Interactive Robot Action Replanning using Multimodal LLM Trained from Human Demonstration Videos

Hori, Chiori; Kambara, Motonari; Sugiura, Komei; Ota, Kei; Khurana, Sameer; Jain, Siddarth; Corcodel, Radu; Jha, Devesh K.; Romeres, Diego; Le Roux, Jonathan

TR2025-034 March 08, 2025

Abstract

Understanding human actions could allow robots to perform a large spectrum of complex manipulation tasks and make collaboration with humans easier. Recently, multimodal scene understanding using audio-visual Transformers has been used to generate robot action sequences from videos of human demonstrations. However, automatic ac- tion sequence generation is not always perfect due to the distribution gap between the training and test environments. To bridge this gap, human intervention could be very effective, such as telling the robot agent what should be done. Motivated by this, we propose an error-correction-based action replanning approach that regenerates better action sequences using (1) automatically generated actions from a pretrained action generator and (2) human error-correction in natural language. We collected single- arm robot action sequences aligned to human action instruction for the cooking video dataset YouCook2. We trained the proposed error- correction-based action replanning model using a pre-trained multimodal LLM model (AVBLIP-2), generating a pair of (a) single-arm robot micro-step action sequences and (b) action descriptions in natural language simultaneously. To assess the performance of error correction, we collected human feedback on correcting errors in the automatically generated robot actions. Experiments show that our proposed interactive replanning model trained in a multitask manner using action sequence and description outperformed the baseline model in all types of scores.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Interactive Robot Action Replanning using Multimodal LLM Trained from Human Demonstration Videos

Chiori Hori[1], *Motonari Kambara*[1,2], *Komei Sugiura*[2], *Kei Ota*[1], *Sameer Khurana*[1],
Siddarth Jain[1], *Radu Corcodel*[1], *Devesh Jha*[1], *Diego Romeres*[1], *Jonathan Le Roux*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA [2]Keio University, Tokyo, Japan

***Abstract*—Understanding human actions could allow robots to perform a large spectrum of complex manipulation tasks and make collaboration with humans easier. Recently, multimodal scene understanding using audio-visual Transformers has been used to generate robot action sequences from videos of human demonstrations. However, automatic action sequence generation is not always perfect due to the distribution gap between the training and test environments. To bridge this gap, human intervention could be very effective, such as telling the robot agent what should be done. Motivated by this, we propose an error-correction-based action replanning approach that regenerates better action sequences using (1) automatically generated actions from a pretrained action generator and (2) human error-correction in natural language. We collected single-arm robot action sequences aligned to human action instruction for the cooking video dataset YouCook2. We trained the proposed error-correction-based action replanning model using a pre-trained multimodal LLM model (AVBLIP-2), generating a pair of (a) single-arm robot micro-step action sequences and (b) action descriptions in natural language simultaneously. To assess the performance of error correction, we collected human feedback on correcting errors in the automatically generated robot actions. Experiments show that our proposed interactive replanning model trained in a multitask manner using action sequence and description outperformed the baseline model in all types of scores.**

***Index Terms*—Robot action generation, Interactive error correction, Human-robot collaboration, Multimodal scene understanding, Multimodal LLM**

I. Introduction

Effective human-robot collaboration for shared goals is necessary for the seamless integration of robots in human daily lives. To realize such effective human-robot collaborative systems, multimodal scene understanding is essential to provide robots with the capability to interpret their environment and interact with humans based on such understanding. An initial attempt to generate an action sequence for a single-arm manipulator using an audio-visual transformer trained from demonstration videos of human cooking was reported earlier in [1]. However, the semantic representation capability for multimodal reasoning was limited because the training data did not cover all possible patterns from the fusion of different modalities. To address this data sparsity issue, we extended BLIP2 [2] into a multimodal large language model (AVBLIP-2) for generating robot actions [3]. We used various features, such as audio, visual, speech, and text, that are embedded into the semantic space of a large language model (LLM) by training a Q-former using both contrastive loss and action generation loss. The multimodal LLM contributes to enhancing the performance of robot action generation [3]. However, automatic action sequence generation is still not perfect when applying the trained model to the real world due to the differences in the training data and test environments. In such situations, human intervention could be useful in correcting the proposed incorrect sequence by providing expert guidance on what should be done.

This paper proposes an interactive planning approach for error correction using multimodal LLMs. We consider the use of a single-arm robot as the manipulator in our proposed work. In general, tasks which can be easily performed by humans need to be properly broken down in micro-step action sequences for a single-arm robot. We found that there is no prior dataset for micro-step action sequence prediction for single-arm robots. We thus first collect single-arm robot action sequences for cooking videos in the YouCook2 dataset [4] by asking Amazon Mechanical Turk (AMT) workers to elaborate human action instructions such as "stir soup" into micro-step action sequences such as "pick up spatula from counter, stir soup in pot, place spatula on counter". We then trained a multimodal LLM model (Audio Visual Bootstrapping Language Image Pre-training 2: AVBLIP-2) [3] for the generation of robot action sequences using the collected data as shown in Fig. 1. Futhermore, robots need to be able to understand human instruction and describe their own actions using natural language to interact with humans. Robots can use natural language to ask humans whether their action planning is correct, and humans can correct their actions before the robot takes action. To implement the confirmation function by robots, we modified AVBLIP-2 to generate a robot action description in natural language aligned to micro-step action sequences. We trained an error correction model using the original instruction in YouCook-2 as pseudo error correction data. To evaluate the performance of the error correction model, we asked AMT workers to rectify the automatic action description by comparing it with the original instructions. The proposed error-correction based action replanning model was trained using the pre-trained AVBLIP-2 that generates robot action description and micro-step action sequences simultaneously.

The main contributions of this work are (1) collecting single-arm micro-step action sequences for instruction videos, (2) proposing an interactive replanning approach for robot action generation based on error correction, that first generates two-style robot actions, (a) single-arm robot micro-step action sequence and (b) action description in natural language, and then feeds them back to the model together with human error correction, and (3) demonstrating the effectiveness of the proposed interactive planning model for robot action sequence correction in the cooking domain.

II. Related work

To build Human Robot Interaction (HRI) systems, there have been some researches on language acquisition by robots to find associations between actions, objects, properties, and effects, and to map those associations to language [5], [6]. However, it is impossible to train models handling a huge vocabulary practically in real situations with all kinds of robots. Thus, we segment robot manipulation into a skill acquisition phase and a knowledge acquisition phase and propose approaches for the knowledge acquisition in this paper.

Recently, Large Language Models (LLMs) have achieved impressive results in creating robotic agents that perform open-vocabulary tasks such as CLIPort [7], SayCan [8]. PROGPROMPT [9] introduces a programmatic LLM prompt structure that facilitates the generation of plans in diverse environments, robot functionalities, and tasks. LLM-POP [10] targets partially observable task planning, leveraging

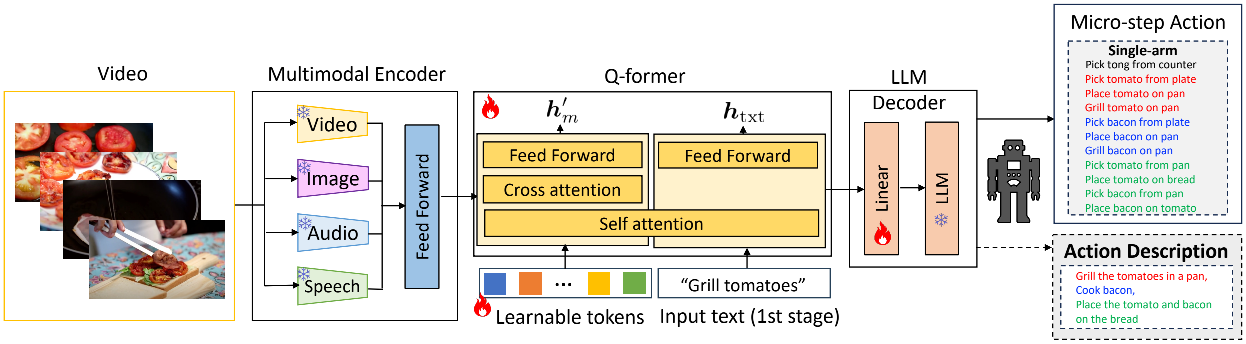

Fig. 1. AVBLIP-2 for cooking demonstration videos [3]. Burning and frozen marks represent the trainable and frozen parameters, respectively. We updated AVBLIP-2 to generate action description in natural language simultaneously with micro-step action sequences. Although the prior work [3] used two-arm action sequences, we collected single-arm robot micro-step actions. The dashed line shows the new elements in this work.

an LLM to gather environmental data through a robot, deduce task states from collected observations, and direct the robot to execute necessary actions. LLMs expand vocabulary and context considerations, while visual grounding LLMs enhance spatial reasoning capabilities. COWP [11] introduces an LLM-based open-world task planning system for robots. Some works explore using LLMs to directly predict a dense sequence of end-effector poses for robot actions with vision models [12]. Recent developments have extended the use of an audio-visual Transformer for multimodal scene understanding to generate a micro-step action sequence based on human demonstrations [1]. To enhance the micro-step action sequence generation using LLM, a multimodal LLM (AVBLIP-2) has been employed to interpret human behaviors [3]. Despite these advances, errors persist in the current action-generation frameworks, with no provision for correcting automatically generated actions. Human-robot collaboration has the potential to enable humans to rectify errors produced by automatic action generators. An interactive method for robot action planning [13] enables LLM analysis and collects missing information by engaging humans through questions, although it requires a substantial number of tokens. Another study [14] explores re-prompting strategies to enhance the executability and accuracy of LLM-generated plans but relies strictly on templated prompts. In our research, we present an error-correction-based action replanning model for robot action generation, leveraging the multimodal LLM AVBLIP-2. Our approach involves automatically generated actions obtained from the multimodal LLM, alongside human action instruction in natural language.

III. Data Preparation

Collection for Micro-step Action: To generate a micro-step action sequence that a single-arm robot could perform, we translated human instructions in YouCook2 [4] into micro-step action sequences. AMT workers generated micro-step action sequences for single-arm robot by selecting words for four placeholders, such as "single-arm action", "target object", "preposition", and "place", to achieve the same actions by humans as shown in Fig. 1. Single-arm actions were selected from the following 12 candidates: Open, Close, Pick, Place, Pour, Stir, TurnOn, TurnOff, Wipe, Cut, Scoop, Squeeze. The target objects were selected as one of the nouns in the human action instruction as much as possible.

Robot Action Description: YouCook2 is already annotated with human instructions in natural language to describe human cooking action steps, as shown in [4] . We used the human action instruction to train a model generating robot action description to interact with humans to confirm its action planning is correct.

IV. Action Generation using AVBLIP-2

In this work, we employ AVBLIP-2 [3], an extension of BLIP-2 [2], a vision-language pre-training method. BLIP-2 bootstraps from a frozen image encoder and a frozen large language model, where a Querying Transformer (Q-former) [15] is trained to bridge the gap between the vision and text modalities. In AVBLIP-2, the image encoder is replaced with audio-visual encoders that encode video, audio, and text feature sequences.

As shown in Fig. 1, AVBLIP-2 mainly consists of two modules: Q-former and LLM Decoder. The Q-former is trained to extract a fixed number of output features from multimodal encoder outputs with different lengths. It has two transformer submodules that share the same self-attention layers: (1) a multimodal transformer that interacts with the frozen audio-visual encoders and (2) a text transformer that works as a text encoder and a text decoder. It also has a set of learnable query embeddings as input to the multimodal transformer. The queries interact with each other through self-attention layers and interact with audio-visual features through cross-attention layers. The queries can additionally interact with the text through the same cross-attention layers. Finally, the queries are converted to an output feature.

The LLM Decoder generates a sequence of micro-step actions for a single-arm robot from multimodal features aligned to language features obtained by the Q-former. The LLM Decoder is constructed with a frozen LLM and a feed-forward layer. By using the LLM as a decoder, it leverages the LLM's inference capabilities when generating action sequences. In this study, we use OPT-2.7B [16] as the LLM.

The training of AVBLIP-2 consists of two stages: (1) vision-language representation learning with frozen multimodal encoders and (2) vision-to-language generative learning with a frozen LLM. In the second stage, we connect the Q-former to the frozen LLM Decoder and perform multimodal action sequence generation. As shown in Fig. 1, we process the multimodal features obtained by the Q-former by using a fully-connected layer. Then, the LLM Decoder generates action sequences from the features. We use the cross-entropy loss function in this stage.

V. Error Correction-based Replanning Model

Although AVBLIP-2 can generate a sequence of actions in various situations by utilizing the generalization capability of LLMs, it can sometimes generate wrong plans due to the distribution gap between training and testing environments. To fill the gap, we extend the AVBLIP-2-based micro-step action sequence generation by introducing an error correction module.

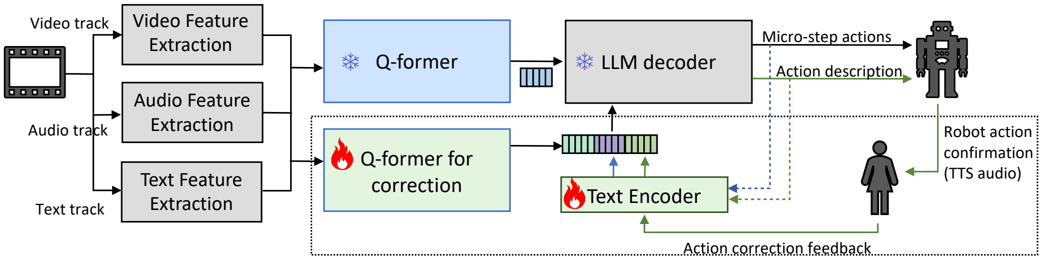

Fig. 2. Error-correction-based action replanning model. AVBLIP-2-based action generation with an error correction module (in the dashed box), that encodes the generated action sequence and human feedback with the text encoder, and then the encoded text and the Q-former output are fed to the LLM as a prompt to regenerate a corrected action sequence. Robot action confirmation utterances are generated using robot action description.

Figure 2 shows the extended system with the additional components in the dashed box. The first pass generates a micro-step action sequence through multimodal feature extraction, Q-former-based feature encoding, and LLM-based sequence generation.

For interactive action replanning, we extend the model to generate a natural language action description in addition to the action sequence to confirm the robot's action to the human. We train AVBLIP-2 to generate the two output sequences. In this work, we compare two approaches: (1) Target sequence concatenation, where we concatenate the action description and the action sequence between which we insert a special token "<a>", and (2) Multitask training, for which we duplicate the training data, where one half has action description targets with prompt "<d>" and the other half has action sequence targets with prompt "<a>".

The error correction pass encodes the generated action description and human error-correction sentence with the text encoder. Then the encoded text and the Q-former output are fed to the LLM as a prompt to regenerate a corrected action sequence. The Q-former for error correction is separately trained to generate correct action sequences from the first-pass outputs and the human error-correction sentence.

The text encoder is trained jointly with the Q-former for error correction, where the multimodal encoders and the LLM remain frozen. The text encoder can be a transformer encoder or just a linear projection on top of the word embedding layer of the LLM. In this work, we use a linear projection.

VI. Experiments

A. Setup

We test the proposed method using YouCook2 consisting of cooking action video clips aligned with human action instruction in natural language. Each video in the dataset consists of 5 to 16 human cooking steps. The dataset contains 2K videos of 89 recipes, divided into training, validation, and test sets, containing 1,333, 457, and 210 videos, respectively. In this work, we used the validation set as the test set because the test set was not publicly available. We collected micro-step action sequences as described in Section III, which contain 2,790 unique phrases of 195 verbs, 2,229 objects, 33 prepositions, and 1002 places. The training and validation sets include 8,928 and 3,122 action sequences, each of which corresponds to a human cooking step. The vocabulary size of the training set is 1037. The out-of-vocabulary rate of the validation set is 6.6%. The average sequence lengths are 25.3 and 25.5 words, respectively.

Multimodal features such as video, image, and audio are extracted using Omnivore [17], Contrastive Language-Image Pre-Training (CLIP) [18], and Audio Spectrogram Transformer (AST) [19], respectively. The image and video features are concatenated and projected to a single video feature sequence before feeding them to the encoder. If a subtitle is available in the video, text features are extracted by Glove word embedding [20]. Otherwise, we feed an embedding vector for the <unk> label instead. The numbers of dimensions of the audio, visual, and text features are 768, 1024, and 300, respectively.

We initialize the Q-former with the pre-trained weights of BERT-base [21], while the cross-attention layers are randomly initialized. We set the number of dimensions of the hidden layers to 768, which results in 188M parameters in total. In the experiments, we use 32 queries, where each query has a 768-dimensional vector, which equals the hidden dimension of the Q-former.

To evaluate the quality of the generated action sequences, we collected human error correction for automatically generated robot actions as shown in Fig. 3. Humans generated sentences to correct the errors based on the original instructions and the automatic action confirmation.

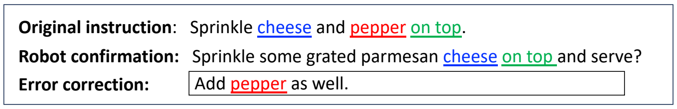
Original instruction: Sprinkle cheese and pepper on top.
Robot confirmation: Sprinkle some grated parmesan cheese on top and serve?
Error correction: Add pepper as well.

Fig. 3. Robot action error correction by AMT workers.

The performance was evaluated using the BLEU-2 and METEOR scores computed between the generated and ground-truth sequences used in the robotics field [22], [23].

B. Results

Table II shows the quality of the action sequences and the action descriptions generated using different models, where "Baseline" denotes the Q-former model trained with the pairs of audio-visual features and their target action sequences. The baseline model only outputs action sequences, while the extended models based on target concatenation (concat) and multitask training (multitask) can generate both action sequences and action descriptions. The multitask training achieves action quality slightly better than the baseline despite generating two sequences. On the other hand, the target concatenation model is not better than the multitask model. Looking at some low-quality outputs, there are over-generated descriptions that are much longer than the references. In addition, the action sequences are prone to be sensitive to action description errors. We also measured the ratio of overlapping words[1] in generated action sequence/description pairs because these two outputs should indicate the same goal. The target concatenation model has a higher overlapping ratio than the multitask model. This is because the model generates each action sequence explicitly conditioned on the description. However, considering that the ratio for the reference sequence pairs was 0.64, the overlapping ratio is high enough with the multitask model. Thus, we decided to

[1] Computed as the ratio of the content words that appeared in both the action sequence and description to all content words in the description.

TABLE I

EXAMPLES OF ERROR CORRECTION. OBJECT-LEVEL ERRORS ARE HIGHLIGHTED IN RED. CORRECTED OBJECT NAMES ARE HIGHLIGHTED IN BLUE.

Video Id	Sequence type	Action sequence or human feedback
xHr8X2Wpmno	Reference	Pick lettuce, Place lettuce on cutting board, Pick knife, Chop lettuce, Place knife on counter, Pick bowl, Place bowl on counter, Pick lettuce, Place lettuce in bowl
	1st-pass output	Pick radish, Place radish on cutting board, Pick knife, Cut radish with knife, Place knife on counter, Pick radish,
	Human feedback	Chop lettuce and place it in a bowl
	Corrected output	Pick lettuce, Place lettuce on counter, Pick knife, Cut lettuce with knife, Place knife on counter, Pick lettuce, Place lettuce in bowl
c9eELn4axpg	Reference	Pick eggs, Add eggs to pan, Place eggs on counter, Pick pepper, Sprinkle pepper, Place salt on counter, Pick pepper, Pour pepper into pan
	1st-pass output	Pick olive oil, Pour olive oil into pan, Place olive oil on counter, Pick salt, Sprinkle salt into pan, Place salt on counter
	Human feedback	Add 2 eggs and season with salt and pepper
	Corrected output	Pick eggs, Pour eggs into pan, Place eggs on counter, Pick salt, Pour salt into pan, Place salt on counter, Pick pepper, Pour pepper into pan
xPiv3hP5888	Reference	Pick frozen peas, Pour frozen peas into pan, Place frozen peas on counter, Pick spatula, Stir ingredients in pan
	1st-pass output	Pick spices, Pour spices into pan, Place spices on counter, Pick spatula, Stir pan
	Human feedback	Add frozen peas to the pan and stir
	Corrected output	Pick frozen peas, Place frozen peas in pan, Pick spatula, Stir peas in pan
RlIWJUvrxEY	Reference	Pick foil, Place foil on counter, Pick sandwich, Place sandwich on foil, Fold foil over sandwich
	1st-pass output	Pick pita, Place pita on grill
	Human feedback	Fold the foil around the sandwich
	Corrected output	Pick foil, Fold foil around sandwich

TABLE II

QUALITY OF GENERATED ACTION SEQUENCE AND DESCRIPTION.

	Action sequence		Action description		
	BLEU-2	METEOR	BLUE-2	METEOR	Word overlap
Baseline	0.356	0.249	-	-	
Concat	0.346	0.243	0.198	0.143	**0.71**
Multitask	**0.370**	**0.257**	**0.220**	**0.158**	0.62

TABLE III

QUALITY OF GENERATED ACTION SEQUENCE AND DESCRIPTION AFTER ERROR CORRECTION FEEDBACK.

	Action sequence		Action description	
	BLEU-2	METEOR	BLEU-2	METEOR
Multitask	0.370	0.257	0.220	0.158
EC w/ 1st-pass feedback	0.375	0.258	0.231	0.161
EC w/ human feedback	**0.408**	**0.281**	**0.398**	**0.303**
EC w/ both	0.379	0.262	0.331	0.246

use the multitask model for the rest of the experiments. For reference, we also evaluated the case of complete feedback, where we used the reference descriptions as human feedback.

Table III shows the quality of the generated sequences. We used the multitask model to generate the 1st-pass action sequence and description. "EC" denotes the error correction model trained with three different feedbacks: feedback from the 1st-pass action sequence output, human feedback, and both. As shown in the Table, the 1st-pass feedback provides a slight improvement on the two metrics over the baseline. However, the gains are very limited without the human feedback, which are less than 2% relative. This is probably because the model was learned from insufficient pairs of error patterns and correct answers. Since we used only single-best hypotheses for training, we need data augmentation to increase the diversity of the error-correction samples. On the other hand, the human feedback significantly improves the quality of action sequences, showing relative gains of 8 – 10%. This result demonstrates that the proposed approach can correct erroneous action sequences by human feedback in natural language. However, error correction with both types of feedback does not show further improvement compared to human feedback only. This also implies that the data sparseness issue exists in the training with only single-best hypotheses. This issue will be addressed in future work.

Table I shows examples of generated action sequences, where each row contains the reference, the 1st-pass output (baseline), the human feedback, and the corresponding corrected (regenerated) sequences. In the first example, the 1st-pass action generation misrecognizes "lettuce" as "radish" and does not emit "bowl". With human feedback "Chop lettuce and place it in a bowl", the errors are corrected in the output sequence, where "lettuce" and "bowl" are placed correctly in a one-hand action sequence. We can also see similar error corrections for object names, e.g., "olive oil" to "eggs" in the second example and "spices" to "frozen peas" in the third example, which are all corrected as the human feedback contained such keywords. In the fourth example, the object "pita" is corrected with "foil" and "sandwich", but the output contains the human feedback as it is. This is a failure example of error correction since the actions are not in the one-hand style and, therefore, the robot cannot follow the sequence.

VII. CONCLUSIONS

This paper proposed a method for error-correction-based action replanning approach that feeds (1) actions generated by a multimodal LLM and (2) human error-correction in natural language. We collected micro-step action sequences for single-arm robot using the cooking video dataset YouCook2[2]. We trained the proposed error-correction-based action replanning model using a pre-trained multimodal LLM model (AVBLIP-2) generating a pair of (a) single-arm robot micro-step action sequences and (b) robot action description in natural language simultaneously. Experiments show that our proposed interactive replanning model trained in a multitask manner using action sequence and description outperformed the baseline model for all types of scores. This work is the first attempt to apply multimodal understanding trained from human demonstration videos for robot action planning aligned with action description in natural language. We will open the data and our baseline to accelerate research in this direction.

[2] Hiroto Takeuchi, a student at Rochester Institute of Technology, collected robot action sequences as an intern at MERL.

References

[1] C. Hori, P. Peng, D. Harwath, X. Liu, K. Ota, S. Jain, R. Corcodel, D. Jha, D. Romeres, and J. Le Roux, "Style-transfer based Speech and Audio-visual Scene understanding for Robot Action Sequence Acquisition from Videos," in *Proc. Interspeech*, 2023, pp. 4663–4667.

[2] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. ICML*, 2023.

[3] "Human action understanding-based robot planning using multimodal LLM," in *Proc. ICRA Workshop for "Cooking Robotics: Perception and motion planning"*, 2024.

[4] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proc. AAAI*, 2018.

[5] G. Saponaro, L. Jamone, A. Bernardino, and G. Salvi, "Beyond the self: Using grounded affordances to interpret and describe others' actions," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 2, pp. 209–221, 2020.

[6] G. Salvi, L. Montesano, A. Bernardino, and J. Santos-Victor, "Language bootstrapping: Learning word meanings from perception–action association," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 3, pp. 660–671, 2012.

[7] M. Shridhar, L. Manuelli, and D. Fox, "CLIPort: What and where pathways for robotic manipulation," in *Proc. CoRL*, 2021.

[8] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do as I can, not as I say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[9] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," in *Proc. ICRA*, 2023.

[10] L. Sun, D. K. Jha, C. Hori, S. Jain, R. Corcodel, X. Zhu, M. Tomizuka, and D. Romeres, "Interactive planning using large language models for partially observable robotic tasks," in *Proc. ICRA*, 2024.

[11] Y. Ding, X. Zhang, S. Amiri, N. Cao, H. Yang, A. Kaminski, C. Esselink, and S. Zhang, "Integrating action knowledge and LLMs for task planning and situation handling in open worlds," *Autonomous Robots*, vol. 47, no. 8, pp. 981–997, 2023.

[12] T. Kwon, N. Di Palo, and E. Johns, "Language models as zero-shot trajectory generators," *IEEE Robotics and Automation Letters*, 2024.

[13] K. Hori, K. Suzuki, and T. Ogata, "Interactively robot action planning with uncertainty analysis and active questioning by large language model," in *Proc. SII*, 2024.

[14] S. S. Raman, V. Cohen, E. Rosen, I. Idrees, D. Paulius, and S. Tellex, "Planning with large language models via corrective re-prompting," in *NeurIPS Foundation Models for Decision Making Workshop*, 2022.

[15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020.

[16] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan *et al.*, "OPT: Open Pre-trained Transformer Language Models," *arXiv preprint arXiv:2205.01068*, 2022.

[17] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, "Omnivore: A Single Model for Many Visual Modalities," in *Proc. CVPR*, 2022.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.

[19] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech*, 2021, pp. 571–575.

[20] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[22] A. Nguyen, D. Kanoulas, L. Muratore, D. G. Caldwell, and N. G. Tsagarakis, "Translating videos to commands for robotic manipulation with deep recurrent neural networks," in *Proc. ICRA*, 2018.

[23] X. Xu, K. Qian, B. Zhou, S. Chen, and Y. Li, "Two-stream 2D/3D residual networks for learning robot manipulations from human demonstration videos," in *Proc. ICRA*, 2021.