TOWARDS NATURAL LANGUAGE-DRIVEN INDUSTRIAL ASSEMBLY USING FOUNDATION MODELS

Omkar Joglekar*, **Shir Kozlovsky***, **Vladimir Tchuiev***, **Zohar Feldman***
Bosch Center for Artificial Intelligence
Haifa, Israel
{joo1tv, kos1tv, tcv1tv, fez1tv}@bosch.com

Tal Lancewicki*†
Tel Aviv University
lancewicki@mail.tau.ac.il

Dotan Di Castro
Bosch Center for Artificial Intelligence
Haifa, Israel
did1tv@bosch.com

ABSTRACT

Large Language Models (LLMs) and strong vision models have enabled rapid research and development in the field of Vision-Language-Action models that enable robotic control. The main objective of these methods is to develop a generalist policy that can control robots with various embodiments. However, in industrial robotic applications such as automated assembly and disassembly, some tasks, such as insertion, demand greater accuracy and involve intricate factors like contact engagement, friction handling, and refined motor skills. Implementing these skills using a generalist policy is challenging because these policies might integrate further sensory data, including force or torque measurements, for enhanced precision. In our method, we present a global control policy based on LLMs that can transfer the control policy to a finite set of skills that are specifically trained to perform high-precision tasks through dynamic context switching. The integration of LLMs into this framework underscores their significance in not only interpreting and processing language inputs but also in enriching the control mechanisms for diverse and intricate robotic operations.

1 INTRODUCTION

The advent of Large Language Models (LLMs) and strong vision models, triggered the development of strong Vision-Language Models (VLMs). These strong Foundation Models (FMs) are powered by the strength and versatility of the transformer model (Vaswani et al., 2023). More recently, in the dynamic realm of industrial robotics, works based on these FMs, such as Octo, RT-X, PaLM-E (Ghosh et al.; Collaboration et al., 2023; Driess et al., 2023) showcases sophisticated technological prowess and heightened operational efficiency. This signifies a radical overhaul in robotics, where theoretical models are skillfully intertwined with practical robotic applications, thereby unlocking new, unprecedented capabilities.

Specific methods such as Octo (Ghosh et al.), RT-2 (Brohan et al., 2023), PaLM-E (Driess et al., 2023) and ChatGPT for Robotics (Vemprala et al., 2023) showcase ingenious methods to decode the complex learned tokens in LLMs such as T5 (Raffel et al., 2019), GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022), to generate control actions for robots with minimal fine-tuning. This expansion from the initial fortes of LLMs in language processing to the complexities of robotic contexts illustrates their immense potential. The versatility and robustness of LLMs are crucial in reshaping robot manipulation, navigation, and interaction, presenting groundbreaking solutions to tasks previously deemed unfeasible. The unique challenges these models encounter, such as data

*All authors contributed equally to this work.
†This work was done while the author was an intern at the Bosch Center for Artificial Intelligence (BCAI).

scarcity, task variability, and the demand for real-time responsiveness, are extensively explored by Wang et al. (2024) in their survey. This survey offers a thorough insight into the current landscape and future prospects of foundation models and LLMs in robotics.

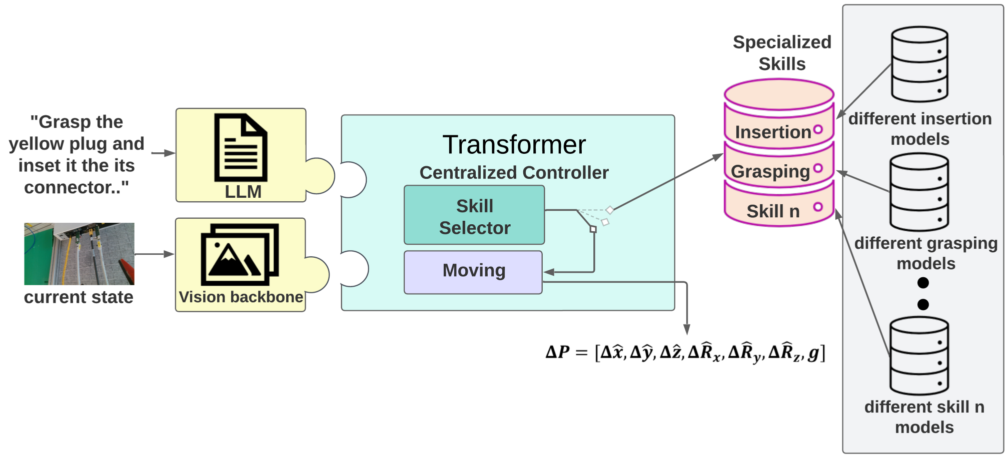


Figure 1: We introduced a user-friendly industrial assembly model based on natural language prompt and LLM, which is easily adaptable to new environments and modular in adding or adjusting components to meet specific needs. The modular nature is evident in the system's inputs and outputs. Regarding input flexibility, the model is designed to support various types of specialized skills' NN models

The implications of these advancements are especially significant in industrial assembly, where robots are expected to perform complex, high-precision tasks. Foundation models, with their enhanced decision-making, planning, and execution abilities, are pivotal for the intricate demands of industrial assembly. These models herald a shift towards smarter, more adaptive, and more efficient industrial processes, enabling robots to tackle complex and unstructured tasks that surpass the limitations of conventional automated systems. Our proposed method presents an approach that processes natural language-based goals from an assembly pipeline, such as "Carefully grasp the yellow plug and insert it into the appropriate socket." to output control instructions based on image observations. In principle, our solution comprises a centralized goal-conditioned model that initiates readily available independent specialized skills by handing over the control to specialized models based on context switching. The centralized module is responsible for inferring the skill to be executed and moving the robot to an initial pose, enabling the specialized control model (skill model) to perform precise manipulation. Specifically, the model outputs two control signals, namely:

1. 7 Degree of Freedom (7DoF) pose for the default "move" action.
2. Skill-based context class.

The summary of our presented method can be seen in Fig 1. The "skills" mentioned above are small pre-trained networks specializing in fine-grained control tasks such as insertion (Fu et al., 2022; Spector et al., 2022). We use the fine-grained skill "grasping for insertion", as an example of specialized skill, to test proper context switching. Our method is modular in terms of the text and image encoders and the type of policy used (we use a simple Multi-Layer Perceptron). In addition, our model is modular regarding specialized skills by simply fine-tuning the context classifier and calling the relevant policy to execute.

2 Related Work

2.1 Vision-Language-Action modeling

Multi-modal data processing has emerged as a new trend of research in the deep learning community. This is enabled by the versatility of the transformer architecture (Vaswani et al., 2023) to process tokens from various data sources. The versatility of transformers is highlighted in works such as ViLT (Kim et al., 2021) and Flamingo (Alayrac et al., 2022), where tokens from images and text are successfully processed by a large transformer model for various downstream applications. Until recently, Vision-Language Models (VLMs) were studied rigorously for various applications. Notably, works such as VisualBERT, ViLBERT, MetaLM (Li et al., 2019; Lu et al., 2019; Hao et al., 2022) and others, as outlined by Gan et al. (2022), are enabled by accelerated research in large language models (LLMs) such as Devlin et al.; Brown et al. (2020); OpenAI et al. (2023); Raffel et al. (2019) and vision models like Dosovitskiy et al. (2020). VLMs are applied in multiple domains like visual question answering Zhou et al. (2019); Zellers et al. (2021), captioning Hu et al., optical character recognition Li et al. (2021), and object detection Chen et al. (2021).

More recently, embodied VLMs, also known as Vision-Language-Action (VLA) models, have gained popularity. These models can perform complex tasks using visual observations conditioned on text prompts. Works such as those introduced by Ahn et al. (2022); Driess et al. (2023); Ha et al. (2023); Hu et al. (2023) augment VLMs by adding support for robotic actions. In parallel, developing expansive models for precise low-level control is an actively researched field, with significant contributions by Brohan et al. (2022; 2023). These models face challenges related to scalability and data requirements, as discussed byCollaboration et al. (2023) and Kalashnikov et al. (2021). These challenges highlight the need for a training framework that balances actionable outputs with scalability. Recent studies, such as those by Reed et al. (2022), Yang et al. (2023), and Kumar et al. (2023), have focused on models that map robot observations directly to actions, exhibiting impressive zero-shot or few-shot generalization capabilities across various domains and robotic systems. Notably, Bousmalis et al. (2023) have made strides in managing different robot embodiments for goal-conditioned tasks. The RT-X model by Collaboration et al. (2023) demonstrates proficiency in language-conditioned manipulation tasks across multiple robot embodiments, signaling a move towards more adaptable robotic systems. Very recently introduced by Ghosh et al., Octo enables VLAs for multiple action spaces by leveraging the versatility of the transformer architecture. This modularity allows the adaptation of robotic FMs to different robot types by swapping the encoders and decoders depending on the embodiment of the robot.

Our work specifically addresses the challenge of converting language and vision inputs into robotic actions, pushing the boundaries of Transformers in robotic control. This extends beyond text-based conditioning, as explored in studies by Gupta et al. (2022), Jang et al. (2022), and Jiang et al. (2022), demonstrating the models' effectiveness across various robotic designs and task specifications. Distinguishing our research, we have developed a centralized foundational model that combines multi-modality with pre-trained large models for specific sub-skills and tasks and offers flexibility in incorporating new skills or tasks. This model effectively connects to language commands and maintains robustness through feature modulation, marking a significant advancement in using Foundation Models in robotics.

2.2 LLM In Robotics

With recent success LLMs, most notably the GPT family by OpenAI (Brown et al., 2020; OpenAI et al., 2023), and the open-source Llama 2 model by Meta (Touvron et al., 2023), researchers aim to leverage the leap in reasoning capability for robotics applications. Examples include LM-Nav (Shah et al.), which tackles the problem of robotic navigation from language commands by utilizing CLIP for image and text interface, GPT-3 for text-based reasoning, and ViNG (Shah et al., 2021), a model for visual navigation. Similarly, VLMaps (Huang et al., 2022) utilizes LLMs to translate natural-language commands to a sequence of navigation goals for a robot. Liang et al. (2023) utilized an LLM fine-tuned on code data to create policies for robotic manipulation tasks via Python scripts that call task-based APIs. Song et al. proposed an LLM-based close-loop planning scheme for mobile robots for robust execution of multi-step tasks. ProgPrompt (Singh et al., 2023) tackles in detail prompting for LLMs for planning and robotic tasks with a programmatic approach.

In combination with VLA models, LLMs are used for visual-based reasoning and planning. VLA models such as RT-1 (Brohan et al., 2022)), RT-2 (Brohan et al., 2023), and PaLM-E (Driess et al., 2023) use tokens from pretrained LLMs as a major component in the VLA model and show impressive results. Other works, such as ChatGPT for Robotics (Vemprala et al., 2023), use GPT-3 (Brown et al., 2020) and clever prompting to generate actions for robots in simulated environments. Additionally, FMs, specifically LLMs, and VLMs, see active research for direct and actionable guidance, a topic that Huang et al. (2023) have explored. Still, modeling real-world physical dynamic systems remains an open challenge. Ghosh et al. use T5-based (Raffel et al., 2019) encoder to process language instructions.

One major consideration in this design is the ability to update the LLM used in the VLA model with ease and minimal fine-tuning. All the aforementioned VLA models introduce various techniques to address this issue. Our method also allows effective swapping of the LLM, which ensures the longevity of the method in light of drastic LLM improvements in the future. In our method, we consider this as an important design decision to ensure effective future-proofing.

3 METHODS

We introduce a centralized control model, employing the transformer architecture, that adeptly switches between specialized control skills from a predefined set, guided by natural language objectives and vision-based inputs.

This centralized controller fulfills two primary functions:

1. Direct the robot to a specified location based on the text prompts
2. Identify and predict the necessary specialized skill, such as grasping or insertion, based on the textual prompt and the robot's current state

The first function, which we term the general ”moving” skill, doesn't necessitate a highly precise 6 Degrees of Freedom (6DoF) pose estimation. The specialized tasks mentioned in the second function demand greater accuracy and involve intricate factors like contact engagement, friction handling, and refined motor skills. Additionally, they might integrate further sensory data, including force or torque measurements, for enhanced precision. Distinct from our core model, these specialized skills are developed independently, utilizing data specifically tailored to their requirements.

In this preliminary version of our work, we assume that these special skills work accurately, given that the robot meets certain constraints, such as being placed in an initial position that is in the proximity region of the manipulated object. The goal is be specified using a natural language prompt, for example, ”Carefully grasp the yellow plug and insert it into the appropriate socket.”

As outlined in Fig. 2, our transformer model accepts language instruction tokens that are encoded by strong language models such as T5 (Raffel et al., 2019), BLIP (Li et al., 2022) and CLIP (Radford et al., 2021), that are pre-trained, frozen, and specialize in text encoding, and generate text instruction tokens. In addition, we use pre-trained vision encoders, such as ResNet-50(He et al., 2015) or ViT (Dosovitskiy et al., 2020), to generate vision tokens that embed information from the observations. For a more detailed discussion on the effect of each chosen encoder model, please refer to Section 6. We pad the input with learnable ”readout tokens”, as described in Octo (Ghosh et al.). In our implementation, the transformer implements a Markovian policy, wherein the action depends solely on the current observation and is independent of past observations. In alignment with our dual-purpose model, we bifurcate the action into two categories: the skill action, denoted as a_s, which pertains to the type of skill being executed, and the moving action, denoted as a_m, which relates to the movement skill. The problem can be formally defined as follows:

$$\begin{aligned} a_s &= \pi_s(s), \\ a_m &= \pi_m(s), \end{aligned} \tag{1}$$

where s is the state vector that encodes information about the current state ($image(t)$) and the general text prompt.

As described in Fig. 2, both policies largely share their weights and architecture but differ in their decoder models. Both policies mentioned are deterministic and based on the Multi-Layer Perceptron(MLP) architecture. The policy π_s functions as a high-level controller, predicting the required skill by classifying predefined skills as follows:

0. Terminate
1. Moving (handled by the centralized controller)

2. Skill 1 (specialized)
3. Skill 2 (specialized)
4. Skill 3 (specialized)

$\vdots$

"Terminate" indicates that the robot has reached its goal per the provided text prompt. When $a_s =$ skill n, the control is handed over to the model specialized in skill n. When a_s predicts the "moving" skill (denoted as "1"), the low-level controller's (π_m) action is executed. Additional specialized skills can be integrated by adding another context class and fine-tuning the classification head with data pertinent to the new skill.

The action space of a_m is defined as a 7-dimensional vector, trained to predict a unit vector in the direction of the delta ground truth of the desired object or task using MSE loss. It is formulated as:

$$\Delta \mathbf{p} = [\Delta x, \Delta y, \Delta z, \Delta R_x, \Delta R_y, \Delta R_z, g] \tag{2}$$

In this formulation, $\Delta x, \Delta y, \Delta z$ represent the translation components, while $\Delta r_x, \Delta r_y, \Delta r_z$ denote the orientation components represented in axis-angles, and g corresponds to the opening of the gripper. This 7-dimensional vector is trained in a supervised manner using the Mean Squared Error (MSE) ($\mathbf{L}_{mse}$).

We define *active domain* as the region that enables the successful execution of a specialized skill. The boundary of this active domain is assumed to be an abstract threshold ε. This threshold varies for different skills and is not solely dependent on distance. For instance, when guiding a grasped plug to a socket for insertion, the context should revert to the "grasping for insertion" specialized skill if the plug's position becomes unfavorable for insertion. A classifier head is trained to estimate this abstract threshold and facilitate context switching accordingly. This multi-class classifier head is trained using Categorical Cross Entropy loss ($\mathbf{L}_{ce}$).

4 Experimental setup

We evaluate the performance of our method using three methods:

1. The quality of the move action - to evaluate the performance of π_m.
2. The accuracy of context switching - to evaluate how accurately the high-level controller transfers from one skill to another
3. The overall system performance - This preliminary version of our work focuses on evaluating the performance of the full systems only for grasping tasks. It measures the system's ability to accurately identify and approach the right plug according to the text prompt, engage our independently trained grasping network, and grasp the plug correctly. We only present qualitative results for this experiment, and the presentation of this specialized skill will be included in future works.

The first method evaluates the quality of the moving action. Specifically, we measure whether the high-level control can bring the gripper within the active domain of the "grasping for insertion" skill. The second method evaluates whether the model accurately switches context when it is inside the active domain to perform "grasping for insertion". For this particular skill, the constraints can be summarized by Eq 3 and 4

$$||[\Delta x, \Delta y, \Delta z]|| \leq 4[cm] \tag{3}$$

$$||[\Delta r_x, \Delta r_y, \Delta r_z]|| \leq 20[deg] \tag{4}$$

To evaluate the robustness of our model, we introduce several perturbations to our experimental setup:

- Plug Distractions: In addition to the four plugs initially trained on, we introduce two additional plugs of a different type into the scene to act as distractions.
- Object Distractions: We include large objects not present in the training scenes.

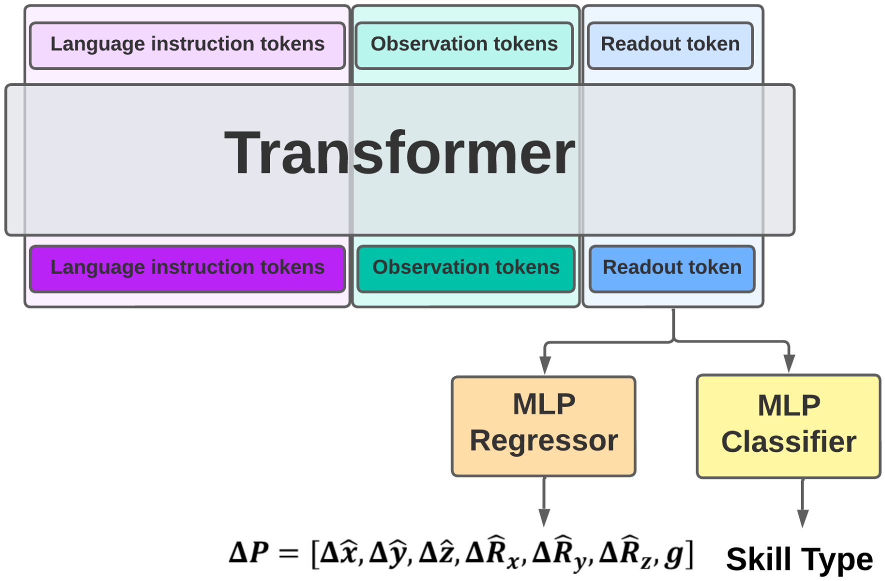


Figure 2: The centralized controller architecture: In each time step, Language tokens, Observation tokens, and the Readout token are concatenated together, where a mask is applied to ensure that the readout token attends all other tokens but is unattended by them. The MLP classifier then determines the appropriate skill to execute, and if the moving skill is selected, a small step is taken in accordance with the MLP Regressor's output.

- Missing Plug: We remove one of the four plugs to examine if its absence influences the success rate for the remaining plugs.
- Unseen Background: We modify the background to an unseen variant.

Example scenarios for each perturbation can be seen in Fig. 3

5 Results

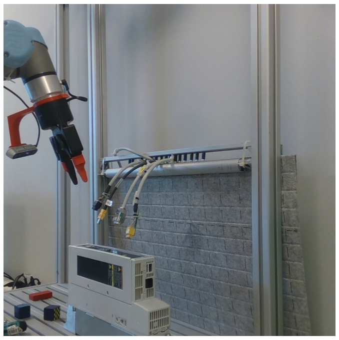
5.1 Quantitative Results

Table 1 shows how many times the moving policy π_m successfully brings the gripper into the active domain (as described in Section 4) for the skill. We observe that the model performs well on the plugs within the model's training set.

Perturbation	**Success Rate**
Baseline (No Perturbation)	$33/40(82.5\%)$
Plug Distractions	$9/19(47\%)$
Object Distractions	$20/23(87\%)$
Missing Plug	$26/36(72\%)$
Unseen Background	$12/17(70.6\%)$

Table 1: Success Rates of Central Controller Under Various Environment Perturbations

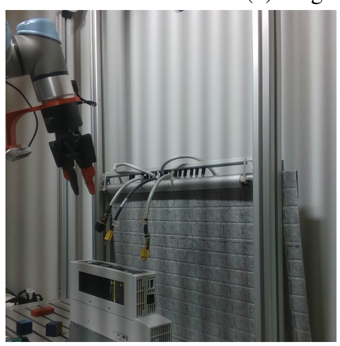
The context switching is modeled as a multi-class classification problem (Section 3). The Fig. 4c and Fig. 4d show the accuracy of this multi-class classification. We see that the final context switching accuracy is $\approx 95\%$ in validation.



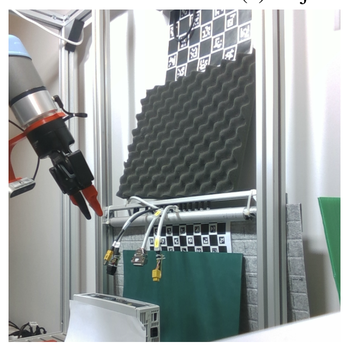
(a) Baseline scene

(b) Plug Distractions

(c) Object Distractions

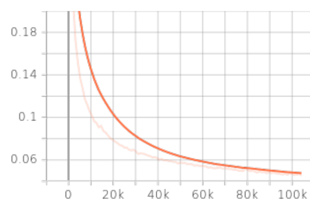


(d) Missing Plug

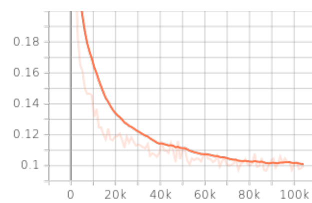


(e) Unseen Background

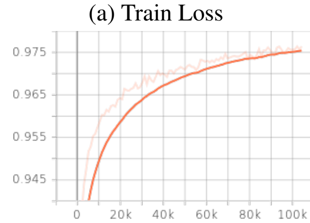
Figure 3: Examples of perturbations introduced in the experimental setup to evaluate the robustness of our model. Each sub-figure illustrates a different type of perturbation.



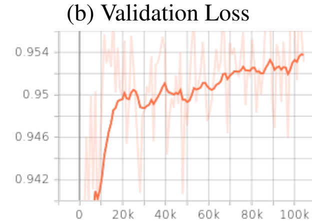
(a) Train Loss



(b) Validation Loss



(c) Context Switching accuracy (Train)



(d) Context Switching accuracy (Validation)

Figure 4: Convergence Graphs and Context Switching Accuracy

5.2 QUALITATIVE RESULTS

In the additional materials, we upload videos of the prediction results of the trained model. We present demo videos of reaching and the ”grasping for insertion” skill. All videos can be found in the project website: `https://sites.google.com/view/roboticfm`. The detailed description of the ”grasping for insertion” skill is outside the scope of this work and will be presented in future methods.

6 ABLATION STUDY

The major contribution of our work is a unified transformer model that fuses language and image data to generate control signals for robotic operation. We performed experiments to choose the most effective transformer architecture. Recently as a part of the Segment Anything Model (SAM) Kirillov et al. (2023), the authors introduced a new attention mechanism termed as ”two-way attention”, in their mask decoder. This attention mechanism differs from regular cross-attention mechanisms in that it updates the context and inputs in each layer. We refer curious readers to this paper for more information about this technique. Motivated by their success in performing complex segmentation tasks, we also test a transformer network that uses a stack of two-way attention blocks described in SAM, along with the standard transformer-decoder network that uses a stack of self-attention blocks (Vaswani et al. (2023)).

While using the two-way attention layers, we input the image embeddings as the ”dense tokens” and the readout and language tokens as ”sparse tokens”. The presence of two cross-attention layers in each such layer drastically increases the memory requirement and processing time. We set the depth of the transformers such that both mechanisms have a similar number of parameters ($\sim 20M$). We find that using a two-way transformer of the same size does not generalize well to points outside the dataset and drastically reduces success rates.

The inherent modularity presented in our model opens up a discussion about the choice of image and text encoders. Broadly speaking, we test two types of text encoders:

1. Encoders pre-trained on Vision-Language tasks such as CLIP or BLIP text encoders
2. Encoders pre-trained Language generation such as the suite of T5 text encoders

Both the T5-small and CLIP text encoders show promising performance. However, CLIP outperforms the T5-small model due to its experience with pairing images to text. The results presented in 1 are for the model that uses CLIP. The size of the text encoder affects the stage in the pipeline where the data fusion begins. In other words, if the text encoder is very deep, then the image-language fusion begins much later on in the pipeline. This effect becomes prominent when swapping the CLIP text encoder for the BLIP one, which is much deeper. Specifically, the model’s performance drops when we use the BLIP text encoder. All the text encoder tests were done while keeping the same image encoder (ResNet-18 pre-trained on ImageNet).

To determine the best image encoder, we fix the CLIP text encoder. We try using ResNet-50 pre-trained on ImageNet and ResNet-50 pre-trained as a part of CLIP. Unsurprisingly, the latter image encoder shows better performance, generalization, and robustness. Table 1 presents results using this type of image encoder.

7 CONCLUSIONS AND FUTURE WORK

This work presents a Robotics Foundation Model that uses text-based task specification and image-based observations to generate control actions for robotic assembly tasks. The highlighting features of our model include modularity, ease of use, and the ability to add any finite number of specialized skills, such as insertion, to the *skill library* that the centralized model can control. We present a comprehensive architecture, training protocol, and qualitative & quantitative results of the described method.

We wish to emphasize that this project is still in the research phase. In future works, we aim to introduce a bigger database of compatible skills and support for task specification using goal images. Success rate improvement is also an active research topic, and implementing a full pipeline for more complex task specifications to deal with collisions and occlusions is also under exploration. We also plan to add datasets from various robotic setups and assembly lines to improve the generalization ability for additional assembly tasks.

REFERENCES

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 J 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, Not As I Say: Grounding Language in Robotic Affordances. 4 2022. URL http://arxiv.org/abs/2204.01691.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. 4 2022. URL http://arxiv.org/abs/2204.14198.

Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X. Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, Antoine Laurens, Claudio Fantacci, Valentin Dalibard, Martina Zambelli, Murilo Martins, Rugile Pevceviciute, Michiel Blokzijl, Misha Denil, Nathan Batchelor, Thomas Lampe, Emilio Parisotto, Konrad Żołna, Scott Reed, Sergio Gómez Colmenarejo, Jon Scholz, Abbas Abdolmaleki, Oliver Groth, Jean-Baptiste Regli, Oleg Sushkov, Tom Rothörl, José Enrique Chen, Yusuf Aytar, Dave Barker, Joy Ortiz, Martin Riedmiller, Jost Tobias Springenberg, Raia Hadsell, Francesco Nori, and Nicolas Heess. RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation. 6 2023. URL http://arxiv.org/abs/2306.11706.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics Transformer for Real-World Control at Scale. 12 2022. URL http://arxiv.org/abs/2212.06817.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. 7 2023. URL http://arxiv.org/abs/2307.15818.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. 5 2020. URL http://arxiv.org/abs/2005.14165.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A Language Modeling Framework for Object Detection. 9 2021. URL http://arxiv.org/abs/2109.10852.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling Language Modeling with Pathways. 4 2022. URL http://arxiv.org/abs/2204.02311.

Open X-Embodiment Collaboration, Abhishek Padalkar, Acorn Pooley, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Animesh Garg, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Gregory Kahn, Hao Su, Hao-Shu Fang, Haochen Shi, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jan Peters, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jiajun Wu, Jialin Wu, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jitendra Malik, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Joseph J. Lim, João Silvério, Junhyek Han, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Zhang, Krishan Rana, Krishnan Srinivasan, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Lionel Ott, Lisa Lee, Masayoshi Tomizuka, Max Spero, Maximilian Du, Michael Ahn, Mingtong Zhang, Mingyu Ding, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Pannag R Sanketi, Paul Wohlhart, Peng Xu, Pierre Sermanet, Priya Sundaresan, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shuran Song, Sichun Xu, Siddhant Haldar, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Sudeep Dasari, Suneel Belkhale, Takayuki Osa, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xuanlin Li, Yao Lu, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zhuo Xu, and Zichen Jeff Cui. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. 10 2023. URL http://arxiv.org/abs/2310.08864.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Technical report. URL https://github.com/tensorflow/tensor2tensor.

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. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 10 2020. URL http://arxiv.org/abs/2010.11929.

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. 3 2023. URL http://arxiv.org/abs/2303.03378.

Letian Fu, Huang Huang, Lars Berscheid, Hui Li, Ken Goldberg, and Sachin Chitta. Safe Self-Supervised Learning in Real of Visuo-Tactile Feedback Policies for Industrial Insertion. 10 2022. URL http://arxiv.org/abs/2210.01340.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-Language Pre-training: Basics, Recent Advances, and Future Trends. 10 2022. URL http://arxiv.org/abs/2210.09263.

Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Dorsa Sadigh, Chelsea Finn, Sergey Levine, and Uc Berkeley. Octo: An Open-Source Generalist Robot Policy Octo Model Team Out-of-the-box Multi-Robot Control 800k Robot Trajectories. Technical report. URL https://octo-models.github.io.

Agrim Gupta, Linxi Fan, Surya Ganguli, and Li Fei-Fei. MetaMorph: Learning Universal Controllers with Transformers. 3 2022. URL http://arxiv.org/abs/2203.11931.

Huy Ha, Pete Florence, and Shuran Song. Scaling Up and Distilling Down: Language-Guided Robot Skill Acquisition. 7 2023. URL http://arxiv.org/abs/2307.14535.

Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language Models are General-Purpose Interfaces. 6 2022. URL http://arxiv.org/abs/2206.06336.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. 12 2015. URL http://arxiv.org/abs/1512.03385.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang Microsoft. Scaling Up Vision-Language Pre-training for Image Captioning. Technical report. URL https://github.com/explosion/spaCy.

Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look Before You Leap: Unveiling the Power of GPT-4V in Robotic Vision-Language Planning. 11 2023. URL http://arxiv.org/abs/2311.17842.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual Language Maps for Robot Navigation. 10 2022. URL http://arxiv.org/abs/2210.05714.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. 7 2023. URL http://arxiv.org/abs/2307.05973.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning. 2 2022. URL http://arxiv.org/abs/2202.02005.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: General Robot Manipulation with Multimodal Prompts. 10 2022. URL http://arxiv.org/abs/2210.03094.

Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. MT-Opt: Continuous Multi-Task Robotic Reinforcement Learning at Scale. 4 2021. URL http://arxiv.org/abs/2104.08212.

Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. 2 2021. URL http://arxiv.org/abs/2102.03334.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. 4 2023. URL http://arxiv.org/abs/2304.02643.

Vikash Kumar, Rutav Shah, Gaoyue Zhou, Vincent Moens, Vittorio Caggiano, Jay Vakil, Abhishek Gupta, and Aravind Rajeswaran. RoboHive: A Unified Framework for Robot Learning. 10 2023. URL http://arxiv.org/abs/2310.06828.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. 1 2022. URL http://arxiv.org/abs/2201.12086.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A Simple and Performant Baseline for Vision and Language. 8 2019. URL http://arxiv.org/abs/1908.03557.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. 9 2021. URL http://arxiv.org/abs/2109.10282.

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 *Proceedings - IEEE International Conference on Robotics and Automation*, volume 2023-May, pp. 9493–9500. Institute of Electrical and Electronics Engineers Inc., 2023. ISBN 9798350323658. doi: 10.1109/ICRA48891.2023.10160591.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. 8 2019. URL http://arxiv.org/abs/1908.02265.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya

Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report. 3 2023. URL `http://arxiv.org/abs/2303.08774`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya 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. 2 2021. URL `http://arxiv.org/abs/2103.00020`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. 10 2019. URL `http://arxiv.org/abs/1910.10683`.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A Generalist Agent. 5 2022. URL `http://arxiv.org/abs/2205.06175`.

Dhruv Shah, Bła˙ Zej Osí Nski †!, Brian Ichter, and Sergey Levine. LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action. Technical report.

Dhruv Shah, Benjamin Eysenbach, Gregory Kahn, Nicholas Rhinehart, and Sergey Levine. Ving: Learning Open-World Navigation with Visual Goals. In *Proceedings - IEEE International Conference on Robotics and Automation*, volume 2021-May, pp. 13215–13222. Institute of Electrical and Electronics Engineers Inc., 2021. ISBN 9781728190778. doi: 10.1109/ICRA48506.2021.9561936.

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 *Proceedings - IEEE International Conference on Robotics and Automation*, volume 2023-May, pp. 11523–11530. Institute of Electrical and Electronics Engineers Inc., 2023. ISBN 9798350323658. doi: 10.1109/ICRA48891.2023.10161317.

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. Technical report. URL `https://osu-nlp-group.github.io/LLM-Planner/`.

Oren Spector, Vladimir Tchuiev, and Dotan Di Castro. InsertionNet 2.0: Minimal Contact Multi-Step Insertion Using Multimodal Multiview Sensory Input. In *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 6330–6336. Institute of Electrical and Electronics Engineers Inc., 2022. ISBN 9781728196817. doi: 10.1109/ICRA46639.2022.9811798.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. 7 2023. URL `http://arxiv.org/abs/2307.09288`.

Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. Technical report, 2023.

Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. ChatGPT for Robotics: Design Principles and Model Abilities. 2 2023. URL http://arxiv.org/abs/2306.17582.

Jiaqi Wang, Zihao Wu, Yiwei Li, Hanqi Jiang, Peng Shu, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Huaqin Zhao, Zhengliang Liu, Haixing Dai, Lin Zhao, Bao Ge, Xiang Li, Tianming Liu, and Shu Zhang. Large Language Models for Robotics: Opportunities, Challenges, and Perspectives. 1 2024. URL http://arxiv.org/abs/2401.04334.

Jonathan Yang, Dorsa Sadigh, and Chelsea Finn. Polybot: Training One Policy Across Robots While Embracing Variability. 7 2023. URL http://arxiv.org/abs/2307.03719.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal Neural Script Knowledge Models. 6 2021. URL http://arxiv.org/abs/2106.02636.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified Vision-Language Pre-Training for Image Captioning and VQA. 9 2019. URL http://arxiv.org/abs/1909.11059.