MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Human Action Understanding-based Robot Planning using Multimodal LLM

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

TR2024-066 June 11, 2024

Abstract

In future smart homes, robots are expected to handle everyday tasks such as cooking, replacing human involvement. Acquiring such skills autonomously for robots is highly challenging. Consequently, existing methods address this issue by collecting data by controlling real robots and training models through supervised learning. However, data collection for long-horizon tasks could be very painful. To solve this challenge, this work focuses on the task of generating action sequences for a robot arm from human videos demonstrating cooking tasks. The quality of generated action sequences by existing methods for this task is often inadequate. This is partly because existing methods do not effectively process each of the input modalities. To address this issue, we propose AVBLIP, a multimodal LLM model for the generation of robot action sequences. Our main contribution is the introduction of a multimodal encoder that allows multiple modalities of video, audio, speech, and text as inputs. This allows the generation of the next action to take into account both the speech information by humans and the audio information generated by the environment. As a result, the proposed method outperforms the baseline method in all standard evaluation metrics.

IEEE International Conference on Robotics and Automation (ICRA) 2024

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

Human Action Understanding-based Robot Planning using Multimodal LLM

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

***Abstract*— In future smart homes, robots are expected to handle everyday tasks such as cooking, replacing human involvement. Acquiring such skills autonomously for robots is highly challenging. Consequently, existing methods address this issue by collecting data by controlling real robots and training models through supervised learning. However, data collection for long-horizon tasks could be very painful. To solve this challenge, this work focuses on the task of generating action sequences for a robot arm from human videos demonstrating cooking tasks. The quality of generated action sequences by existing methods for this task is often inadequate. This is partly because existing methods do not effectively process each of the input modalities. To address this issue, we propose AVBLIP, a multimodal LLM model for the generation of robot action sequences. Our main contribution is the introduction of a multimodal encoder that allows multiple modalities of video, audio, speech, and text as inputs. This allows the generation of the next action to take into account both the speech information by humans and the audio information generated by the environment. As a result, the proposed method outperforms the baseline method in all standard evaluation metrics.**

I. Introduction

A robot helper that can perform daily household tasks could be very valuable in future smart homes for assisting older or disabled people. However, it is challenging to design robot agents which can perform such household tasks. Teaching these skills using human expert demonstrations provides a potential solution. For example, in a cooking task, humans could instruct how to cut and heat ingredients in detail. However, teaching step-by-step instructions for long-horizon tasks could be very tedious. Thus, approaches to mitigate tedious human expert demonstrations are very desirable. As statistical models for speech recognition, machine translation, video captioning, and human-machine dialog systems are trained using data aligned between input signal and output text, robot action sequences can be trained from human demonstration videos aligned with action captions [1]. It would be highly efficient to automatically generate action sequences for a robot from these demonstration videos. On the other hand, the quality of action series generated by those existing methods is still insufficient. Therefore, we focus on the task of generating action sequences for a manipulator arm for cooking tasks.

Robot actions can be designed in a cascaded manner. For example, consider (1) a long horizon goal, {cook sandwich} that can be broken down into (2) short horizon steps such as {grill tomato, cook bacon, place tomato and bacon on top of bread}. Furthermore, those are broken down to (3) micro-manipulation steps such as {pick, place, cut}, which can be executed by a robot agent. Figure 1 shows such an example. A video demonstration of a task "cook sandwich" performed by a human is given to the system to generate a sequence "grill tomato, cook bacon, place tomato and bacon on top of bread". Although this task is closely related to video captioning tasks, the difficulty of our task is that the output sequence must be in the order in which a robot arm can execute them. For instance, when the robot has only one arm, it cannot pick tomatoes and a piece of bacon to put on the bread at the same time. Therefore, in that case, it is preferable to repeat the process of grasping and placing one by one. Thus, models are required to predict tasks which can predict subtasks based on their feasibility.

Several existing studies propose action sequence generation models for manipulators from human demonstration videos. Some methods were also proposed to generate an action sequence for a manipulator using an audio-visual transformer trained from demonstration videos [2]. On the other hand, these existing methods do not use large language models (LLMs) and do not take advantage of the common knowledge that LLMs have. Our method *AVBLIP*, Audio Visual Bootstrapping Language Image Pre-training, is distinctive for integrating different perceptual inputs via a multimodal encoder. This encoder processes a diverse array of inputs, including video, speech, and text, facilitating a comprehensive understanding of the task at hand by assimilating both the visual demonstrations and auditory instructions from the environment.

A key innovation in our approach is the deployment of a Query-Transformer (Q-Former) that translates the multimodal sensory input into "text-like" representations that can be ingested by a backend LLM. The LLM conditioned on these "text-like" representations generates actionable sequences for robot manipulation. Unlike BLIP-2, the Q-Former in our work is multimodal. Furthermore, we leverage a LLM as a decoder within our framework. This integration allows us to use the extensive knowledge and inferential capabilities inherent in LLMs to refine the generated action sequences. Through this integration, we showcase the potential of using advanced LLMs for robot manipulation. The use of LLMs in the context of robotics is also showcased in [3]–[5].

Through several experiments on standard benchmarks, we

[1]The authors are with Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, 8th Floor, Cambridge, MA 02139-1955, United States.

[2]The authors are with Keio University, 3-14-1 Hiyoshi, Kohoku, Yokohama, Kanagawa 223-8522, Japan. {motonari.k714}@keio.jp

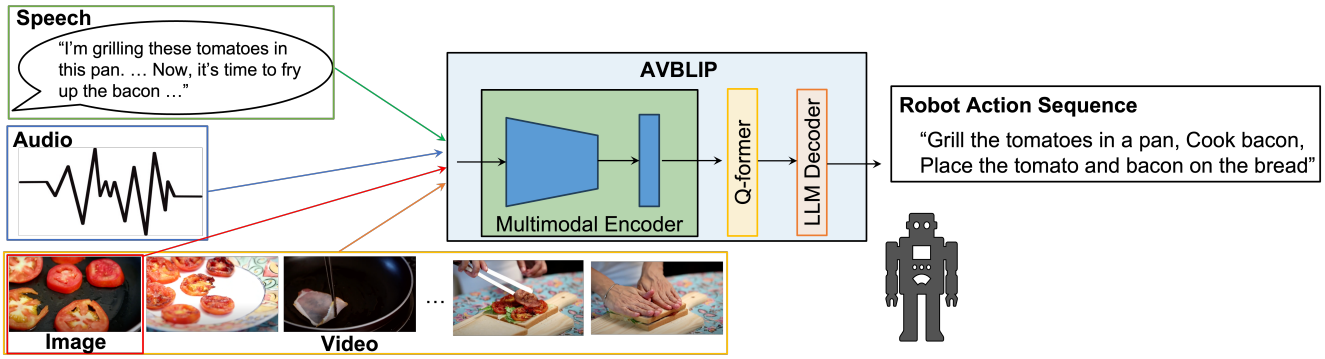

Fig. 1: This work studies how to generate a sequence of actions for a single-arm robot from a human demonstration video. The main contribution is that we propose a novel framework, called AVBLIP, capable of generating robot actions based on human demonstrations. A key component of our approach is the introduction of a multimodal encoder designed to concurrently process video, image, audio, and speech features. By incorporating this encoder, we can effectively leverage additional contextual information such as human speech and environmental sounds from the audio input, thereby enhancing the overall performance of the generated tasks. We also use an LLM as a decoder in the action sequence generation task. This makes it possible to refine the generated actions using the inference capability of the LLM.

showcase the superiority of our method. This underscores the effectiveness of our integrated multimodal Q-Former LLM approach in generating complex action sequences for robots, particularly in the cooking domain. While our immediate application is centered on cooking tasks, the architectural design and underlying principles of AVBLIP have broader applicability in robotic autonomy and assistive technologies.

II. Related Work

Learning robot skills from videos has been an active area of research in robotics and computer vision [6]–[11]. Several methods are used to acquire the robot's ability to perform object manipulation tasks from human demonstration videos [8], [11]. For example, R3m is a representation learning method using the Ego4D dataset [12]. Also, MUTEX is an imitation learning method based on multi-modalities such as speech, text, and video. Because these methods are imitation learning methods that directly predict the trajectory of the manipulator, they are labor-intensive because each trajectory must be taught to acquire skills for various tasks.

Recently, Large Language Models (LLMs) have achieved impressive results in creating robotic agents to perform open-vocabulary tasks [3], [5], [13]–[15]. For example, in Code as Policies [14], LLM writes Python codes to execute natural language directives given using pre-defined APIs. Tidybot [15] also uses LLM to plan action sequences with pre-defined APIS in tidy-up tasks. SayCan [16] expanded the capability of vision-language grounding models for robot action learning using LLMs.

LLM-POP [3] targets partially observable task planning, leveraging an LLM to collect 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 grounded LLMs enhance spatial reasoning capabilities. Recent developments have extended the use of an audio-visual Transformer for multimodal scene understanding to generate a sequence of actions for single-arm robots based on human demonstrations [2].

These methods use natural language instructions or images as input and do not assume video input. Therefore, they cannot acquire knowledge about a task from human demonstration videos of the task. Hence, it is inefficient because it cannot take advantage of existing large video data sets. On the other hand, since the proposed method generates action sequences from demonstration videos, it can utilize existing video datasets.

III. Problem Statement

This study focuses on generating a sequence of robot actions from a human demonstration video. Figure 1 shows an example of this task. Given a video in which a man makes tortillas, it is required to generate an action sequence like "Grill the tomatoes in a pan, cook bacon, place the tomato and bacon on the bread." Input and output in the task are defined as follows:

- Input: A human demonstration video including audio waveform, and speech transcription.
- Output: An action sequence.

IV. Proposed Method

In this work, we propose AVBLIP, an extension of BLIP-2 [17], 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) [18] is trained to bridge the gap between the vision and text modalities. As shown in Fig. 2, AVBLIP mainly consists of three modules: Multimodal Encoder, Q-former, and LLM Decoder. We will detail each part in the following sections. The input to the network is a human demonstration video $\boldsymbol{V} = \{\boldsymbol{v}_i | i = 1, \ldots, T\}$, an audio waveform $\boldsymbol{A}$, and a speech transcription $\boldsymbol{S}$. Here, $\boldsymbol{v}_t$ represents an image at time t.

The AVBLIP training procedure follows the procedure proposed in BLIP-2 and consists of two stages: (1) vision-language representation learning with frozen multimodal encoders and (2) vision-to-language generative learning with a frozen LLM.

Vision-language representation learning. In the first stage, the objective is to align the multimodal feature $\boldsymbol{h}_m$

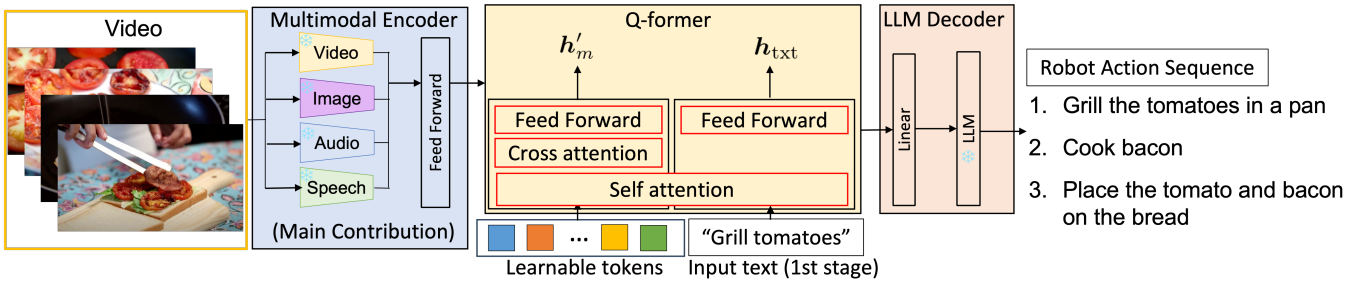

Fig. 2: The overview of AVBLIP. AVBLIP takes a video clip, images included in the video, audio waveform, and speech transcription as an input and predicts an action sequence for a single-arm robot. The main difference from BLIP-2 [17] is the introduction of the Multimodal Encoder, which allows proper handling of four different modalities: video, image, audio, and speech. The snow symbols indicate frozen layers during training.

with the text features obtained from the action sequences in the Q-former.

In the Q-former, the multimodal transformer computes cross-attention between the learnable tokens $\{\boldsymbol{z}_j | j = 1, \ldots, N\}$ and $\boldsymbol{h}_m$, the multimodal feature extracted by the multimodal encoder. Finally, the multimodal transformer outputs $\boldsymbol{h}'_m \in \mathbb{R}^{N \times d}$, where N and d denote the number of learnable tokens and the dimension of the tokens, respectively. On the other hand, a text transformer computes self-attention of an input action sequence $\boldsymbol{T}$. Finally, this transformer outputs the first token of the feature as the text feature $\boldsymbol{h}_{\mathrm{txt}}$.

In this stage, three types of pre-training objectives are employed to align the multimodal features of audio, video, and speech with the language features: Video-Text Contrastive Learning (VTC), Video-grounded Text Generation (VTG), and Video-Text Matching (VTM). The objective function of VTC is shown below:

$$\mathcal{L}_{\mathrm{vtc}} = \frac{1}{2} \left(\mathcal{L}_{CE}(s_{\mathrm{m2t}}, s_{\mathrm{ref}}) + \mathcal{L}_{CE}(s_{\mathrm{t2m}}, s_{\mathrm{ref}}) \right),$$

where $s_{\mathrm{m2t}} = \frac{\max(\boldsymbol{h}'_m \cdot \boldsymbol{h}_{\mathrm{txt}}{}^\top)}{\tau}$, $s_{\mathrm{t2m}} = \frac{\max(\boldsymbol{h}_{\mathrm{txt}} \cdot \boldsymbol{h}'_m{}^\top)}{\tau}$. Furthermore, the s_{ref} denotes the reference labels, specifically the index of the correct pair of action sequences and demonstration videos. VTC maximizes mutual information between multimodal features and text features by using contrastive learning. This involves maximizing the multimodal text feature similarity of positive pairs.

Next, VTG learns to minimize the prediction error of each token when generating action sequences using multimodal features. The objective function of this is as follows:

$$\mathcal{L}_{\mathrm{vtg}} = \mathcal{L}_{CE}(\boldsymbol{T}, f_c(\boldsymbol{h}_{\mathrm{txt}})),$$

where $\mathcal{L}_{CE}(\cdot)$ and $f_c(\cdot)$ represent the cross entropy loss function and a linear layer, respectively, and $\boldsymbol{T}$ is the ground-truth action sequence from a dataset.

Finally, VTM aims to acquire more detailed alignment capabilities than VTC by addressing a binary classification task, predicting which action sequence as a whole is paired with which demonstration video. The objective function of VTM is as follows:

$$\mathcal{L}_{\mathrm{vtm}} = \mathcal{L}_{\mathrm{BCE}}(\boldsymbol{h}'_m),$$

where $\mathcal{L}_{\mathrm{BCE}}(\cdot)$ denotes the binary cross entropy loss function. The loss function at this stage can be written as follows from the above:

$$\mathcal{L} = \mathcal{L}_{\mathrm{vtc}} + \mathcal{L}_{\mathrm{vtg}} + \mathcal{L}_{\mathrm{vtm}}$$

Vision-to-language generative learning. In the second stage, we connect the Q-former to the LLM Decoder and perform multimodal action sequence generation. In this stage, we update parameters included in the layers not marked with frozen in Fig. 2. As shown in Fig. 2, we process the $\boldsymbol{h}'_m$ obtained by the Q-former by using a linear layer. Note that we do not use the text transformer in this stage.

Then, the LLM Decoder generates action sequences from the features. We use the cross-entropy loss function as a loss function in this stage.

A. Model Architecuture

Multimodal Encoder. From the network input, the multimodal encoder extracts four types of features: video, image, audio, and speech (text). An input to this module is a human demonstration video. The output of this module is the intermediate feature $\boldsymbol{h}_m$.

Q-former [17]. This module learns to align $\boldsymbol{h}_m$ with text features obtained from action sequences. The inputs to this module are $\{\boldsymbol{z}_j | j = 1, \ldots, N\}$ and $\boldsymbol{h}_m$. In the first-stage training, described above, $\boldsymbol{T}$ is also input. This module extracts a latent vector $\boldsymbol{h}'_m$.

As shown in Fig. 2, the Q-former has two transformer submodules that share the same self-attention layers: (1) a multimodal transformer and (2) a text transformer that works as a text encoder and a text decoder.

LLM Decoder. This predicts an action sequence $\boldsymbol{y}$ from the text feature $\boldsymbol{h}'_m$ obtained by the Q-former. The LLM Decoder is constructed with a frozen LLM and a learnable feed-forward layer. Using the LLM as a decoder leverages the LLM's inference capabilities when generating action sequences [4]. In this study, we use OPT-2.7B [19] as an LLM.

V. Experiments

We present the results of numerical experiments to show the efficacy of the proposed method. In particular, we answer the question: Can AVBLIP generate higher quality action sequences from human demonstration videos?

TABLE I: Quantitative comparison and ablation studies. The best scores are in bold.

Method	BLEU1	BLEU2	BLEU4	METEOR	ROUGE	CIDEr
Baseline [2]	22.7	8.5	–	10.4	–	–
Ours (w/o 2nd stage)	33.58	18.48	6.46	14.05	31.7	75.88
Ours (full)	**36.80**	**21.54**	**8.35**	**15.27**	**34.67**	**91.06**

TABLE II: Ablation results on input modalities. In the first column, "I", "V", "A", and "S" denote image, video, audio, and speech modalities, respectively, used for training and evaluation. For example, "IV_S" means that we used image, video, and speech features but did not use audio features.

Modality	BLEU1	BLEU2	BLEU4	METEOR	ROUGE	CIDEr
IVAS	36.80	21.54	8.35	15.27	34.67	91.06
_VAS	36.35	20.80	7.72	14.70	33.08	81.11
I_AS	36.82	21.84	8.33	15.32	34.68	90.91
IV_S	34.08	19.63	7.34	13.72	32.59	76.90
IVA_	**37.31**	**22.19**	**8.51**	**15.51**	**35.18**	**93.80**
IV__	32.19	18.27	6.75	12.72	31.05	70.32

A. Can AVBLIP generate higher quality action sequences from human demonstration videos?

Settings. In the first experiment, we quantitatively evaluated our method. We used a method proposed by Hori *et al.* [2] as a baseline since it is one of the action sequence generation models for manipulator robots and is a method for tasks similar to those focused on in this paper. [2] used Audio-Video Transformer to extract relationships between the multiple modalities. On the other hand, because this method did not use an LLM, it was not possible to generate action sequences based on the common sense knowledge possessed by the LLM. We used the YouCook2 dataset [1], a standard dataset for cooking tasks to evaluate the method, containing cooking demonstration videos. Each video in the dataset consists of 5 to 16 human cooking steps, and the annotators annotated each cooking step with a description such as "grill the tomatoes in a pan". The average sentence length is 8.8 words and the vocabulary size is 1.6K words. The dataset contains 2K videos of 89 recipes, divided into training, validation, and test sets, each 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 trained the proposed method on a GeForce RTX4090. We follow the standard metrics used in the video captioning task, namely, BLEU, METEOR, ROUGE, and CIDEr.

Results. Table I shows the quantitative results of the experiments compared to the baseline method. Here, the scores for the baseline in Table I are those reported in [2]. From the table, the proposed method outperforms the baseline method by 14.10, 13.04, and 4.87 points for BLEU1, BLEU2, and METEOR, respectively.

This is believed to be mainly due to the following two points: The Q-former effectively aligned action sequences and multimodal features. In the LLM Decoder, the inferencing capabilities of the LLM enabled more appropriate prediction and generation of action sequences.

To investigate the effectiveness of the two-step training in the proposed method, an evaluation was conducted at the end of each step. Table I shows these results. The table shows that the CIDEr score for the model that made it to the second step exceeded the CIDEr score at the first stage by 15.18 points. The CIDEr scores of the model up to the second step were 15.18 points higher than the CIDEr scores at the first stage. This suggests that the LLM Decoder, with its extensive knowledge, was able to generate action sequences from the multimodal features appropriately.

Table II shows the results of the ablation study on input modalities, where we removed some of the input modalities in both training and evaluation. Results on one missing modality (_VAS, I_AS, IV_S, and IVA_) show the importance of each modality. Compared to the result of all modalities (IVAS), we can see the image and audio modalities contribute significantly to the performance while the speech modality has a negative impact. This could be because the speech features were extracted from the subtitle text, which was sometimes missing or unrelated to the cooking instruction. However, comparing IV_S with IV__, the speech modality still helps if the audio is missing. We can also see the importance of audio modality from the result of IVA_ and IV__. Thus, the results demonstrate that our method AVBLIP can effectively utilize multimodal features from human demonstration videos to improve robot action generation.

VI. Conclusions

A domestic service robot could significantly benefit future smart homes by aiding the elderly or those with disabilities. Nonetheless, creating robots that can execute these domestic tasks presents considerable challenges. One promising solution is through the use of demonstrations by human experts. Therefore, this paper focused on the task of generating action sequences for manipulators from human cooking demonstration videos. The main contributions of this paper are that we proposed a new framework, AVBLIP, a model for generating action sequences for manipulators from human demonstration videos. We introduce a multimodal encoder to process video, audio, speech, and text features to consider the speech information by the human and the audio information generated by the environment. The proposed method outperformed the proposed method in all standard evaluation metrics.

This paper focuses on the action sequence generation and we have not conducted physical experiments with a single-arm manipulator. Therefore, a future study would be to conduct experiments based on the action sequences generated by the proposed method.

REFERENCES

[1] L. Zhou, C. Xu, and J. Corso, "Towards Automatic Learning of Procedures From Web Instructional Videos," in *AAAI*, 2018, pp. 7590–7598.

[2] 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 *Interspeech*, 2023, pp. 4663–4667.

[3] 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 *ICRA*, 2024.

[4] 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. 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 and not as i say: Grounding language in robotic affordances," in *arXiv preprint arXiv:2204.01691*, 2022.

[5] M. Shirasaka, T. Matsushima, S. Tsunashima, Y. Ikeda, A. Horo, S. Ikoma, C. Tsuji, H. Wada, T. Omija, D. Komukai, *et al.*, "Self-Recovery Prompting: Promptable General Purpose Service Robot System with Foundation Models and Self-Recovery," in *ICRA*, 2024.

[6] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "VIP: Towards Universal Visual Reward and Representation via Value-Implicit Pre-Training," in *ICLR*, 2022.

[7] Y. J. Ma, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman, "LIV: Language-Image Representations and Rewards for Robotic Control," in *ICML*, 2023, pp. 23 301–23 320.

[8] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A Universal Visual Representation for Robot Manipulation," in *CoRL*, 2023, pp. 892–909.

[9] A. Nguyen, D. Kanoulas, L. Muratore, D. Caldwell, and N. Tsagarakis, "Translating Videos to Commands for Robotic Manipulation with Deep Recurrent Neural Networks," in *ICRA*, 2018.

[10] 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 *ICRA*, 2021.

[11] R. Shah, R. Martín-Martín, and Y. Zhu, "MUTEX: Learning Unified Policies from Multimodal Task Specifications," in *CoRL*, 2023.

[12] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, *et al.*, "Ego4D: Around the World in 3,000 Hours of Egocentric Video," in *CVPR*, 2022, pp. 18 995–19 012.

[13] Z. Liu, A. Bahety, and S. Song, "REFLECT: Summarizing Robot Experiences for Failure Explanation and Correction," in *CoRL*, 2023.

[14] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as Policies: Language Model Programs for Embodied Control," in *ICRA*, 2023, pp. 9493–9500.

[15] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, "Tidybot: Personalized Robot Assistance with Large Language Models," *Autonomous Robots*, vol. 47, no. 8, pp. 1087–1102, 2023.

[16] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian, *et al.*, "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances," in *CoRL*, 2023, pp. 287–318.

[17] 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 *ICML*, 2023.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *ECCV*, 2020.

[19] 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.