative LLMs and VLMs. We observe a significant performance gain over zero-shot settings. Having said that, we also observe that, even with the in-context learning, the performance of the generative models (with more than 7b parameters) do not reach even close to the performance of the fine-tuned BERT-large model (340M parameters). This suggests that, for the specific affordance prediction tasks from text, finetuning is absolutely essential even for the state-of-the-art LLMs and VLMs.

6 Error analysis

Encoder based models. We conducted a qualitative analysis of the erroneous cases for the two models (BART-large and RoBerta-large) in MNLI settings to understand what are the typical causes of errors. We take examples where accuracy is below 0.3. Consider the representative example below.

> *Sentence: The salt from La Mata is often used as table salt.*
> *Object: table salt*
> Top 5 predicted affordances (according to the probability score) - [‘sitOn’, ‘pourFrom’, ‘grasp’, ‘fix’, ‘lookThrough’]

The model predicts ‘SitOn’ as the top affordance for table salt, implying that the model misinterprets “table salt” with “table”. Similarly, for the object “the window sill”, the model predicts ‘lookThrough’, ‘watch’ as top affordances, which again suggests that the model is confused between “the window sill” and a “window”. In another case, the model predicts [’grasp’, ’writing’, ’typing’, ’lookThrough’, ’throw’] as the top affordance labels for the object “any rock concerts”.

Analysis of generative models. In Appendix Figure 10a, we plot the correlation between error rate made by chatGPT for each affordance classes and the classwise annotator agreement. We observe a moderately negative correlation ($\rho = -0.29$) which suggests that there is a chance that the model is making higher mispredictions where the agreement is low. Similarly we observe that the mispredictions made by chatGPT for the most frequent objects has a moderately negative correlation ($\rho = -0.58$) with the annotator agreement. The correlation is shown in Figure 10b. The trends are similar for the other LLMs. These results together indicate that the those objects and affordance classes which are hard to disambiguate by humans also pose a challenge to the most sophisticated GenAI models in predicting the correct answer.

7 Conclusion

In this paper we introduced a novel text-based affordance dataset GRAFFORD to investigate the affordance knowledge of pre-trained language models and pre-trained vision language models in different zero-shot settings. Our findings suggest that, the state-of-the-art language models, particularly text-to-text models, still exhibit limitations in their ability to reason about object affordances. Finetuning emerges as the only way to improve in such a complex task and here GRAFFORD promises to be a very valuable resource for researchers. By leveraging our dataset, future studies can contribute to enhancing the reasoning capabilities of language models and advancing the understanding of how language is grounded in the context of objects and their affordances.

Limitations

All of our experiments were conducted for English language. The models may act differently in multilingual settings. Our dataset is curated based on a specific set of affordance classes, which may introduce bias in terms of affordance representation. This could limit the generalizability of our findings to other domains or contexts. Despite efforts to train annotators and ensure agreement, subjective interpretations of affordance classes, can introduce noise. Our study primarily relies on textual information for affordance prediction. The absence of grounded visual information may limit the model's ability to accurately predict affordances, as some affordances may be more visually dependent.

Ethics Statement

We used the publicly available XNLI corpus to curate our GRAFFORD dataset. Our dataset does not contain any harmful or offensive contents. Any personal or sensitive information is anonymized and treated with utmost confidentiality. We ensure the protection of participants' privacy and obtain informed consent for data collection, annotation, and analysis. We incentivized all the annotators uniformly throughout the annotation process.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Muhannad Alomari, Paul Duckworth, Majd Hawasly, David C Hogg, and Anthony G Cohn. 2017. Natural language grounding and grammar induction for robotic manipulation commands. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 35–43.

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. PROST: Physical reasoning about objects through space and time. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4597–4608, Online. Association for Computational Linguistics.

Emanuele Bastianelli, Danilo Croce, Andrea Vanzo, Roberto Basili, Daniele Nardi, et al. 2016. A discriminative approach to grounded spoken language understanding in interactive robotics. In *IJCAI*, pages 2747–2753.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*.

Kan Chen, Rama Kovvuri, and Ram Nevatia. 2017. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 824–832.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Abhinav Gupta, Aniruddha Kembhavi, and Larry S. Davis. 2009. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(10):1775–1789.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574.

David Inkyu Kim and Gaurav S Sukhatme. 2014. Semantic labeling of 3d point clouds with object affordance for robot manipulation. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5578–5584. IEEE.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Hedvig Kjellström, Javier Romero, and Danica Kragic. 2011. Visual object-action recognition: Inferring object affordances from human demonstration. *Comput. Vis. Image Underst.*, 115(1):81–90.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not written in text: Exploring spatial commonsense from visual signals. *arXiv preprint arXiv:2203.08075*.

Jinpeng Mi, Song Tang, Zhen Deng, Michael Goerner, and Jianwei Zhang. 2019. Object affordance based multimodal fusion for natural human-robot interaction. *Cognitive Systems Research*, 54:128–137.

Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. 2015. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381. IEEE.

Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. 2016. Detecting object affordances with convolutional neural networks. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2765–2770. IEEE.

Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. 2017. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915. IEEE.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Anirban Roy and Sinisa Todorovic. 2016. A multi-scale cnn for affordance segmentation in rgb images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 186–201. Springer.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding.

Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. 2017. Weakly supervised affordance detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2795–2804.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Mohit Shridhar and David Hsu. 2018. Interactive visual grounding of referring expressions for human-robot interaction. *arXiv preprint arXiv:1806.03831*.

Yu Sun, Shaogang Ren, and Yun Lin. 2014. Object–object interaction affordance learning. *Robotics and Autonomous Systems*, 62(4):487–496.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olmpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2022. Z-LaVI: Zero-shot language solver fueled by visual imagination. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1186–1203, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhibin Yu, Sangwook Kim, Rammohan Mallipeddi, and Minho Lee. 2015. Human intention understanding based on object affordance and action classification. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE.

Yuke Zhu, Alireza Fathi, and Li Fei-Fei. 2014. Reasoning about object affordances in a knowledge base representation. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 408–424. Springer.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Data annotation

A.1 Instruction page on the Toloka platform

Figure 2 shows the guidelines/instructions, that the annotators had to follow for labelling.

A.2 Interface for labelling

A sample task interface is shown in Figure 3.

A.3 Annotators demographics

Figure 6 provides the demographic information about the annotators. We can observe that a large number of annotators (36%) are from Russia and most of the annotators having the age in between 20-35.

A.4 Phasewise annotator agreement

We plot the soft agreement[9], hard agreement[10] in Figure 7, which shows gradual increase in agreement scores.

A.5 Incentive details

During the pilot study, we provided USD 0.05 per task-suite where in each task-suite, there were 10 examples (15 affordance labels for each example) to be answered. We attempted to take feedback from the tolokers who had answered randomly (e.g., mark all the values as 0), to understand their requirements properly. Most of them suggested that a wage of $0.1 to $0.15 would be ideal for the survey.

During the main study we provided USD 0.25 per task-suite, where in each task-suite there were 5 examples to be answered. Some of them were consistently providing good answers and few of them also suggested improvement on the objects. We awarded them with an additional bonus of USD 0.5. Overall, we spent USD 777 for the annotation process.

A.6 Correlation of affordances

In Figure 9 we show the correlation between the different affordance classes.

A.7 Most frequent objects

Figure 8a shows the most frequent 15 objects in the GRAFFORD dataset.

B Correlation of ChatGPT accuracy and average human agreement

We provide the figures corresponding to the generative model analysis in Figure 10.

C Prompt selection

We use intuitive prompts for each of the setups, which are suitable for affordance related to object.

D Model implementation details

The language models and the ViLT are built on top of the huggingface API[11]. For NLI based zero-shot prediction, we use the zero-shot classification

[9] Soft agreement: Mapping Likert scale ratings to binary labels for measuring agreement by applying a threshold value.

[10] Hard agreement: Treating each Likert scale rating as a distinct label.

[11] https://huggingface.co/

Figure 2: The instruction used for annotators in the Toloka platform

pipeline [12]. We adapted the CLIP model from the OpenAI's public repo [13], and we select the ViT/B32 as the image encoder. For ViLT, we select the vilt-b32-mlm [14] model. For generative LLMs and VLMs we apply the models available on hugging-

[12] https://huggingface.co/docs/transformers/main_classes/pipelines

[13] https://github.com/openai/CLIP

[14] dandelin/vilt-b32-mlm

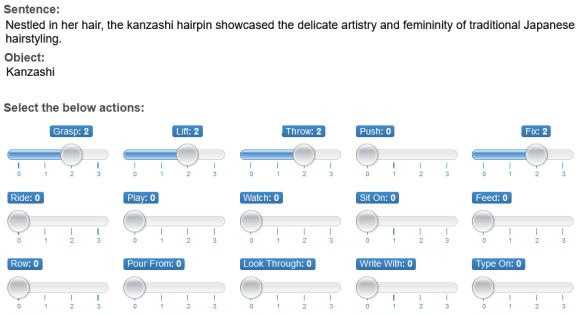


Figure 3: The sample task interface used for the annotators in the Toloka platform

Model	Prompt used
FLAN-T5	consider {sentence}. Now, can human {affordance} the {object_name}? Answer Yes or No:
Falcon	"""You are a helpful AI assistant. Answer only "Yes" or "No" for the question based on the given context. Context:sentence \n »QUESTION« Can human {affordance} the {object_name}? \n »ANSWER«""".strip()
I-BLIP, IDEFICS, LLaVA	consider the sentence {sentence}. Now from this information, can human {affordance} the {object_name}? Accompanying this query is an image of the object_name. Note that the image may contain noise or variations in appearance. Given the textual description and the image, answer Yes or No whether the human can {affordance} the {object_name}. Answer: "

Table 7: Prompt format used by different models for the prediction. I-BLIP: InstructBLIP.

face [15]. All the experiments were conducted on 2x NVIDIA RTX 4090 GPU server.

E Dataset creation time

Annotating affordances about the object from a text itself is a difficult and very subjective task. It took approximately 5 months for completing the extraction of noun-phrases from xnli data, filtering objects, selecting skillful tolokers and training, and then final phase-wise annotation after rigorous review process.

F Sample dataset

Figure 11 shows a sample of GRAFFORD dataset

G Additional experiments

G.1 Qualitative analysis of generated images

We conducted a qualitative analysis on 50 randomly sampled objects and their corresponding generated images. Two annotators (one Phd student and one

[15] https://huggingface.co/models

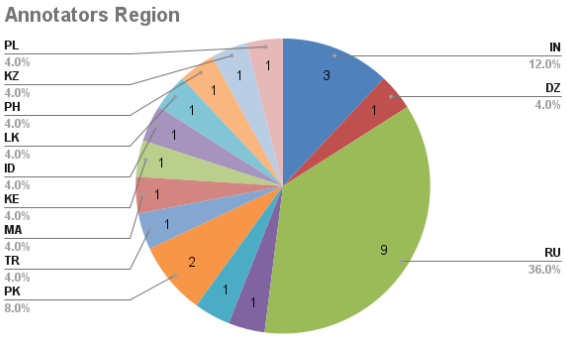


Figure 4: Country distribution of the annotators

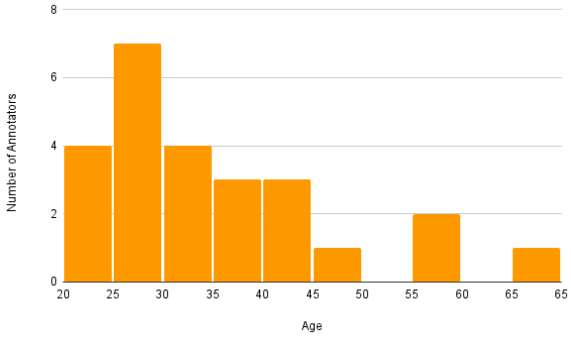


Figure 5: Age distributions of the annotators

Figure 6: The Annotators Demographics

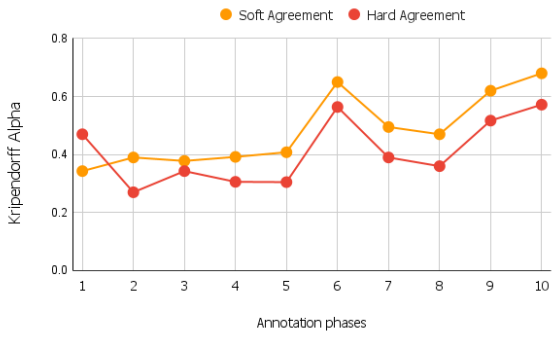
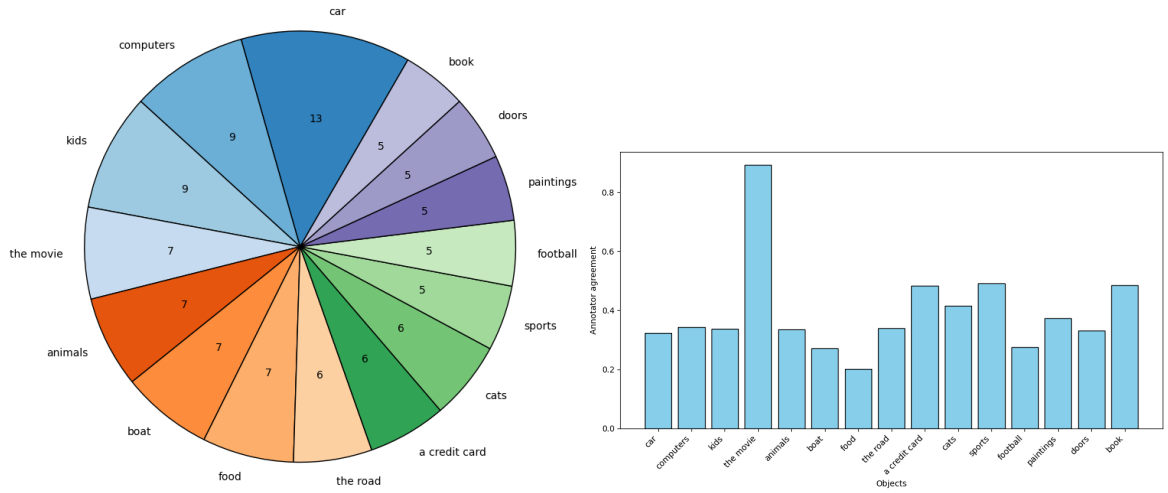


Figure 7: Phase-wise annotator agreement.



(a) (b)

Figure 8: (a)Most frequent 15 objects and their corresponding frequency in the GRAFFORD dataset. (b)Annotator agreement for the most frequent 15 objects.

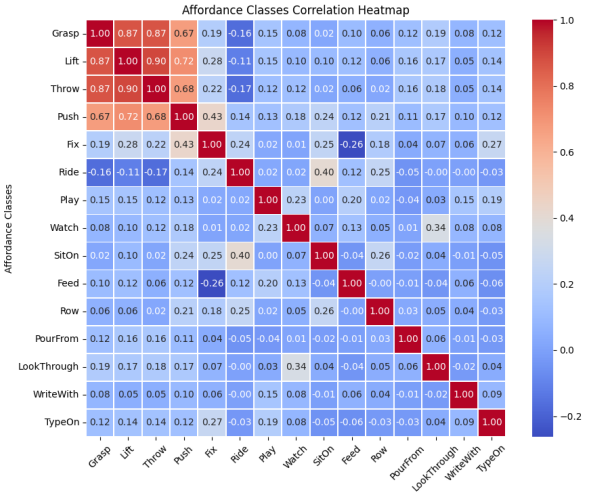
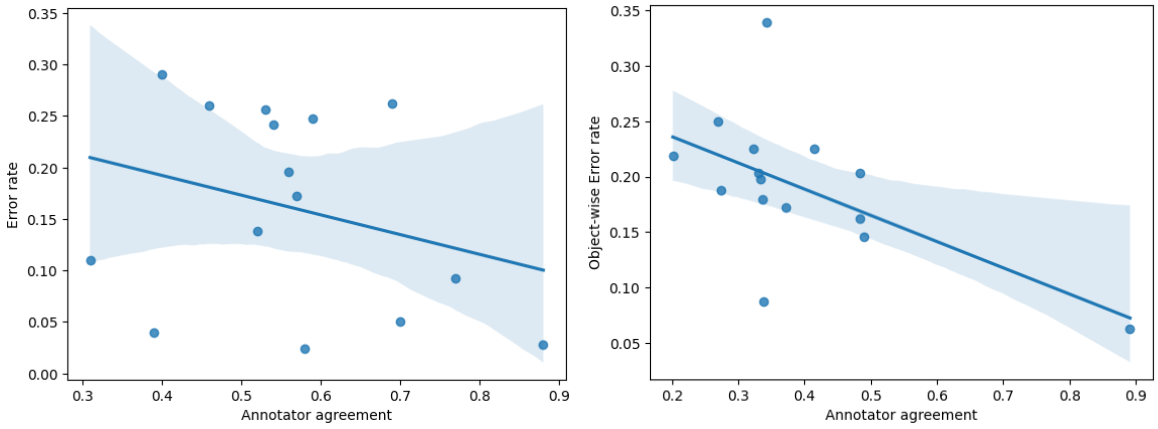


Figure 9: Correlation between each of the affordance classes.

undergrad student) marked each of the 5 generated images as 1 or 0 according to their relevance and non-relevance to the object respectively. We considered the image as relevant if both of the annotators marked that image as 1. We achieved an Acc@1 of 0.2, Acc@5 of 0.88 and an MAP@5 of 0.36. Which suggests that in most of the cases there are relevant images in the top-5 generated images. In our pursuit of assessing the statistical significance of our sampled data (i.e., the 50 examples), we embarked upon a rigorous hypothesis testing procedure utilizing the binomial distribution. Within our specific context, we accorded greater significance to the top-5 accuracy metric, which demonstrated an impressive achievement of 0.88. This signifies that among the 50 selected examples, in 44 instances, at least one of the five generated images displayed relevance to the object under consideration.

Guided by this success rate, we proceeded to conduct a meticulous hypothesis test employing the binomial distribution. We assumed an expectation of success at 0.75. The outcome of this statistical analysis revealed a p-value of less than 0.02, thereby underscoring the statistical significance of our success rate.



(a) (b)

Figure 10: (a) Correlation between average classwise error rate made by chatGPT and the annotator agreement. ($\rho = -0.29$) (b) Correlation between frequent object wise error rate made by chatGPT and the annotator agreement. ($\rho = -0.58*$). *indicates a p-value < 0.05.

Sentence	Object	Grasp	Lift	Throw	Push	Fix	Ride	Play	Watch	SitOn	Feed	Row	PourFrom	LookThrough	WriteWith	TypeOn
This diablo only comes out to slaughter the cattle .	cattle	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0
Delivery points should include at least a bench and a locked storage compartment .	bench	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0
There are four fences , and you can only go past the second one if you are a member of the imperial family , or a high-ranking priest .	fences	1	0	1	1	0	0	0	1	0	0	0	0	0	0	0
Users are excited about being able to share their own events on the calendar page .	calendar page	1	1	1	1	0	0	0	1	0	0	0	0	1	0	0
White ran towards where the people were hitting each other with swords .	swords	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
The cat ate every kind of fish except tuna .	fish	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0
The snake was hissing underneath the deck .	deck	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
On the higher levels of the town hall , Umbrian and Tuscan paintings are on show .	the town hall	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
He couldn 't follow up because his mouth was gagged by a group of mercenaries .	mercenaries	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0
A gristle gun is featured .	gristle gun	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 11: Example snapshot of GRAFFORD dataset.