PhyGrasp: Generalizing Robotic Grasping with Physics-informed Large Multimodal Models

Author Names Omitted for Anonymous Review. Paper-ID 475

Abstract—**Robotic grasping is a fundamental aspect of robot functionality, defining how robots interact with objects. Despite substantial progress, its generalizability to counter-intuitive or long-tailed scenarios, such as objects with uncommon materials or shapes, remains a challenge. In contrast, humans can easily apply their intuitive physics to grasp skillfully and change grasps efficiently, even for objects they have never seen before.**

This work delves into infusing such physical commonsense reasoning into robotic manipulation. We introduce PhyGrasp, a multimodal large model that leverages inputs from two modalities: natural language and 3D point clouds, seamlessly integrated through a bridge module. The language modality exhibits robust reasoning capabilities concerning the impacts of diverse physical attributes on grasping, while the 3D modality comprehends object shapes and parts. With these two capabilities, PhyGrasp is able to accurately assess the physical properties of object parts and determine optimal positions and angles for grasping. Additionally, its language comprehension enables it to interpret human instructions, facilitating the output of grasping poses aligned with human preferences. For training PhyGrasp, we construct a dataset PhyPartNet with 195K object instances with varying physical properties, alongside their corresponding language descriptions of physical properties and human preferences. Extensive experiments conducted in both simulators and real robots demonstrate that PhyGrasp achieves state-of-the-art performance, particularly in long-tailed cases, *e.g.*, about 10% improvement over GraspNet. More demos and information are available on our anonymous webpage.

I. Introduction

Human-like embodied intelligence represents an important milestone in the realm of robotic manipulation, offering practical applications such as household robots designed to assist with our daily tasks. Despite notable advancements that have been made [13, 46], the current capabilities of robots still lag far behind humans, particularly in physical commonsense reasoning and generalizability [4]. Humans possess inherent multimodal reasoning abilities and an intuitive sense of physics, enabling them to plan actions accurately by leveraging such commonsense knowledge, and easily generalize to uncommon even counterfactual objects or situations. For example, as illustrated in Figure 1, humans intuitively understand the need to grasp the base when lifting a monitor and recognize the fragility of the display, realizing that mishandling it could lead to breakage. Existing robot grasping techniques lacking physical common sense may inadvertently disregard these principles, potentially resulting in damage. Incorporating physical common sense into robotic systems can mitigate this issue. How to empower robots with such capabilities to handle long-tailed objects and scenarios becomes an important challenge.

Previous methods for robotic grasping and manipulation generally fall into two streams. 1) The first stream directly estimates low-level robotic actions or trajectories for execution [6, 7, 51]. These methods typically rely on large-scale data for training, resulting in models that struggle to generalize to novel scenarios or robotic platforms. 2) To improve the generalizability, the second stream [18, 17] proposes to employ analytical methods or learned models to predict affordance maps or grasping pose proposals. Subsequently, it plans low-level robotic actions based on these estimated affordances or poses. The underlying motivation is that grasp poses are easier to generalize than robotic action sequences. Nonetheless, existing grasping pose detection algorithms often concentrate on the analysis of 3D shapes and semantics of objects, while overlooking part information, physical senses, and constraints. Consequently, they still face challenges in generalizing to objects with diverse physical properties and long-tailed scenarios. The incorporation of physical commonsense remains a fundamental aspect that is largely unexplored within existing robotic grasping frameworks.

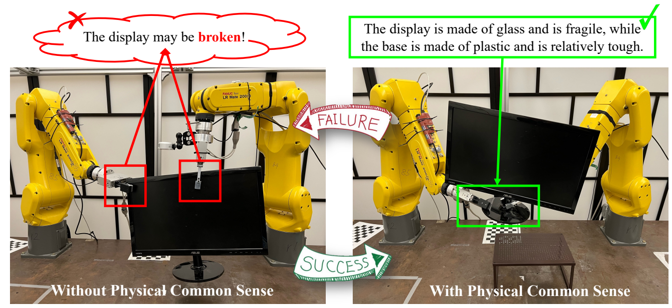


Fig. 1: Motivation of our PhyGrasp. Current general robot grasping policies (left) typically predict the pose and position for grasping based solely on the object's 3D shape, neglecting its physical properties. This oversight can lead to potential damage to the display. In contrast, integrating physical common sense into robotic systems (right) can address this issue effectively.

In recent times, the rapid evolution of large language models (LLMs), such as ChatGPT, has showcased robust understanding and generalization capabilities, holding promise for physical commonsense reasoning. However, these models lack perceptual environmental information, such as detailed parts and shapes from 3D vision, which poses a challenge in utilizing LLMs for real grasping applications. While some vision-language models (VLMs) have been proposed to provide vision information for LLMs, their focus has predominantly been on visual question-answering tasks, leaving them ill-equipped to reason effectively about the physical world, particularly within domains like robotic grasping and manipulation. Considering 3D models such as PointNet and PointNet++, which offer substantial insight into object shapes and poses in the physical world, an intuitive solution emerges: building a multimodal model that bridges the 3D and language modalities. This integration aims to facilitate a comprehensive physical reasoning of objects in robotic grasping tasks.

In practice, it's non-trivial to train an interface between the 3D and language modalities, due to its data-intensive nature and underrepresentation in standard multimodal pretraining datasets. Existing datasets and benchmarks typically either concentrate solely on grasping without considering the underlying physical concepts (*e.g.*, material, fragility, mass, friction) [16], or they focus on high-level physical understanding without addressing low-level grasping estimation [20], restricting their usefulness for robotic grasping and manipulation tasks. Our objective is to address this problem from both sides.

Motivated by the above observations, in this work, we construct a physically grounded 3D-language dataset, termed PhyPartNet. It contains 195K unique object instances featuring various physical properties across their parts based on PartNet. For each object instance, we sample physical attributes, such as material, fragility, mass, density, and friction, for individual parts of the object. Subsequently, we generate corresponding grasping probability maps using analytical grasping solutions, along with machine-generated language instructions and preferences, which are then verified by humans.

Based on PhyPartNet, we introduce PhyGrasp, a multimodal model designed to serve as an interface between LLMs and 3D encoders, effectively bridging and grounding high-level physical semantics and language into low-level grasping maps. PhyGrasp employs frozen PointNext [54] and Llama 2 [62] as its encoders, coupled with a carefully crafted bridge module capable of integrating information from language, visual local, and visual global representations to generate final predictions. It offers several appealing benefits. Firstly, it predicts grasping poses based on both language descriptions and 3D information regarding an object's physical properties, such as material, fragility, mass, density, and friction. Secondly, its language comprehension enables the interpretation of human instructions, facilitating the output of grasping poses aligned with human preferences. Lastly, it demonstrates strong generalizability to long-tailed, unseen, and even counterfactual objects.

Our primary contribution is PhyGrasp, which generalizes robotic grasping through the integration of physics-informed large multimodal models. For the first time, we facilitate grasping pose detection by leveraging the object's part-level physical properties. We conduct experiments in both simulators and real robots to demonstrate the effectiveness of PhyGrasp. Another contribution is our PhyPartNet dataset, a comprehensive collection of large-scale 3D mesh instances featuring diverse part-level physical attributes and corresponding language annotations. We aspire for our work to inspire future research in robot grasping, particularly among those inclined towards physical reasoning and interactions.

II. Related Work

1) Physical Reasoning: Previous work's focus was on estimating the physical properties of objects through visual perception, using interaction data as a primary source of learning [68, 69, 34]. A distinct body of research has shifted towards developing representations that encapsulate physical concepts, going beyond direct property estimation [25, 72]. Notably, methods [41, 33, 20] explore physical reasoning using LLMs and VLMs, *e.g.*, [20] introduces a dataset specifically designed to quantify and enhance object-centric physical reasoning capabilities. Moreover, OpenScene [52] employs CLIP [56] to discern objects within scenes based on attributes like material composition and fragility. However, they focus on high-level physical understanding without addressing low-level grasping estimation, restricting their usefulness for robotic grasping and manipulation tasks. This work introduces PhyPartNet, which not only underpins our methodology but also facilitates advancements in robotic manipulation by providing a more nuanced understanding of physical properties and their implications for robotics grasping.

2) Large Multimodal Models: The community has witnessed the emergence of multimodal large language models (MLLMs), designed to augment the capabilities of traditional language models by incorporating the ability to process and understand visual information [81, 80, 82, 70, 61, 2, 86, 32, 30, 74, 11, 31, 79, 35, 75]. Among these, Flamingo [2] stands out by utilizing both visual and linguistic inputs to demonstrate impressive few-shot learning capabilities, particularly in visual question-answering tasks. Building on this foundation, advancements have been made with the introduction of models like GPT-4 [50], the LLaVA series [39, 42, 38], and MiniGPT-4 [83], enhancing visual language large models (VLLMs) through visual instruction tuning. This innovation has significantly improved these models' ability to follow instructions, a crucial aspect for applications requiring precise interaction with visual content. Simultaneously, a new wave of models [67, 53, 3, 66, 10] has been developed to strengthen the visual grounding capabilities of VLLMs. These advancements facilitate more nuanced tasks such as detailed region description and precise localization, underscoring the growing sophistication of these systems in interpreting and interacting with visual data. Despite these significant strides in the development of MLLMs and their enhanced ability to integrate and interpret multimodal data, there remains a notable gap in their application to physical reasoning, particularly in the context of robotic grasping. This gap highlights a pivotal area for future research, where the potential for MLLMs to contribute to the understanding and execution of complex physical interactions can be further explored and realized.

3) Large Models for Robot Learning: Leveraging large pre-trained models holds promise for creating capable robot agents. Numerous works focus on using language models for planning and reasoning in robotics [22, 1, 9, 59, 21, 57, 60, 37, 64, 14, 15, 77, 43, 65, 55]. To enable language models to perceive physical environments, common approaches include providing textual descriptions of scenes [23, 78, 59] or access to perception APIs [36]. Vision can also be incorporated by decoding with visual context [24] or using multi-modal language models that directly take visual input[15, 49, 48, 73]. In this work, we leverage the strong capabilities of vision and language models for physical common sense reasoning, thereby for the first time, enabling physics-informed robotic grasping.

4) Grasp Pose Detection: The domain of vision-guided grasp pose detection has become a focal point in robotics research, representing a shift from traditional top-down grasping techniques to the thorough exploration and implementation of six degrees of freedom (6 DOF) grasping methods. This evolution is underscored by notable contributions in the field, exemplified by advancements documented in [44, 85, 45] for planar grasping. It is further propelled by the introduction of sophisticated 6 DOF methodologies in studies such as [5, 26, 84].Central to this progression is the development of state-of-the-art 6 DOF grasp pose detection models, particularly exemplified by AnyGrasp [18]. AnyGrasp extracts and encodes geometric features of objects from point clouds, achieving a success rate in object grasping that parallels human capabilities. Leveraging the grasp poses identified by AnyGrasp, subsequent research endeavors have been proposed, concentrating on specific object grasping [40, 27]. These efforts have extended to articulated object manipulation tasks [76] as well. However, these investigations often assume fixed physical parameters of objects or aim to identify a universally robust grasp amidst varying physical uncertainties. Such assumptions may lead to impractical or hazardous grasping scenarios, particularly when dealing with delicate object parts, a challenge exacerbated by the limited ability of vision sensors to discern material properties.To address these challenges, an innovative approach is proposed: integrating physical parameters into the grasp planning algorithm through natural language descriptions from human guidance. This approach allows the network to adjust its planning outcomes based on the articulated physical characteristics of objects, thereby enhancing the practicality and safety of robotic grasping operations.

III. Dataset Generation

We develop a dataset that enables robots to learn physical reasoning for grasping objects. This dataset includes object point clouds for visual processing, language summaries, and corresponding analytical grasping solutions. The left side of Figure 3 summarizes the data generation process. For each object, we generate multiple instances where different parts of the object have different physical attributes (e.g., material, density, mass, friction). Section III-A provides details of the statistics of the dataset. We use analytical methods (refer to Section III-B) to calculate force closure grasp pairs and construct a grasping affordance map, which can serve as the ground truth grasping solution for robot learning. As described in Section III-C, we use OpenAI's ChatGPT API to provide descriptive summaries of objects, highlighting different physical attributes in each grasp instance.

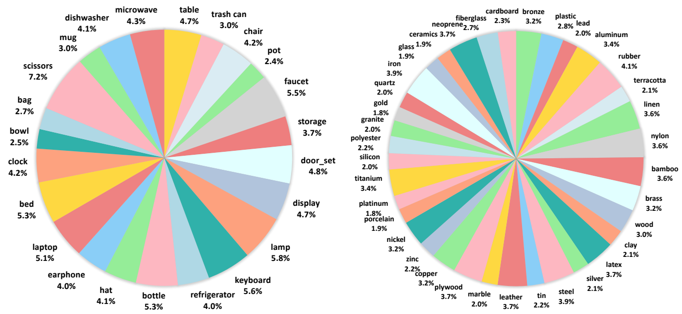


Fig. 2: Statistics for the dataset. The left and right figures denote instance distributions among objects and materials, respectively.

A. Dataset Statistics

We build our dataset based on the PartNet dataset [47], which comprises 28,599 objects across 24 categories, each featuring part segmentation. For every object, we generate multiple instances, varying the materials of different parts. We introduce 16 materials, each associated with unique physics attributes: density, friction, and fragility. These attributes enable us to compute the mass, center of mass, and maximum normal force applicable to each surface of the object. Additionally, we assign varying levels of grasping probability to each part, reflecting human common sense (e.g. human will not grasp knife blade). In total, we create 193,856 unique instances, with equal distribution among objects and materials (refer to Figure 2).

The training, validation, and testing set have 173,856, 10,000, and 10,000 instances, respectively. In addition, we prompt ChatGPT to pick a "hard set" that is a subset of the general testing set and contains 370 the most counter-intuitive instances.

B. Analytical Grasping Solutions

A grasp, denoted by g, achieves force-closure if, for any external wrenches (i.e., forces and torques, F_{ext}) applied to the object, there exist contact forces f_c within the contact friction cone K_g that counterbalance the external wrenches, satisfying $Gf_c = F_{ext}$. Here, G represents the grasp mapping matrix, which is contingent upon the grasp's location, g, and the magnitude of f_c can be arbitrarily large [58].

In this study, we identify potential grasp candidates by employing a ray-shooting technique around the object to determine contact pairs, thereby conceptualizing a parallel grasp, g. The grasp mapping matrix G is then formulated based on g's positioning relative to the object's center of mass (CoM). To assess the force-closure property of a grasp, we employ the following optimization problem:

$$
\begin{aligned}
&\min_{f_c} ||f_c|| \\
&s.t.\ Gf_c = F_{ext} \\
&\quad f_c \in K_g
\end{aligned}
\tag{1}
$$

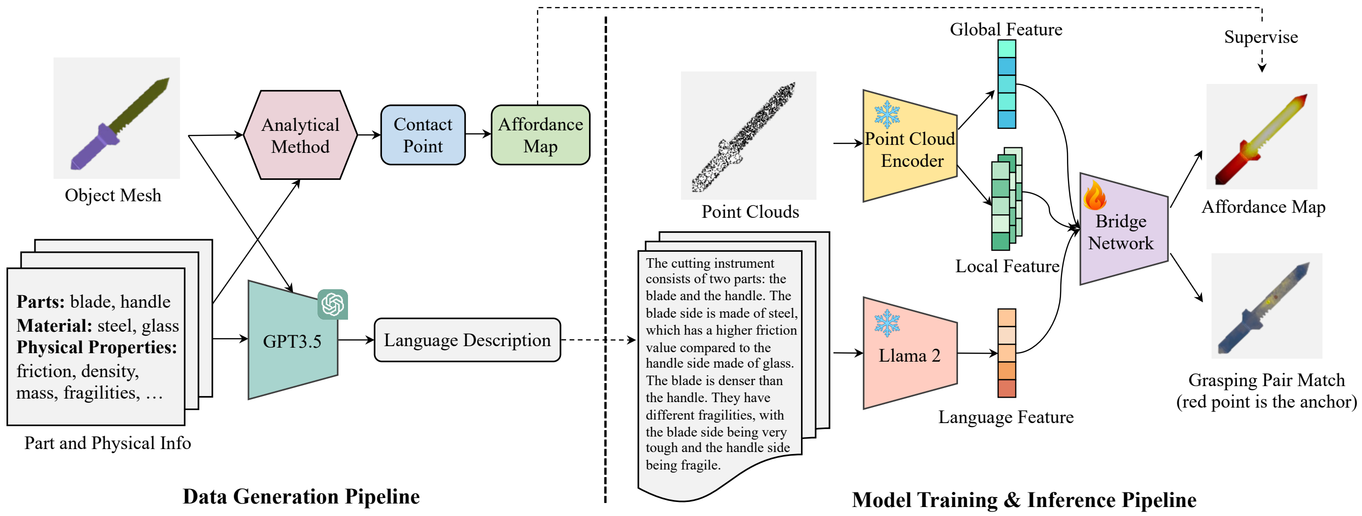


Fig. 3: An overview of our PhyPartNet generation pipeline and our PhyGrasp framework. Given object meshes sampled from PartNet, we leverage GPT3.5 and an analytical method to automatically generate the grasping affordance map and language descriptions for the object instance. The generated data is then human-verified, forming our PhyPartNet. We freeze PointNext [54] and Llama 2 [63] and tune the bridge network during training on PhyPartNet. After training, PhyGrasp is able to generalize to novel 3D point clouds and new natural language instructions.

While we can ascertain the force-closure status using simpler methods, as indicated in [44, 84], the formulation in Eq. (1) enables the incorporation of additional constraints reflective of the object's physical characteristics. Specifically, we consider the maximum permissible contact force ($|f_c| \leq \epsilon$), the variation in friction coefficient across the object's surface ($K_g \propto \mu$), and adjustments to the object's center of mass ($G \propto$ CoM).

We compute the feasibility of the solution to Eq. (1) to verify whether a grasp pair is force-closure and complies with other physical prerequisites.

we construct a grasp affordance map by assigning a Gaussian distribution on each grasping location. We normalize the resulting sum, so the grasping probability of each point on the mesh is under a mixture of Gaussian distribution. The left column of Figure 5 illustrates the resulting affordance map. For each instance, we sample 2,048 points on the surface of the object mesh as the input for vision processing.

With analytical grasp pairs, we create a grasp affordance map by allocating a Gaussian distribution to each grasping location. The normalized sum of each point on the object mesh represents the grasping probability and follows a mixture of Gaussian distributions. The left column of Figure 5 illustrates the resulting affordance map. We also save 2,048 points sampled from the surface of each object mesh as object point cloud for each instance for future vision processing.

C. Language Summary Generation

For objects composed of multiple parts with varying materials and physics attributes, we utilize OpenAI's ChatGPT API to generate language descriptions that summarize each instance, emphasizing the distinct physical attributes in every grasp scenario. We prompt ChatGPT to create a list of common materials, each characterized by specific values for density, friction, and fragility using its common sense. After we generate all instances based on the material list, we supply ChatGPT with manually crafted prompts and examples (refer

```
Role: you are a grasping analytical assistant, skilled
in summarizing the features of different objects and
materials with a natural language.
  You should provide as much information as possible with
minimal words.
  You focus on the important features of every part with
their materials, rather than the specific values.
  You will be given a paragraph describing the object and
its parts with their materials, densities, frictions,
fragilities, and human grasp probabilities hint.
  You should follow such rules:
1. Names: Describe the object & material names precisely.
2. Densities: Point out the densest part or the lightest
part. If the density difference is not obvious, you can
ignore it.
3. Frictions: Point out the part with the highest
friction and the lowest friction. If the friction
difference is not obvious, you can ignore it.
...
  I will give you some examples.
  examples ...
Instruction: Please process the following paragraph.
Output in one paragraph.
```

Listing 1: An example of the prompt used for GPT3.5.

```
Input: There is a faucet, it has several parts including a
 switch, a frame, and a spout. The material of each part
 is plastic, brass, and fiberglass, with friction: 0.4,
 0.38, 0.6, density: 1400, 8530, 2020, fragility: normal,
 tough, normal.

Output: The faucet has three parts: switch, frame, and
 spout. The spout is made of fiberglass with the highest
 friction. The switch's material is plastic and the frame
 is made of brass with the highest density.
```

Listing 2: A human example of the language description for GPT3.5 prompt.

to List 1 and List 2) that effectively illustrate the instances. This approach helps ChatGPT in understanding the relevant terms and enables it to accurately describe the remaining instances in our dataset.

IV. Learning Methods

With our dataset, we are able to train a neural network in the intricacies of robot grasping grounded with physical reasoning. The training regimen starts with a large vision model and a large language model (see Section IV-A), both of which work in tandem to encode our dataset into visual and linguistic features. We then construct a bridge network (see Section IV-B) that takes these features as input and yields a grasping affordance map, as well as a complementary classifier to generate an array of corresponding grasping pairs for each point on an object's point cloud. Section IV-C details the losses we use for training.

A. Feature Extraction

1) Vision Encoder: We use the PointNeXt architecture [54] to transform an object's point cloud into global and local visual features. With a PointNeXt encoder pre-trained on the ModelNet40 dataset [71], we extract a global feature vector with a shape of (1024,) for each object's point cloud. Since ModelNet40 dataset contains different objects from those in our dataset, these global features facilitate our model's ability to generalize to objects out of the domain of our dataset. For the extraction of local features, we leverage the PointNeXt encoder in conjunction with its part segmentation decoder, outputting local features of dimension (64,) for each point within the point cloud. The encoder-decoder pair, having been trained on PartNeXt—the same dataset that underpins our work—embeds detailed part segmentation information within the local features, enhancing our network's capacity to discern the variations among different parts of an object.

2) Language Encoder: We utilize Llama [62] to encode the language descriptions of each instance into linguistic features. Opting for the representation from the model's 20th layer, as indicated by the findings in [87], which demonstrated optimal outcomes for feature extraction, we obtain features with a dimension of (4096,).

B. Bridge Network

Our bridge network uses the extracted features to predict grasping solutions. Figure 4 illustrates the structure of our bridge network. We use a multilayer perceptron (MLP) to compress both the global visual and linguistic features down to a dimension of (128,) and mix them with another MLP to generate a global feature of (64,). In a parallel process, we refine the local visual feature through an MLP and amalgamate it with the global features and the object's point cloud, culminating in a composite feature vector of (64+64+3,) for each point on the point cloud. We then deploy two distinct MLPs: one functions as a predictor to generate a grasp affordance map, while the other acts as a classifier to identify corresponding grasp pairs using embeddings.

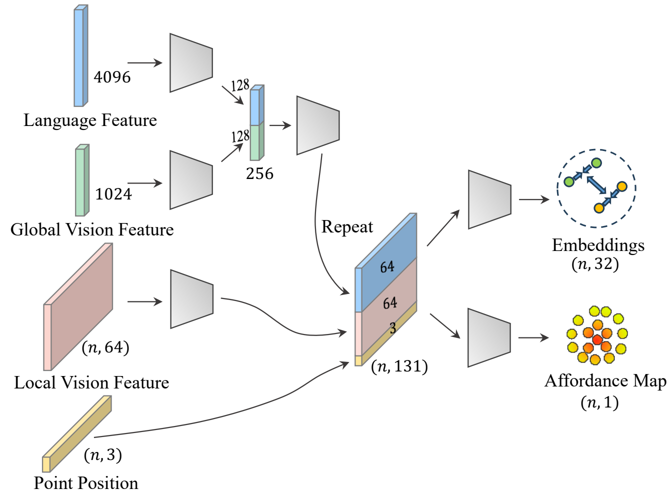


Fig. 4: The architecture for the bridge module of PhyGrasp. It outputs the grasping probability (affordance map) and the pair embedding for each point.

C. Losses

We introduce 3 different loss functions and balance them with Automatic Weighted Loss (AWL) [29] in Eq. (5). We use the following definitions: N is the number of instances, K_i^{p} is the number of positive grasp pairs for ith instance, while K_i^{n} is for negative pairs. δ_p and δ_q are respectively the margins for the positive and negative embedding loss. $||\cdot||$ is the L1 or L2 distance, and $[x]_+ = \max(0, x)$ denotes the hinge.

The first loss function is global loss L_{g}, where G_i is the ith output affordance map by our model and G_i^{gt} is the corresponding ground truth in our dataset.

$$L_{\mathrm{g}} = \frac{1}{N}\sum_{i=1}^{N}||G_i - G_i^{\mathrm{gt}}|| \tag{2}$$

The second loss function is L_{emb}, which is a linear combination of its positive part $L_{\mathrm{emb}}^{\mathrm{p}}$ and negative part $L_{\mathrm{emb}}^{\mathrm{n}}$ in Eq. (3). $Q_{i,k}$ is our model's embeddings output for the kth grasp pair of ith instance.

$$\begin{aligned} L_{\mathrm{emb}}^{\mathrm{p}} &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K_i^{\mathrm{p}}}\sum_{k=1}^{K_i^{\mathrm{p}}}[||Q_{i,k,1} - Q_{i,k,2}|| - \delta_p]_+^2 \\ L_{\mathrm{emb}}^{\mathrm{n}} &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K_i^{\mathrm{n}}}\sum_{k=1}^{K_i^{\mathrm{n}}}[\delta_n - ||Q_{i,k,1} - Q_{i,k,2}||]_+^2 \end{aligned} \tag{3}$$

$$L_{emb} = \lambda_p \cdot L_{\mathrm{emb}}^{\mathrm{p}} + L_{\mathrm{emb}}^{\mathrm{n}}$$

The third loss function L_{seg} is for segmentation with embeddings. We use a MLP classifier to segment the pairs by their grasping probability leveraging the output embeddings. This classifier takes a pair of embeddings as input, whose dimension is (64,), and outputs a grasp score M. Therefore, $M_{i,k}$ is the grasp score for the kth pair of ith instance, and

TABLE I: The grasping success rate (%) evaluated in the simulation for baseline models and our model.

Method	General Set		Hard Set	
	Top1	Top5	Top1	Top5
Analytical (upper bound)	78.0	92.1	70.0	87.6
GraspNet [16]	56.4	83.2	50.5	77.6
VGN [5]	34.1	45.9	33.6	43.5
PhyGrasp (Ours)	**61.5**	**86.0**	**59.7**	**79.2**

TABLE II: Ablation study of our model. We report the grasping success rate (%) evaluated in the simulation.

Method	General Set		Hard Set	
	Top1	Top5	Top1	Top5
Ours	**61.5**	86.0	**59.7**	79.2
Ours w/o Local	46.4	81.3	44.6	76.5
Ours w/o Global	60.2	86.3	53.5	**79.4**
Ours w/o Language	61.0	**86.7**	55.1	77.8

$M^{\text{gt}}_{i,k}$ is the ground truth, which is 1 for positive pair and 0 for negative one.

$$L_{\text{seg}} = -\sum_{i=1}^{N} \sum_{k=1}^{K_i^{\text{p}}+K_i^{\text{n}}} \log p(M_{i,k} = M^{\text{gt}}_{i,k}) \quad (4)$$

$$L = \text{AWL}(L_{\text{g}}, L_{\text{emb}}, L_{\text{seg}}) \quad (5)$$

V. Experiments and Evaluation

We conducted experiments in both simulation (Section V-A) and real-world (Section V-B) to evaluate the performance of our method in robot grasping.

A. Simulation Experiments

1) Settings: We conducted simulated experiments using PyBullet [12]. We used two gripper fingers to pinch the object at the predicted grasping positions, directed towards each other. For each instance, we evaluated the top-n predictions from models, considering any trial where the object remained secure between the fingers as a successful grasp.

2) Baselines:

- Analytical (upper bound) refers to the analytical grasping solutions in each instance. Evaluating this baseline helps in quantifying the gap between analytical predictions and their practical simulation outcomes.
- GraspNet [16] is a baseline for general object grasping. It uses a convolutional neural network to predict grasp instances directly from point clouds, providing a comprehensive and efficient approach to robot grasping.
- Volumetric Grasping Network (VGN) [5] constructs a Truncated Signed Distance Function (TSDF) representation of the scene and outputs a volume of the same spatial resolution, similar to the grasping affordance map.

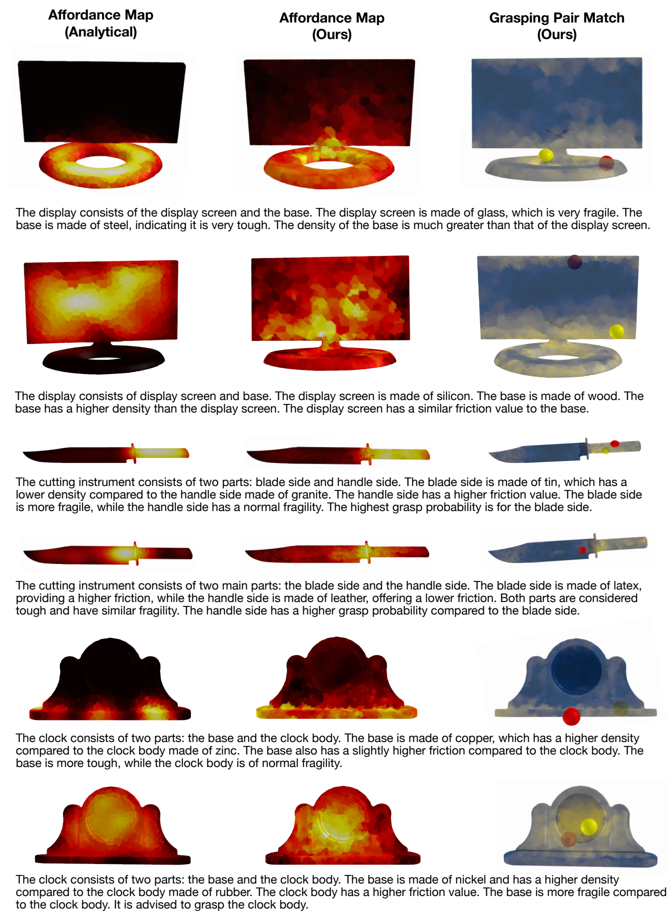
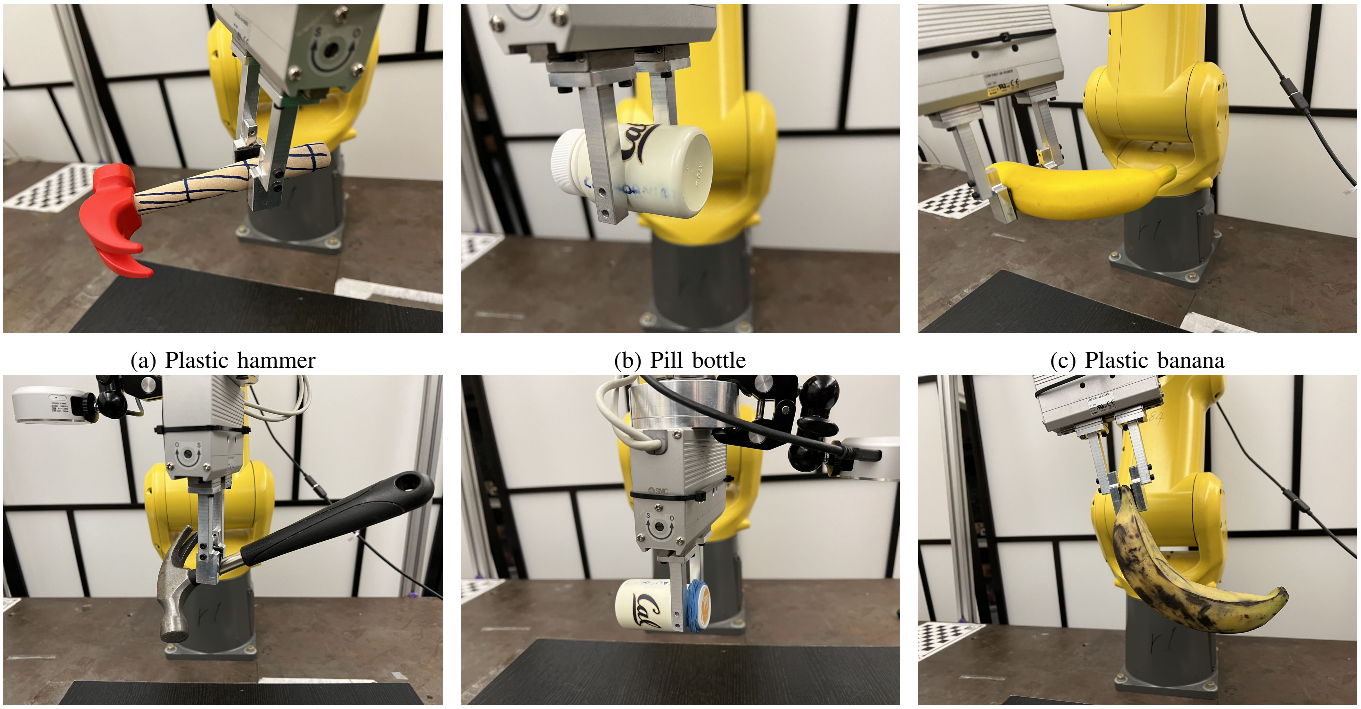


Fig. 5: Visualizations of the affordance map and grasping pair match map for our method. The left column is the affordance map of the analytical method (ground truth), the middle column is our affordance map, and the right column is the grasping pair match map. We observe that our affordance map prediction exhibits high quality and closely resembles the ground truth. In the match map, the intensity of yellow coloration indicates the confidence level, with the red point representing an anchor and the yellow point representing the top-1 prediction to be paired with the anchor.

3) Results: Table I summarizes the grasping success rate evaluated in simulation for the baseline models and our model. Our model outperforms all baselines in every metric. VGN underperforms on our dataset, particularly with large objects with multiple surfaces like tables, chairs, and beds, due to the difficulty in constructing TSDFs for these items. Additionally, its heavy reliance on visual features makes it prone to failure in scenarios where physical factors alter grasping strategies. GraspNet exhibits slightly lower performance than ours in the general test set. However, its performance drops by more than 5% on the hard set, whereas our model maintains its effectiveness. Since the hard set includes the most counter-intuitive examples, this indicates that our model effectively comprehends language descriptions and reasons about physical attributes to adapt its grasping strategy. In contrast, GraspNet,



(a) Plastic hammer (b) Pill bottle (c) Plastic banana

(d) Steel hammer (e) Pill bottle with rubber band (f) Ripe banana

Fig. 6: Our real-world experiments. We select three representative objects with different physical properties. Our model accurately predicts locations that align with our expectations during setting these testing scenarios. For instance, it effectively estimates the center of gravity of a hammer with various materials and plans graspings. Moreover, it recognizes that grasping the rubber portion of the pill bottle provides greater stability.

TABLE III: Comparisons of grasping affordance map accuracy under different metrics.

Method	KLD ↓	SIM ↑	AUC-J ↑
VGN	5.2622	0.4452	0.5026
Ours	0.3783	0.7306	0.8545

which depends solely on vision, is likely to struggle with these long-tailed edge cases.

We provide qualitative results for our model's predictions of afforance map and grasping pair match in Figure 5. Visually, these predictions closely resemble analytical solutions. We further explore the impact of physical attributes on the same object, as exemplified by two clocks: the top clock features high friction and low fragility at its base, while the base of the bottom clock is low in friction and is fragile. Our model successfully captures this information and identifies the correct part to grasp. The grasping pair match highlights the efficiency of our embedding and classifier, with the anchor and query points forming a force-closure grasp, thereby enhancing the grasping success rate.

In Table III, we report the Kullback-Leibler Divergence (KLD) [19], the Similarity metric (SIM) and the Area Under the Curve (AUC-J) [8, 28] to evaluate the effectiveness of the predictions of affordance map. These metrics evaluate the discrepancy in the distribution of heatmaps or affordance maps in relation to grasping probability for both our method and VGN. The results indicate that our method outperforms VGN in generating more accurate grasping affordances.

4) Ablations:

- Ours w/o Local: Eliminating local vision features significantly impacts our model's capability to discern part segmentation information. This limitation hinders its ability to prioritize grasping parts with a higher probability of successful grasp, leading to the most notable performance drop.
- Ours w/o Global: Excluding global features results in a relatively minor impact on our model's performance. This is understandable since the encoder is pretrained on ModelNet40, which differs from our objects. While this approach aids in generalizing to unseen objects, as demonstrated in our real-world experiments, it wasn't explicitly evaluated in simulation tests.
- Ours w/o Language: Omitting language features leads to minimal performance changes in the testing set but results in failure in the hard set. In more general instances, the model can rely on vision features for identifying safe grasps. However, in the counter-intuitive instances, language information becomes crucial to ensure successful grasping.

TABLE IV: The top 5 grasping success rate (%) evaluated in the real world for GraspNet and our model.

Method	Scenario	Banana	Hammer	Bottle	Overall
GraspNet	Normal	0.2	0.2	1.0	0.5
	Challenging	0.0	0.2	0.0	0.2
Ours	Normal	0.4	0.6	1.0	0.7
	Challenging	0.6	0.6	1.0	0.7

B. Real-world Experiments

1) Settings: In our experiments, we compared our method with GraspNet using two bananas, a pill bottle, and two hammers, representing both standard and challenging grasping scenarios (refer to Figure 6). For bananas, the standard scenario was unrestricted grasping, whereas the challenge was grasping only the stem of an overly ripe banana without causing damage. The pill bottle's challenge involved a rubber-banded cap and a low-friction body, requiring cap grasping rather than the body. The two hammers, one with a uniform mass distribution (plastic head and wood handle) and the other with a mass-concentrated steel head, presented varied center of mass (COM) challenges. The robot had to grasp the head of the steel hammer due to its limited gripper wrench capacity.

We used Reality Composer on an iPhone 13 Pro to create the objects' meshes and sampled point clouds from these meshes for input into both GraspNet and our model. In normal scenarios, our model received simple object descriptions, while in the challenging situations, we provided with detailed language descriptions outlining our specific grasping requirements. Our experiments operated under the assumption of known accurate object pose, as pose estimation was not the focus of this study. We used PyBullet for motion planning and commanded a FANUC Robot LR Mate 200iD/7L to grasp object at the predicted grasp positions.

2) Results: In real-world tests assessing grasping success rates, our model consistently surpassed GraspNet across a range of objects and scenarios. Table IV presents a summary of these success rates. Our method achieved an impressive success rate of 70% in both normal and challenging scenarios, whereas GraspNet attained 50% in normal conditions and 20% in challenging ones. This highlights our method's efficacy and dependability in real-world grasp generation.

Figure 6 illustrates the resulting grasping poses. The successful grasping of bananas and hammers further exemplifies our model's ability to generalize to objects that are unseen in our dataset.

VI. Conclusion

This study delves into the integration of physical commonsense reasoning into robotic grasping. We introduce PhyGrasp, a large multimodal model that combines inputs from two modalities: natural language and 3D point clouds, seamlessly connected through a bridge module. The language modality demonstrates robust reasoning capabilities regarding the impacts of diverse physical attributes on grasping, while the 3D modality comprehends object shapes and parts. By leveraging these two capabilities, PhyGrasp accurately evaluates the physical properties of object parts and determines optimal positions and angles for grasping. Moreover, its language understanding enables it to interpret human instructions, facilitating the generation of grasping poses aligned with human preferences. To train PhyGrasp, we curate our PhyPartNet dataset comprising 195,000 object instances with varying physical properties, along with corresponding language descriptions of these properties and human preferences. We anticipate that our dataset and models will prove to be valuable resources for the community, particularly for those interested in advancing physical reasoning and grasping.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736, 2022.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446):eaat8414, 2019.

[5] Michel Breyer, Jen Jen Chung, Lionel Ott, Roland Siegwart, and Juan Nieto. Volumetric grasping network: Real-time 6 dof grasp detection in clutter. In *Conference on Robot Learning*, pages 1602–1611. PMLR, 2021.

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2:

Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[8] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

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

[10] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[11] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2022.

[12] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2019.

[13] Jinda Cui and Jeff Trinkle. Toward next-generation learned robot manipulation. *Science robotics*, 6(54): eabd9461, 2021.

[14] Yan Ding, Xiaohan Zhang, Chris Paxton, and Shiqi Zhang. Task and motion planning with large language models for object rearrangement. *arXiv preprint arXiv:2303.06247*, 2023.

[15] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.

[16] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020.

[17] Hao-Shu Fang, Minghao Gou, Chenxi Wang, and Cewu Lu. Robust grasping across diverse sensor qualities: The graspnet-1billion dataset. *The International Journal of Robotics Research*, 2023.

[18] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 2023.

[19] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2139–2147, 2018.

[20] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. *arXiv preprint arXiv:2309.02561*, 2023.

[21] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023.

[22] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022.

[23] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[24] Wenlong Huang, Fei Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Pete Florence, Igor Mordatch, Sergey Levine, Karol Hausman, et al. Grounded decoding: Guiding text generation with grounded models for robot control. *arXiv preprint arXiv:2303.00855*, 2023.

[25] Michael Janner, Sergey Levine, William T. Freeman, Joshua B. Tenenbaum, Chelsea Finn, and Jiajun Wu. Reasoning about physical interactions with object-oriented prediction and planning. In *International Conference on Learning Representations*, 2019.

[26] Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. *arXiv preprint arXiv:2104.01542*, 2021.

[27] Yuanchen Ju, Kaizhe Hu, Guowei Zhang, Gu Zhang, Mingrun Jiang, and Huazhe Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. *arXiv preprint arXiv:2401.07487*, 2024.

[28] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009.

[29] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[30] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

[31] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang,

Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[32] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[33] Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Qi Liu, Lingpeng Kong, and Xu Sun. Can language models understand physical concepts? *arXiv preprint arXiv:2305.14057*, 2023.

[34] Yunzhu Li, Toru Lin, Kexin Yi, Daniel Bear, Daniel L.K. Yamins, Jiajun Wu, Joshua B. Tenenbaum, and Antonio Torralba. Visual grounding of learned physical models. In *ICML*, 2020.

[35] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*, 2023.

[36] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as Policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

[37] Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023.

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

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

[40] Peiqi Liu, Yaswanth Orru, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Ok-robot: What really matters in integrating open-knowledge models for robotics. *arXiv preprint arXiv:2401.12202*, 2024.

[41] Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M. Dai. Mind's eye: Grounded language model reasoning through simulation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=4rXMRuoJlai.

[42] Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*, 2023.

[43] Yujie Lu, Pan Lu, Zhiyu Chen, Wanrong Zhu, Xin Eric Wang, and William Yang Wang. Multimodal procedural planning via dual text-image prompting. *arXiv preprint arXiv:2305.01795*, 2023.

[44] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *arXiv preprint arXiv:1703.09312*, 2017.

[45] Jeffrey Mahler, Matthew Matl, Vishal Satish, Michael Danielczuk, Bill DeRose, Stephen McKinley, and Ken Goldberg. Learning ambidextrous robot grasping policies. *Science Robotics*, 4(26):eaau4984, 2019.

[46] Matthew T Mason. Toward robotic manipulation. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:1–28, 2018.

[47] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019.

[48] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pretraining via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023.

[49] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[50] OpenAI. Gpt-4 technical report, 2023.

[51] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[52] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, 2023.

[53] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[54] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 35:23192–23204, 2022.

[55] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9, 2019.

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[57] Shreyas Sundara Raman, Vanya Cohen, Eric Rosen, Ifrah Idrees, David Paulius, and Stefanie Tellex. Planning with large language models via corrective re-prompting. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

[58] Máximo A Roa and Raúl Suárez. Grasp quality measures: review and performance. *Autonomous robots*, 38: 65–88, 2015.

[59] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. ProgPrompt: Generating situated robot task plans using large language models. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE, 2023.

[60] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023.

[61] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

[62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[63] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[64] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. Chatgpt for robotics: Design principles and model abilities. *Microsoft Auton. Syst. Robot. Res*, 2:20, 2023.

[65] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[66] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023.

[67] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *NeurIPS*, 2023.

[68] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. *Advances in neural information processing systems*, 28, 2015.

[69] Jiajun Wu, Joseph J Lim, Hongyi Zhang, Joshua B Tenenbaum, and William T Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, page 7, 2016.

[70] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

[71] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[72] Zhenjia Xu, Jiajun Wu, Andy Zeng, Joshua B Tenenbaum, and Shuran Song. Densephysnet: Learning dense physical object representations via multi-step dynamic interactions. In *Robotics: Science and Systems (RSS)*, 2019.

[73] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. Octopus: Embodied vision-language programmer from environmental feedback. *arXiv preprint arXiv:2310.08588*, 2023.

[74] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9, 2023.

[75] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-docowl: Modularized multimodal large language model for document understanding, 2023.

[76] Qiaojun Yu, Junbo Wang, Wenhai Liu, Ce Hao, Liu Liu, Lin Shao, Weiming Wang, and Cewu Lu. Gamma: Generalizable articulation modeling and manipulation for articulated objects. *arXiv preprint arXiv:2309.16264*, 2023.

[77] Haoqi Yuan, Chi Zhang, Hongcheng Wang, Feiyang Xie, Penglin Cai, Hao Dong, and Zongqing Lu. Plan4mc: Skill reinforcement learning and planning for open-world minecraft tasks. *arXiv preprint arXiv:2303.16563*, 2023.

[78] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

[79] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[80] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

[81] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and

Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.
[82] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.
[83] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
[84] Xinghao Zhu, Lingfeng Sun, Yongxiang Fan, and Masayoshi Tomizuka. 6-dof contrastive grasp proposal network. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6371–6377. IEEE, 2021.
[85] Xinghao Zhu, Yefan Zhou, Yongxiang Fan, Lingfeng Sun, Jianyu Chen, and Masayoshi Tomizuka. Learn to grasp with less supervision: A data-efficient maximum likelihood grasp sampling loss. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 721–727. IEEE, 2022.
[86] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023.
[87] Xueyan Zou, Linjie Li, Jianfeng Wang, Jianwei Yang, Mingyu Ding, Zhengyuan Yang, Feng Li, Hao Zhang, Shilong Liu, Arul Aravinthan, et al. Interfacing foundation models' embeddings. *arXiv preprint arXiv:2312.07532*, 2023.