Joint Action Language Modelling for Transparent Policy Execution

Theodor Wulff, Rahul Singh Maharjan, Xinyun Chi and Angelo Cangelosi Department of Computer Science The University of Manchester Manchester, United Kingdom {theodor.wulff,rahulsingh.maharjan,xinyun.chi,angelo.cangelosi}@manchester.ac.uk

Abstract—An agent's intention often remains hidden behind the black-box nature of embodied policies. Communication using natural language statements that describe the next action can provide transparency towards the agent's behaviour. We aim to insert transparent behaviour directly into the learning process, by transforming the problem of policy learning into a language generation problem and combining it with traditional autoregressive modelling. The resulting model produces transparent natural language statements followed by tokens representing the specific actions to solve long-horizon tasks in the Language-Table environment. Following previous work, the model is able to learn to produce a policy represented by special discretized tokens in an autoregressive manner. We place special emphasis on investigating the relationship between predicting actions and producing high-quality language for a transparent agent. We find that in many cases both the quality of the action trajectory and the transparent statement increase when they are generated simultaneously.

Index Terms—Behaviour Transparency, Vision Language Action Models, Robotics

I. INTRODUCTION

The field of robotics is progressing towards robots with higher degrees of autonomy [1]. Eventually, robots could be able to collaborate with humans after receiving only very little instruction on how to complete a certain task. However, as a robotic agent's degree of autonomy increases, its actions tend to remain opaque until the moment they are executed [2]. This is especially important for robots deployed in reallife scenarios where the interacting human might not have experience with that agent. In cases where the robot's behavior does not align with the expectations of the human, this can lead to a loss of trust in the robotic agent and, as a consequence, hinder the effectiveness of collaboration [3]. To avoid these situations and establish a common ground, humans naturally utilize language, among other behaviours, to coordinate tasks

The authors gratefully acknowledge funding from the EU and UKRI in the context of Horizon Europe under the MSCA grant agreement No 101072488 (TRAIL). Special thanks also to Yilun Du for publishing the additional short-term language annotations of the Language-Table dataset, and the team at the Computational Shared Facility at the University of Manchester for providing the resources to train our models.

© 2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

and solve problems effectively [4]. We hypothesize that autonomous robotic agents should exhibit similar behavior to collaborate in human-robot teams effectively.

Current robotics research focuses primarily on the training of agent behavior by evaluating policy execution and success rates within specific environments [1], [5]–[7]. Learning to provide transparency at the same time is rarely a consideration for the development of most agentic systems. Behavior transparency can be achieved by different means, for example, by providing an outline of the action trajectory before execution or by generating a detailed plan of subsequent actions [8]. In this work, we chose the domain that is the most commonly used for communication between humans and robots for transparency: natural language [8].

The success of Foundation Models, which have been trained on a huge amount of data, especially in the domains of language and vision, has led to researchers in robotics adopting such models in their systems [1], [5]–[7], [9]. However, they are mostly used either to learn a policy directly, which is no longer transparent, or to map an observation to a lower-level statement, which is then executed by a separately learned policy. The capabilities of producing language for communication are then considered separately, if at all. We argue that for a robotic agent to exhibit transparent behavior, learning to be transparent should be incorporated into the learning process from the beginning.

In this work, our aim is to increase the transparency of robot behavior by utilizing a single model to generate the next action and simultaneously provide a natural language statement. Modern Vision-Language Models (VLMs) form a base for effective grounded communication, as they have shown excellent qualities in image understanding and question-answering tasks [10]. We leverage the general language understanding and generation capabilities of the VLM to a) communicate the agent's subsequent action in natural language, and b) utilize specific action tokens to execute the agent policy by turning the problem into a full language learning task. This also allows us to specify additional contextual information like the robot's state into the query by mapping it to specific tokens or providing a corresponding language utterance.

Our key contributions can be summarized as:

• Joint Action Language Generation formulated as a transparent language-learning problem. We transform

the problem of learning the agent's policy to steer the actuators into a natural language processing task which inherently produces transparent statements within the same output. Contrary to prior work [5], [6], [9] that uses VLMs for policy learning, we specifically investigate the interplay between the production of language statements and low-level actions.

• Actions benefit from transparency. Our results show that autoregressively generating a transparent language statement alongside action tokens positively impacts the predicted trajectories as well as the language output.

The remainder of this paper is organized as follows: Section II covers current work on Vision-Language-Action Models and transparency in robotics. Section III describes the details of our approach. In Sections IV and V we present our experimental setup and results. The paper concludes with a discussion and brief summary in Sections VI and VII.

II. RELATED WORK

A. Vision-Language-Action Models

Based on their high generalizability, recent works have started to utilize VLMs in the field of robotics. Applied as an interface for training language-conditioned vision-based policies, they are termed Vision-Language-Action Models (VLA). This is accomplished by either defining a textual representation of the actions and treating the problem as a conditional language generation task or using specific policy heads that produce the action.

Driess et al. [11] embed robot state, pixel- and object-level visual input, and language in a multimodal token sequence, which is processed by a Large Language Model (LLM). A separate control policy performs the actions based on the language guidance provided by the model. They train their PaLM-E model [11] on various multimodal tasks to increase the quality of extracted features. Gosh et al. [12] propose to train a transformer-based generalist agent, called Octo, by training on a wide variety of mixed-modality data. Li et al. [13] predict coordinates of gripper-related objects in the field of view using specific queries to the VLM. Shridhar et al. [14] utilize 3D Voxels as input and goal specification to the Perceiver-Actor model. Kim et al. [6] propose OpenVLA, an open-source VLA that generates action tokens using an LLM backbone, which processes language and visual features. A specific decoder de-tokenizes the output tokens into low-level actions. The model is trained on a multitude of embodiments for generalizability across domains [6]. Black et al. [9] utilize a pretrained PaliGemma Vision-Language Model [10] and attach an action expert to it for policy execution.

The RT-1 [1] model embeds language and visual input into a joint token representation and produces discrete actions using a transformer head. Its successor RT-2 [5] utilizes a central LLM, which processes visual and language input to predict action tokens and was trained across Visual Question Answering among other tasks. RT-H [7] employs a twostep querying strategy by first letting the model break down the current task into a short-term action which serves as the context for predicting the next action using the same model. Similarly, Zhao et al. [15] break down an abstract task description into concretely executable actions whose execution is learned separately.

Although many models have been pretrained on languagegeneration tasks, little to no emphasis is put on utilizing the language-generative capabilities of the model to simultaneously make the actions more transparent. The aim is to create an improved language-conditioned policy, in contrast to a policy with explicitly high-quality language output. We aim to move research a step forward towards agents that have inherently learned to behave transparently.

B. Explainability and Transparency

Explainability and transparency in robotics are related topics, but differ slightly in their goals. Explainability aims to answer the *why*, *what and how* of robot behavior while transparency is mostly concerned with the *what and how*, which facilitates inferring the *why* without explicitly providing it [8]. Although we focus on transparency in this work, techniques that train a model to provide an explanation contain aspects of transparency as well. Both concepts have in common that they aim to provide answers on the *what* and *how* with respect to the agent's behavior.

Work in interactive explanation learning [16] aims to improve model-generated explanations with a human-in-the-loop who acts as a critic of the model's explanations, similar to how LLMs are trained with Reinforcement Learning from Human Feedback (RLHF) [17]. Stammer et al. [18] automate the learning process by replacing the human with a surrogate model that acts as the critic in the human's place. Alternatively, Duan et al. [19] train their VLM to provide natural language explanations on robot failures, after specifically pretraining on such cases.

Leveraging existing LLM or VLMs, Chain-of-thought methods can provide transparency by outlining multiple steps that lead to some solution. However, since these are generated, they can be erroneous themselves. The basic idea is to prompt an LLM or VLM to think step-by-step instead of solely producing an answer to a question and was proposed by Wei et al. [20]. As such, many variations of this process have been proposed that incorporate different modalities into the intermediate stepby-step thought chain [21], [22] or explore different possible paths along consecutive thought chains [23].

Other methods like Kerzel et al. [24] equip a robot with transparent behavior using a variety of modalities that highlight different decisions that take place in the system. Besides the progress of transparency in machine learning and robotic applications, transparency is an active topic of research in the fields of psychology and human-robot interaction [8]. Other research investigates the use of different modalities to provide transparency on robot behavior, even though language and speech are among the most prominent choices [25].

III. METHOD

Our training procedure consists of two main steps. First, we pretrain our models on visual question answering to generate transparent statements in robotic settings. Then, we train on robotic tasks to generate actions in addition to the transparent statements. Figure 1 visually summarizes our approach.

A. Problem Formulation

To facilitate the learning process and to show that the model can learn both the transparent statement in conjunction with the policy, we train the model to imitate an expert's behavior and augment its actions with the natural language statement given by the ground truth caption. Similar to Jang et al. [26], we train our models in a supervised way using languageconditioned behavior cloning. We extend the problem of language-conditioned behavior cloning [27] with the additional requirement of producing transparent statements. This problem is often modelled as a partially observable Markov decision process [28]. Conventially, the goal is to learn to predict the next action that follows an expert's policy given a language instruction and some observation of the environment. We extend this formulation by providing a long-term language instruction for the input and expanding the output to generate a transparent statement in the form of a short-term natural language description alongside the action.

TABLE I PROMPT DEFINITIONS

Prompt	Target				
Next action?	Action tokens				
Immediate next step and action?	Description and action tokens				
Immediate next step?	Description				
Context Definitions					
Current task is: <instruction>. <prompt></prompt></instruction>					
Civian (State) Current tool is (Instruction) (Durant)					

Given *<State>*. Current task is: *<Instruction>*. *<Prompt>*

In our case, the observations consist of a camera input from the robot, the current end-effector state, and the language query that contains the long-term goal instruction and the corresponding question. To embed the action and state vectors into the language prompts, we discretize each dimension of the continuous action/state vector representations by applying a binning strategy and mapping each associated bin onto corresponding special tokens. Refer to Section III-E for the detailed tokenization process.

Prior works have not explicitly investigated the quality of statements which have been produced as intermediate output. While the positive effects of intermediate outputs have been seen, e.g., in Chain-of-Thought-like mechanisms [5], [7], [20], the quality of these explanations is not clear. Here, we specifically investigate the output quality concerning ground truth statements.

B. Datasets

1) RoboVQA: The RoboVQA [29] dataset is a recent VQA dataset specific to robotic settings. It contains video scenes

of different agents operating in robotic scenarios annotated with question-answering pairs. The questions are asking for different forms of planning, success recognition, and scene understanding. We utilize the freeform planning questions to pretrain our models. Here, the goal is to predict a single natural language statement describing the "immediate next step". In total, we utilize 94,997 freeform planning question-answerimage triplets to pretrain our models.

2) Language Table: We split the Language-Table [30] dataset into training, validation and test subsets on an episode level. An episode consists of a long-horizon goal provided in natural language. Furthermore, each episode is split into multiple sub-episodes which are annotated with captions that describe the current high-level action. A sub-episode ends once the caption changes. We utilize the captions as our ground truth labels for the agent's transparent statements. The action trajectory consists of a sequence of low-level actions, in this case, 2D vectors referring to the translation of the robot arm's pointer across the board. This leads to a dataset with 23,019 episodes and 399.846 captions. We load a batch of episodes and sample N observations for each caption to form our minibatch before shuffling and passing it into the network. To sample the individual frames, we always select the first and last frames and uniformly draw N-2 frames from the sequence in between these.

C. Pretraining

We pretrain our models on the freeform planning subset of the RoboVQA dataset, which asks the model to predict the subsequent step given the current frame and context. We hypothesize that this extra fine-tuning step is beneficial since robotic data is sparse and rarely encountered in general webscraped text-image datasets used to pretrain VLMs [31]. The answer to the question is a transparent statement: a natural language description of the immediate next action. Queries are of the form: "Current goal is: *<goal description>*, immediate next step?" In some later configurations, we prepend the state information to the query following the tokenization strategy described in Section III-E. Refer to Table I for an exhaustive overview of prompts used in our setup.

D. Model

Taking inspiration from Black et al. [9], we use the PaliGemma model [10] for our experiments, due to its relatively small size with ca. 3 billion parameters, which reduces training and inference times, and the effective generalizability across different multimodal tasks, which require language grounding in visual inputs. We stick to an input resolution of 224x224 pixels for the images, since higher resolutions increase computational requirements, and previous work has not found a worthwhile performance increase. While we chose the PaliGemma model for the reasons above, our method can be applied to any other Vision-Language Model.

E. Action and State Tokenization

We limit the range of possible action trajectories to (-0.05, 0.05) along each dimension and state trajectories to (-0.3, 0.35)



Fig. 1. Method Overview. We utilize the Vision-Language Model PaliGemma to produce a transparent statement and action tokens given an input prompt, describing the current task, and the visual observation of the environment. The model is pretrained on visual question answering in robotic settings. We discretize the state and action vectors into special tokens to embed these directly into the input prompt and target strings.

along the x-axis and (0.2, 0.6) along the y-axis. These ranges ensure that we capture at least one standard deviation of each trajectory dimension based on the distribution of our training data. We discretize the two-dimensional action and state space by mapping each dimension of the trajectory and robot state to special tokens, which we add to the vocabulary of our model and tokenizer. We associate each of the action and state bins with a number to create the special tokens "a0" - "aN" and "s0" - "sN" for the tokenized actions and states. Additionally, we surround the tokenized state and actions with the special tokens "[state]" or "[action]" to mark the beginning and end of the state and actions. We perform the same procedure on the robot state as in RT-2 [5] and OpenVLA [6].

IV. EXPERIMENTS

We train our models on an Nvidia A100 GPU using the Adam optimizer. The training duration of our models is three epochs, each containing 160,000 examples. The minibatches contain 8 samples, but we accumulate the gradients between minibatches to reach an effective batch size of 256. We sample 3 frames from each sub-episode, following the procedure outlined in Section III-B, to ensure that our model sees a high variety of captions during training. To calculate the MSE and Cosine Similarity between the generated and target trajectories, we map the generated action tokens back to their respective discrete bins and compare the mean value of the bin with the continuous trajectory. Additionally, we investigate the ability to produce coherent language output using common metrics found in natural language processing: namely the

BLEU [32] and ROGUE-1 [33] scores. All measurements presented in Section V are presented with the corresponding standard deviation across the test set.

V. RESULTS

We provide the results on the language table dataset using supervised imitation learning. We investigate both the model's ability to reproduce the expert's action and the quality of the language statements. Previous work has already shown that VLMs can generate actions for execution on robotic systems [1], [5]–[7]. Of special interest is the effect that producing the verbal statement alongside a trajectory has on the quality of the trajectories.

TABLE II Action Before Language

Actions First	ROUGE↑	BLEU↑	CosSim↑	MSE↓
×	0.5087	0.1511	0.0106	0.0024
1	0.5247	0.1334	0.0377	0.0024

A. How well does the model produce transparent statements?

Based on the output of our test set, the model learns to generate comprehensive statements in natural language alongside a trajectory. Figure 3 presents a collection of sample input-output pairs. We include three positive and three negative samples. A sample is considered positive when the meaning of the generated transparent statement semantically resembles the ground truth statement or hints towards a similar action.



Fig. 2. Comparison between different orders of joint output: action tokens before or after the language statement.

The results further highlight the difficulty of evaluating learned transparent behavior. Even though the statement provided by the agent seems like a logical step towards approaching the provided goal, it can differ drastically from the ground truth wording. Sample 2 presents a good example of this, where the generated verb move is semantically similar to the target place. Word and n-gram-based metrics, like BLEU and ROUGE, do not reflect this behavior, even though they can give an intuition, and can lead to lower scores. In addition, the target direction of the pushing action is generated as towards center while the ground truth refers to diagonal to the green star. Without supplementing visual input, it is also difficult to determine whether these two utterances refer to the same location. When it comes to the negative samples, it can be challenging to determine whether the language output does present a viable option for achieving the goal despite the output not matching the ground truth. Regarding the actions, the predicted action tokens rarely exactly equal the ground truth tokens. However, this does not mean that the generated trajectories are of low quality. We refer to our other analyses which investigate the quality of trajectories after detokenizing the action tokens again.

B. Action Before Language

We find that producing the action tokens first and the statement last leads to more accurate action tokens than vice versa. The results in Table II show a slightly lower BLEU score but a higher ROUGE score when producing the language statement last. The increased ROUGE score means that this model generates more words that are part of the target statement than in the setting that produces the statement last. However, it also generates more words that are not part of the ground truth, resulting in a lower BLEU score. Figure 2 further highlights this observation.

TABLE III						
STATE	INCLUSION					

State	ROUGE↑	BLEU↑	CosSim↑	MSE↓
X	0.4463	0.1360	-0.0045	0.0024
1	0.4596	0.1495	-0.0196	0.0024

C. Robot State Inclusion

From the results in Table III and Figure 4a we find that the inclusion of state tokens in our input prompt has a slightly

beneficial impact on the language generation capabilities of our model. The quality of the action remains the same, as can be seen in Figure 4b. It should be noted that introducing additional tokens for the robot state possibly increases the size of the word embedding, leading to increased computational cost.

D. Tokenization Resolution

We investigate the influence of different resolutions of the action tokens on the quality of the produced trajectories and language statements. We hypothesize that a lower resolution (fewer discretization bins) results in higher scores. We can observe this when producing the joint output of the language statement and the action trajectories, as shown in Figure 5. When only producing the action tokens we do not observe this decreasing trend, and the action trajectory quality stays roughly within the same range. The precise measurements can be found in Table IV. In addition, generating both the transparent statement and the action trajectory can be observed to have a positive impact on the quality of the action trajectories.

TABLE IV TOKENIZATION RESOLUTION

Resolution	Output	ROUGE↑	BLEU↑	CosSim↑	MSE↓
10	Action	-	-	0.0060	0.0024
25	Action	-	-	-0.0113	0.0023
50	Action	-	-	0.0152	0.0023
10	Full	0.5471	0.1696	0.1454	0.0021
25	Full	0.5338	0.1560	0.0965	0.0022
50	Full	0.5255	0.1511	0.0106	0.0024

E. Pretraining Influence

As illustrated in Table V pretraining our models on the RoboVQA dataset leads to a higher quality of action trajectories than without pretraining. At the same time, we notice slightly higher scores in the language metrics. Note that when only producing the action tokens as output, the pretraining on a language generation task does not increase the accuracy of the action tokens. A reason for this could be the fact that the task of action generation introduces new tokens to the vocabulary that have not been encountered during the pretraining phase, lowering its effectiveness.

TABLE V Pretraining

Checkpoint	Output	ROUGE↑	BLEU↑	CosSim↑	MSE↓
None	Action	-	-	0.0152	0.0023
None	Full	0.5255	0.1511	0.0106	0.0024
RoboVQA	Action	-	-	0.0108	0.0023
RoboVQA	Full	0.5309	0.1541	0.1601	0.0020

F. Freezing the Vision Encoder

We observe little differences on the language generation performance (± 0.04 on the ROUGE and BLEU scores) when



Fig. 3. Sample outputs of our model on our test set including positive and negative samples. We removed the surrounding prompt-specific tokens for readability.

training only the LLM part of the model compared to the full model, as visible in Table VI. The performance of generating action trajectories increases slightly when training the full model. It is likely that the model learns to pay more attention to different novel visual features when having to generate the next trajectory, resulting in this behavior.

VI. DISCUSSION

Learning transparency is an inherently difficult task and depends on many factors. In this work, we opt for learning natural language statements that were provided by large-scale annotations. While learning by imitating in this way is certainly

TABLE VI Results Weight and Output Configuration

Training	Output	ROUGE↑	BLEU↑	CosSim↑	MSE↓
LLM	Action	-	-	-0.0069	0.0023
LLM	Full	0.4462	0.1360	-0.0044	0.0024
Full	Action	-	-	0.0121	0.0023
Full	Full	0.4436	0.1375	-0.0018	0.0024

an option, datasets with ground truth language annotations for transparent statements in robotics are sparse and expensive to create, calling for novel methods to train transparent policies. Even when these annotations are given, creating both language utterances and actions does not mean that these two outputs are semantically aligned; e.g., there is no guarantee that the description adheres to the executed trajectory. Future work could address this by explicitly investigating measures that evaluate the quality and alignment of actions and language. With respect to the semantic differences between generated and ground truth language, our results further highlight in Section V-A the need for more advanced metrics to evaluate transparency via natural language. In addition, transparency should incorporate the human's perspective [34]. What one human might perceive as transparent could not have the same effect on another. Further research could investigate incorporating the human as context to provide different statements depending on the user's preference.



Fig. 4. Effects of including the tokenized state vector in the input prompt.

A problem we faced when training the model to predict an action embedded into the language output using the traditional Cross-Entropy Loss is that this loss assigns the same importance to every token, which leads to the model not differentiating between different degrees of error when generating the action tokens. For example, if the goal action was "a20 a25", the corresponding error will not necessarily reflect the trajectory deviation for the output of "a5 a45" (large difference in trajectory) and the output of "a21 a24" (similar trajectory). Addressing this could prove highly useful in future work using VLMs for policy generation.



Fig. 5. Effects of varying action tokenization resolutions on action quality.

Regarding a practical implementation, it is not necessary to provide a new transparent statement at each timestep, but rather once the described action has been performed. Such a mechanism could be implemented with a specific prompt or an external module which only asks for a new transparent statement when certain conditions are met.

VII. CONCLUSION

Driven by the need for more transparent robots, we presented a method to train agents to be transparent about their behavior using natural language while simultaneously learning low-level action trajectories. Our model learns transparent behavior alongside a policy by combining both tasks into a single supervised language generation problem. While we show promising results on the Language-Table dataset and find that transparency can benefit the model's action quality, we also highlight a need for methods to semantically analyze the quality of transparent agents, which we leave for investigation in further research.

REFERENCES

- [1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "RT-1: Robotics Transformer for Real-World Control at Scale," 2023. [Online]. Available: http://arxiv.org/abs/2212.06817
- [2] T. Sakai and T. Nagai, "Explainable autonomous robots: a survey and perspective," Advanced Robotics, vol. 36, no. 5-6, pp. 219–238, 2022. [Online]. Available: https://doi.org/10.1080/01691864.2022.2029720
- [3] M. T. Gervasio, K. L. Myers, E. Yeh, and B. Adkins, "Explanation to avert surprise." in *IUI Workshops*, ser. CEUR Workshop Proceedings, A. Said and T. Komatsu, Eds., vol. 2068. CEUR-WS.org, 2018. [Online]. Available: http://dblp.unitrier.de/db/conf/iui/iui2018w.html#GervasioMYA18
- [4] V. Rieser and J. Moore, "Implications for generating clarification requests in task-oriented dialogues," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, K. Knight, H. T. Ng, and K. Oflazer, Eds. Ann Arbor, Michigan: Association for Computational Linguistics, 2005, pp. 239–246. [Online]. Available: https://aclanthology.org/P05-1030
- [5] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," 2023. [Online]. Available: http://arxiv.org/abs/2307.15818
- [6] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "OpenVLA: An Open-Source Vision-Language-Action Model," 2024. [Online]. Available: https://arxiv.org/abs/2406.09246
- [7] S. Belkhale, T. Ding, T. Xiao, P. Sermanet, Q. Vuong, J. Tompson, Y. Chebotar, D. Dwibedi, and D. Sadigh, "RT-H: Action Hierarchies Using Language," 2024. [Online]. Available: http://arxiv.org/abs/2403.01823
- [8] S. Y. Schött, R. M. Amin, and A. Butz, "A Literature Survey of How to Convey Transparency in Co-Located Human–Robot Interaction," *Multimodal Technologies and Interaction*, vol. 7, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2414-4088/7/3/25

- [9] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, " π_0 : A Vision-Language-Action Flow Model for General Robot Control," 2024. [Online]. Available: https://arxiv.org/abs/2410.24164
- [10] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, T. Unterthiner, D. Keysers, S. Koppula, F. Liu, A. Grycner, A. A. Gritsenko, N. Houlsby, M. Kumar, K. Rong, J. Eisenschlos, R. Kabra, M. Bauer, M. Bosnjak, X. Chen, M. Minderer, P. Voigtlaender, I. Bica, I. Balazevic, J. Puigcerver, P. Papalampidi, O. J. Hénaff, X. Xiong, R. Soricut, J. Harmsen, and X. Zhai, "PaliGemma: A versatile 3B VLM for transfer," *CoRR*, vol. abs/2407.07726, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2407.07726
- [11] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, "PaLM-E: An Embodied Multimodal Language Model," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 8469–8488. [Online]. Available: https://proceedings.wlr.press/v202/driess23a.html
- [12] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo, Y. L. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An Open-Source Generalist Robot Policy," 2024. [Online]. Available: http://arxiv.org/abs/2405.12213
- [13] X. Li, M. Zhang, Y. Geng, H. Geng, Y. Long, Y. Shen, R. Zhang, J. Liu, and H. Dong, "ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), June 2024, pp. 18 061–18 070.
- [14] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation," in *Proceedings of The* 6th Conference on Robot Learning, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 14–18 Dec 2023, pp. 785–799. [Online]. Available: https://proceedings.mlr.press/v205/shridhar23a.html
- [15] C. Zhao, S. Yuan, C. Jiang, J. Cai, H. Yu, M. Y. Wang, and Q. Chen, "ERRA: An Embodied Representation and Reasoning Architecture for Long-Horizon Language-Conditioned Manipulation Tasks," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3230–3237, 2023, conference Name: IEEE Robotics and Automation Letters. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10097850
- [16] S. Teso and K. Kersting, "Explanatory Interactive Machine Learning," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics,* and Society, ser. AIES '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 239–245. [Online]. Available: https://doi.org/10.1145/3306618.3314293
- [17] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proceedings of the 36th International Conference* on Neural Information Processing Systems, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.
- [18] W. Stammer, F. Friedrich, D. Steinmann, M. Brack, H. Shindo, and K. Kersting, "Learning by Self-Explaining," *Transactions* on *Machine Learning Research*, 2024. [Online]. Available: https://openreview.net/forum?id=bpjU7rLjJ7
- [19] J. Duan, W. Pumacay, N. Kumar, Y. R. Wang, S. Tian, W. Yuan, R. Krishna, D. Fox, A. Mandlekar, and Y. Guo, "AHA: A Vision-Language-Model for Detecting and Reasoning Over Failures in Robotic Manipulation," in 2nd CoRL Workshop on Learning Effective Abstractions for Planning, 2024. [Online]. Available: https://openreview.net/forum?id=d3nmdJHIIS
- [20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

- [21] F. Ni, J. Hao, S. Wu, L. Kou, J. Liu, Y. Zheng, B. Wang, and Y. Zhuang, "Generate Subgoal Images Before Act: Unlocking the Chain-of-Thought Reasoning in Diffusion Model for Robot Manipulation with Multimodal Prompts," in 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13991–14000.
- [22] R. Liu, J. Wei, S. S. Gu, T.-Y. Wu, S. Vosoughi, C. Cui, D. Zhou, and A. M. Dai, "Mind's Eye: Grounded Language Model Reasoning through Simulation," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=4rXMRuoJlai
- [23] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: deliberate problem solving with large language models," in *Proceedings of the 37th International Conference* on Neural Information Processing Systems, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.
- [24] M. Kerzel, J. Ambsdorf, D. L. Becker, W. Lu, E. Strahl, J. Spisak, C. Gäde, T. Weber, and S. Wermter, "What's on Your Mind, NICO?" *KI - Künstliche Intelligenz*, vol. 36, pp. 237 – 254, 2022. [Online]. Available: https://doi.org/10.1007/s13218-022-00772-8
- [25] S. Wallkötter, S. Tulli, G. Castellano, A. Paiva, and M. Chetouani, "Explainable Embodied Agents Through Social Cues: A Review," J. Hum.-Robot Interact., vol. 10, no. 3, Jul. 2021. [Online]. Available: https://doi.org/10.1145/3457188
- [26] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning," in *Proceedings of the 5th Conference on Robot Learning*. PMLR, Jan. 2022, pp. 991–1002, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v164/jang22a.html
- [27] S. Nair, E. Mitchell, K. Chen, brian ichter, S. Savarese, and C. Finn, "Learning Language-Conditioned Robot Behavior from Offline Data and Crowd-Sourced Annotation," in 5th Annual Conference on Robot Learning, 2021. [Online]. Available: https://openreview.net/forum?id=tfLu5W6SW5J
- [28] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, 1st ed. USA: John Wiley & Sons, Inc., 1994.
- [29] P. Sermanet, T. Ding, J. Zhao, F. Xia, D. Dwibedi, K. Gopalakrishnan, C. Chan, G. Dulac-Arnold, sharath maddineni, N. Joshi, P. Florence, W. Han, R. Baruch, Y. Lu, S. Mirchandani, P. Xu, P. Sanketi, K. Hausman, I. Shafran, brian ichter, and Y. Cao, "RoboVQA: Multimodal Long-Horizon Reasoning for Robotics," in *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023. [Online]. Available: https://openreview.net/forum?id=R1I94rrgDz
- [30] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence, "Interactive Language: Talking to Robots in Real Time," *IEEE Robotics and Automation Letters*, pp. 1–8, 2023.
- [31] A. O'Neill et al., "Open X-Embodiment: Robotic Learning Datasets and RT-X Models: Open X-Embodiment Collaboration," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024, pp. 6892–6903.
- [32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the* 40th Annual Meeting of the Association for Computational Linguistics, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040
- [33] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out.* Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013
- [34] A. Weller, "Transparency: Motivations and challenges," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds. Springer International Publishing, 2019, pp. 23–40. [Online]. Available: https://doi.org/10.1007/978-3-030-28954-6_2