# Artificial Intelligence in Industry 4.0: The future that comes true: AI

Karabegović, Isak; editor

2024-09-17

https://bastina.anubih.ba/handle/123456789/791

DOI: 10.5644/PI2024.215.05

# Advancements in Robotic Intelligence: The Role of Computer Vision, DRL, Transformers and LLMs

Lejla Banjanović-Mehmedović*[1], Anel Husaković[2], Azra Gurdić Ribić[3], Naser Prljača[1], Isak Karabegović[4]

***Abstract:*** *In recent advancements in robotics, Artificial Intelligence (AI) methods such as Deep Learning, Deep Reinforcement Learning (DRL), Transformers, and Large Language Models (LLMs) have significantly enhanced robotic capabilities. Key AI models driving advancements in robotic vision include Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), the DEtection Transformers (DETR), the YOLO family of algorithms, segmentation techniques, and 3D vision technologies.*
*Deep Reinforcement Learning (DRL), an AI technique where agents learn optimal behaviors through trial and error interactions with their environment, enables robots to perform complex tasks autonomously. Transformers, originally developed for natural language processing, have been adapted to robotics for tasks involving sequence prediction and data understanding, improving perception and decision-making processes. LLMs leverage vast amounts of text data to enhance robot-human interaction, enabling robots to understand and generate human-like language, thus improving their communicative and collaborative abilities in various applications. The integration of these AI methods enhances the adaptability, efficiency, and overall performance of robotic systems, paving the way for more sophisticated and intelligent autonomous agents.*

***Keywords:*** *Artificial Intelligence,Robotics, Computer Vision, Deep Learning, Deep Reinforcement learning, Transformers, Large Language models, Perception, Control, Decision Making, Sequence Prediction*

## 1. Introduction

The field of robotic intelligence has witnessed significant advancements in recent years, driven by the integration of cutting-edge technologies such as Deep Learning (DL), Deep Reinforcement Learning (DRL), Transformers, and Large Language Models (LLMs)[1-4]. These technologies are revolutionizing the way robots learn, perceive, and interact with their environments, leading to unprecedented levels of autonomy and efficiency.

*[1]University of Tuzla, Faculty of Electrical Engineering, Tuzla, Bosnia and Herzegovina
[2]Eacon doo, Zenica, Bosnia and Herzegovina
[3]Foundation for Innovation, Technology and Transfer of Knowledge, Bosnia and Herzegovina
[4]Academy of Sciences and Arts of Bosnia and Herzegovina, Sarajevo,Bosnia and Herzegovina
E-mail: lejla.mehmedovic@untz.ba, naser.prljaca@untz.ba, anel@eacon.ba, azra.gurdic@fet.ba, isak1910@hotmail.com

Deep Reinforcement Learning (DRL) has emerged as a pivotal technique in robotic learning, combining the principles of reinforcement learning with the powerful representation capabilities of deep neural networks. DRL allows robots to learn optimal behaviors by interacting with their environments and receiving feedback in the form of rewards or penalties, facilitating the trial-and-error learning process.This approach has proven particularly effective in complex and dynamic tasks, such as robotic manipulation, navigation, and autonomous.

Originally developed for natural language processing (NLP), transformers have been adapted for various applications in robotics because of their capability to handle sequential data and understand long-range dependencies. The self-attention mechanism in transformers enables robots to concentrate on significant elements of their sensory inputs, making them highly suitable for tasks requiring the integration of multimodal information. This capability enhances robots' understanding and decision-making processes, enabling more sophisticated and context-aware interactions with their environments driving [5-8].

Large Language Models (LLMs) are transforming the interface between humans and robots by providing advanced natural language understanding and generation capabilities. LLMs enable robots to comprehend and execute complex instructions given in natural language, allowing for more intuitive and flexible human-robot interactions [4,9-10]. These models can interpret context, generate detailed action plans, and integrate information from multiple sources, significantly improving the robots ability to perform intricate tasks autonomously.

## 2. Fundamental Modules in Robotics

In robotics, fundamental capabilities include *perception, control, decision-making, planning and interaction.*

***Perception*** is a crucial skill for robots, involving the use of sensors to collect environmental data. This capability enables robots to recognize objects, interpret visual information, and understan spatial relationships [11]. By leveraging advanced algorithms and AI models, computer vision enables robots to process visual information, identify objects, understand scenes, and make informed decisions. Techniques such as Convolutional Neural Networks (CNNs) [12] and Vision Transformers (ViTs) [6] have revolutionized this field, providing robust and accurate solutions for tasks like image recognition, object detection, and scene segmentation. As a result, robots equipped with sophisticated vision capabilities are increasingly capable of performing complex tasks in dynamic and unstructured environments [13], from industrial automation and autonomous driving to healthcare and service robotics [14].

Transformers have transformed robotic perception by improving the ability to process and interpret sensory data. Initially designed for natural language processing, transformers have been adapted for vision tasks [13], such as object

detection and scene understanding [15-16], through Vision Transformers (ViTs). ViTs can handle large amounts of visual data and extract meaningful features, which is crucial for robots to recognize and respond to their surroundings effectively. This advanced perception allows robots to perform tasks such as autonomous navigation and object manipulation with greater accuracy and reliability [6].
A variety of multi-modal models have been developed to tackle tasks like visual question answering, image captioning, and speech recognition [17].The related technologies for perception used in robotics are Large Vision Models (LVMs), that can "see" clearly, but are "blind" in reasoning [18], Multimodal Large Language Model (MLLM), where LLMs and LVMs work together for discriminative and generative tasks and Vision-language-action (VLA) model, where pre-existing vision-language models, are trained without any new parameters to output text-encoded robot actions [19].
The control module is vital for managing robotic actions. Reinforcement Learning (RL) significantly contributes to robot control by enabling robots to determine optimal actions through trial and error.Deep Reinforcement Learning (DRL) algorithms allow robots to develop control policies that maximize cumulative rewards over time. This learning process is particularly beneficial for tasks requiring precise motor control and adaptation to changing environments. DRL has been successfully applied in robotic arms for tasks like picking and placing objects, where the robot learns to control its movements to achieve the desired outcomes efficiently [20]. The paper explores enhancing human-robot team performance by using Q-learning algorithms to adjust task loads based on real-time physiological data analysis [21].

***The planning module*** in robotics is essential for enabling autonomous robots to make decisions and execute tasks effectively [22]. It involves the creation of strategies or sequences of actions that allow robots to achieve specific goals while navigating their environments [23]. Key components of robotic planning include *path planning, motion planning, and task planning.* The paper [24] presents the use of AI methods such as machine learning algorithms, heuristic search techniques, and optimization algorithms to enhance the efficiency and effectiveness of autonomous disinfection robots in navigating complex environments to combat the spread of COVID-19. The application of the Grey Wolf Optimization algorithm is presented in the paper [25] to significantly enhance the efficiency and effectiveness of service robot path planning, improving their operational performance in various environments.

Large Language Models (LLMs), like GPT-4, have advanced robotic planning by enabling robots to understand and execute complex instructions given in natural language. LLMs can translate high-level commands into detailed action sequences, making it easier for robots to plan and perform tasks autonomously.

This capability is enhanced when combined with visual input, allowing LLMs to create visual-semantic plans that integrate information from multiple modalities. Such integration improves the robots' ability to generate precise and context-aware plans, essential for tasks like autonomous navigation and complex manipulations. [26-27].

***The decision-making module*** is a key component of robots, allowing them to make informed choices and plan tasks based on their current state and environment. As the core of a robot, decision-making integrates information from the perception module to produce suitable actions, bridging the gap between previous inputs and future responses. Sequence Precision involves executing tasks in a specific order with high accuracy, crucial for applications like assembly lines and surgical robots where precise timing and order are essential for success. Multimodal Decision Making involves for example LLMs, which can integrate information from various sources (e.g., visual, auditory, and textual data) to make context-aware decisions. This integration allows robots to generate detailed and accurate action sequences based on comprehensive situational understanding [4, 27].

***Interaction*** is a crucial module that allows robots to engage with both their environment and humans. To improve their capability to interact effectively in the physical world, robots typically undergo extensive training [28-30].

The use of LLMs has also significantly improved human-robot interaction. LLMs' advanced natural language understanding allows robots to engage in more intuitive and meaningful conversations with humans. This ability to comprehend and respond to nuanced language makes robots more user-friendly and capable of performing tasks based on verbal instructions. Furthermore, the integration of LLMs with other AI technologies enables robots to provide personalized and context-aware interactions, enhancing their utility in various applications, from customer service to personal assistance [4].

## 3. AI Transform Robotics

AI is revolutionizing robotics by allowing robots to perform complex tasks with enhanced autonomy, accuracy, and flexibility [31]. Machine learning (ML) algorithms enable robots to learn from data and improve their performance over time, while computer vision systems provide them with the ability to perceive and understand their environment.

Natural language processing (NLP) helps robots communicate with humans more effectively, enhancing their usability in diverse applications. This transformation is paving the way for a future where intelligent robots seamlessly integrate into everyday life, providing innovative solutions and enhancing productivity across various sectors.

### 3.1. Advancements in Computer Vision in Robotics Based on AI Models

Convolutional Neural Networks (CNNs) have played a crucial role in the advancement of image recognition [32]. These networks are built to automatically and adaptively learn the spatial hierarchies of features from input images, which makes them extremely effective for tasks like image classification and object detection. CNNs are widely used in robotics for tasks like sorting objects, quality control in manufacturing, and facial recognition in security systems [33].

Vision Transformers (ViTs) represent a newer approach, leveraging transformer architectures that have proven highly effective in capturing long-range dependencies in data. ViTs have demonstrated superior performance in various image recognition tasks and are increasingly being adopted in robotics for applications requiring high accuracy and efficiency [16].

The DEtection TRansformer (DETR) is a novel approach that uses transformers for object detection. It has shown promising results in accurately detecting and recognizing objects in images, making it suitable for complex robotics applications [16].

The YOLO (You Only Look Once) family of algorithms has revolutionized real-time object detection and recognition [34]. YOLO models divide images into grids and predict bounding boxes and class probabilities for objects within these grids simultaneously, allowing for rapid and accurate object detection. This capability is crucial in robotics for tasks such as autonomous navigation [35], where robots need to detect and respond to objects in their environment in real time.

Segmentation involves partitioning an image into multiple segments to simplify its representation and make it more meaningful. Segmentation techniques, including Fully Convolutional Networks (FCNs), U-Net, Mask_R-CNN which provide detailed pixel-wise classification, essential for applications such as robotic surgery and automated inspection systems [36-37] .

3D vision technologies enable robots to perceive and interact with their environment in three dimensions. This is achieved through techniques such as stereo vision, depth sensors, and LiDAR. In the realm of 3D vision, algorithms like PointNet and VoxelNet [38-39] process 3D point clouds to enable precise object detection and spatial understanding, facilitating tasks such as robotic grasping, 3D mapping, and autonomous driving.

### 3.2. Reinforcement Learning Algorithms

Reinforcement Learning (RL) aims to solve the sequential decision-making problem by taking actions that maximize expected rewards. Essentially, the agent follows a policy to make a series of decisions (i.e., take actions) in different states of the environment [40-41]. The sequence of these states and

actions creates a trajectory. Each decision within this policy is evaluated based on the accumulated rewards over the trajectory to determine the policy's effectiveness. By evaluating these trajectories, the agent improves the policy by increasing the likelihood of decisions that yield higher expected rewards. Through this iterative process of trial and error, the agent continually refines the policy until it reaches an optimal state.

To improve the policy, we utilize the Bellman optimality equations to update the value functions by choosing the action that provides the highest value, rather than considering all possible actions.

Various RL techniques have been proposed from different perspectives to optimize the policy. A detailed classification of RL algorithms is presented in Figure 1.

(Deep) Reinforcement Learning (RL) can be divided into model-free and model-based algorithms [41,42]. The main distinction between these categories lies in whether the agent has access to a model of the environment, specifically the transition function and the reward function.*Model-based algorithms*, like AlphaZero, obtain or learn a model of the environment to predict future values or states.

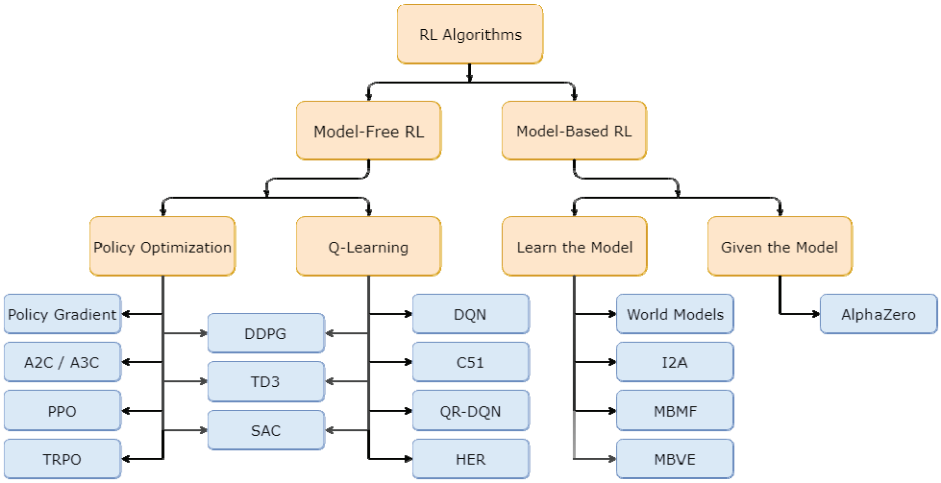


*Figure 1. Classification of Reinforcement Learning [42]*

In many RL scenarios, the reward and transition functions are unknown due to the environment's complexity and intricate mechanisms. As a result, agents often use *model-free methods*, learning the policy solely from the experience gained through interactions with the environment.

In its early stages, reinforcement learning (RL) was suited to scenarios with discrete and limited state and action spaces, enabling agents to record information in tables. However, modern tasks like playing Go or autonomous

driving involve large discrete state-action spaces or continuous values, rendering table-based methods impractical. To overcome this challenge, Deep Reinforcement Learning (DRL) integrates RL and deep learning (DL), with RL setting the problem and optimization objectives, while DL models the policy and expected rewards. Depending on the role of deep neural networks (DNN) in DRL, it can be classified into three categories [43].

*In value-based methods*, deep neural networks (DNN) learn a value function for discrete action spaces to assess each potential action. DNN do not participate in policy decision-making; instead, they are used to estimate the policy's performance. Figure 2 presents the schematic of the DQN model [44].

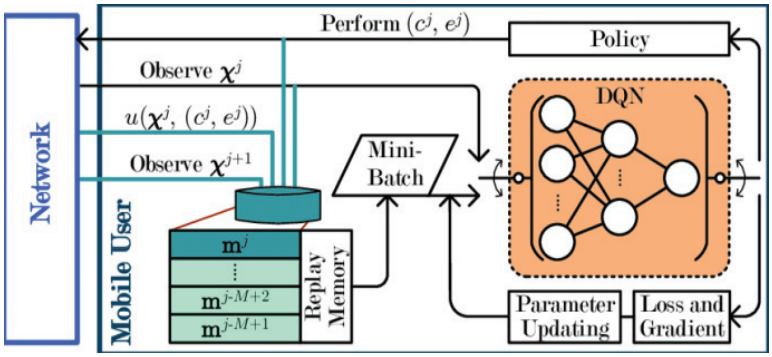


*Figure 2. Schematics of the DQN model [44]*

*In policy-based methods*, deep neural networks (DNN) are directly involved in selecting actions.Compared to value-based methods, policy-based algorithms generally offer better convergence and can learn stochastic policies, whereas value-based methods select actions deterministically based on the maximum Q value. *Policy-based algorithms*, like *Proximal Policy Optimization (PPO)*, offer a continuous action space and aim to directly map states to actions by developing a representation of the actual behavior policy. Figure 3 illustrates the network structure utilized for the PPO [45].

Value-based methods tend to be less stable and exhibit poorer convergence compared to policy-based approaches, as they rely on approximating the Q-function. However, value-based methods are more sample-efficient. On the other hand, policy-based methods are more prone to getting stuck in local optima due to the large search space.

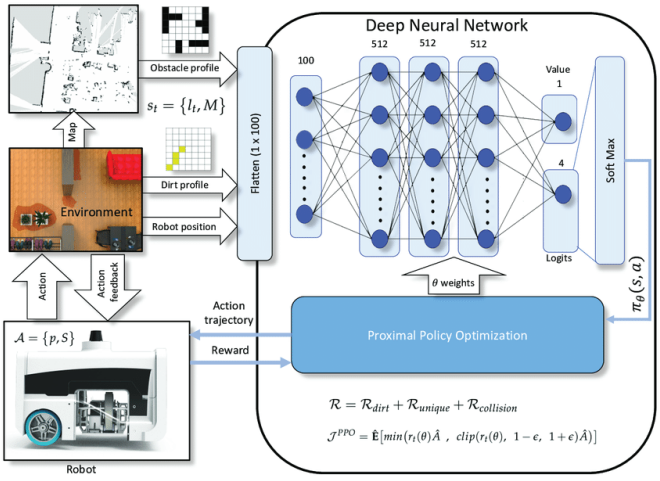


*Figure 3. Schematic of the PPO model [45]*

*The hybrid actor-critic model* integrates both approaches, using two distinct deep neural networks (DNNs) named the actor and the critic. During each training iteration, the actor evaluates the current state and the policy to decide on an action. The environment then transitions to a new state and provides a reward. The critic updates its parameters based on this feedback and rates the actor's action. Subsequently, the actor adjusts its policy based on the critic's evaluation. Recent examples of algorithms that employ this model include *Deep Deterministic Policy Gradient (DDPG)* and *Asynchronous Advantage Actor-Critic (A3C)*, as depicted in Figure 4 [43].

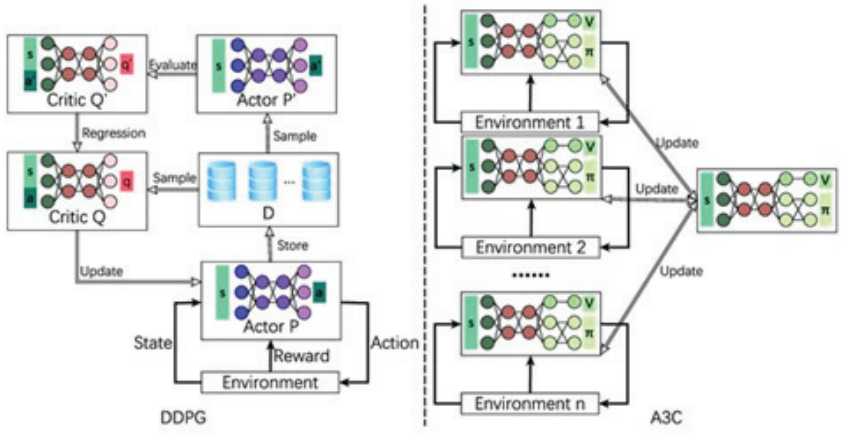


*Figure 4. Schematic of the DPPG and A3C models. [43]*

DDPG incorporates the concepts of target deep neural networks (DNNs) and experience replay memory [46,47]. Despite these improvements, DDPG's performance is hindered by the operation of experience replay. This limitation can be overcome by *asynchronous* training approachin A3C (Asynchronous Advantage Actor-Critic) [48]. Namely, multiple replicas of the global network interact with their respective environments. Each local deep neural network (DNN), which includes both the actor and the critic, is trained independently(as shown on the right side of Figure 4).During training, the local DNNs do not update their own parameters directly but instead modify the global model. The local models synchronize with the global DNN after several steps. This setup allows the global network to be refined through the independent, concurrent training of local models, thereby accelerating the training process.

A3C's multithreaded implementation also greatly boosts learning efficiency. However, A3C struggles in complex environments due to its fixed learning rate, resulting in less robust performance.

To tackle this problem, *DPPO (Distributed Proximal Policy Optimization)* was introduced [49], as shown in Figure 5. DPPO incorporates a penalty term, which mitigates the impact of an inappropriate learning rate by ensuring a more balanced update proportion.

The algorithm features a global network alongside multiple local networks and employs a centralized learning approach combined with decentralized execution. The global network updates the actor and critic parameters, while the local networks gather sample data. During training, the local networks interact independently and simultaneously with the environment based on the global network's strategy.

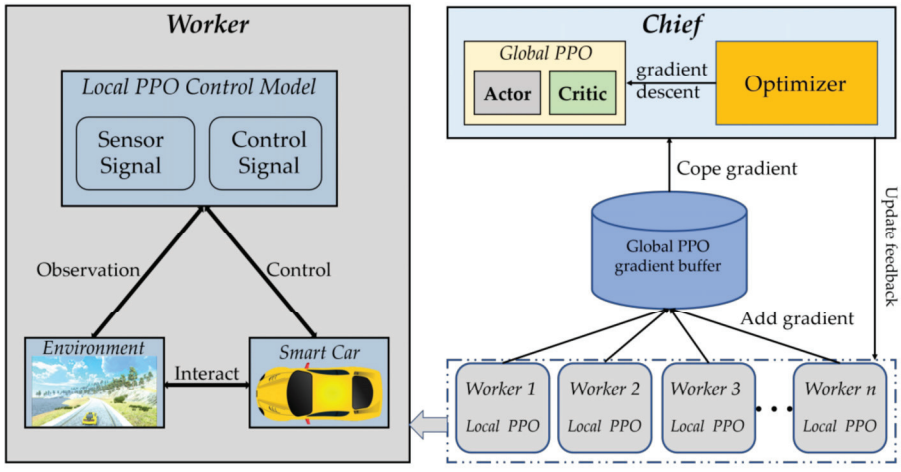


*Figure 5. Schematic of the DPPO model [49]*

They collect data, compute the policy gradient and store it in a shared gradient area. Once this shared area accumulates enough data, the global network retrieves gradient information for learning and updates its parameters. Subsequently, the updated policy parameters are shared with the local networks, which continue to gather data in their respective environments and repeat the process until the maximum number of training steps is reached. This setup involves multiple agents in parallel training within the simulation environment, allowing them to communicate and exchange feedback, thus significantly enhancing training efficiency and addressing issues like slow convergence.

## 3.3. Large Language Models and their Applications in Robotics

### 3.3.1 The working principle of LLM

A Large Language Model (LLM) is a foundational system designed to comprehend, interpret, and generate text in human language [50]. It accomplishes this by analyzing datasets and recognizing patterns and grammatical structures within the data to create text in a conversational manner. LLMs consist of numerous layers and contain millions or even billions of parameters. They are trained on extensive amounts of data, which allows them to understand complex relationships between words and predict the next word in a sentence.

These models utilize self-supervised learning, continuously processing the data until achieving a high level of accuracy. The performance of a language model is heavily influenced by the quality of its training data. The specific steps of the LLM model are shown in Figure 6.

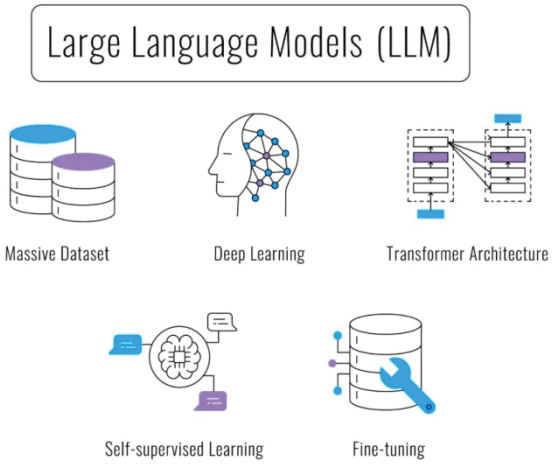


*Figure 6. The custom steps of the LLM structure [50]*

The first step in creating a large language model is to determine the type of LLM you intend to develop. This process starts with gathering a vast and diverse dataset of text from multiple sources, which serves as the foundation for training the model.
After collection, the text data undergoes *preprocessing*, which involves tasks such as *tokenization* (breaking text into words or subwords), converting text to lowercase, removing punctuation, and encoding the text into numerical formats suitable for machine learning. During this phase, each token (word or subword) is converted into a vector representation known as an *embedding*. Embeddings capture the semantic information of words, allowing the model to understand and learn the relationships between them.
Large Language Models generally use neural network architectures known as *transformer architectures*, which employ self-attention mechanisms to capture relationships between words irrespective of their positions in the input sequence. Since transformer architectures do not inherently account for the order of words, *positional encodings* are added to provide information about each token's position in the sequence, enabling the model to understand the sequential structure of the text.
*Training* a large language model (LLM) involves feeding sequences of tokens into the model and adjusting its parameters to minimize the difference between predicted and actual next tokens. These models are trained on extensive datasets over numerous epochs, progressively enhancing their performance. After the initial training phase, *fine-tuning* may be performed on more specific tasks or domains to customize the model for particular applications.

### 3.3.2 Examples of LLMs

To date, numerous foundational models or LLMs have been developed. Notable examples include BERT, Roberta, GPT-3, GPT-4, LlaMA, OPT (an open-source pre-trained transformer language model from Meta), Falcon 2 (an open-source LLM and VLM), Bloom (an open-source LLM from BigScience), and Mistral-7B. These models represent significant advancements in the field of artificial intelligence (AI) in recent years.

- *GPT*– OpenAI's Generative Pretrained Transformer (GPT) is perhaps the most renowned large language model (LLM). Among these, GPT-3 stands out with its 175 billion parameters, enabling it to generate coherent and contextually relevant text across a wide range of domains, including translation, question-answering, and cloze tasks. Additionally, it excels in tasks requiring real-time reasoning or domain adaptation, such as unscrambling words, using novel words in sentences, or performing three-digit arithmetic [52]. GPT-4(V), released on March 14, 2023, represents a significant advancement in the evolution of language

models. A notable innovation in GPT-4 is its multimodal capabilities, allowing it to process images as input (GPT-4V) and generate detailed descriptions, classifications, and analyses across various media types. This multimodal functionality expands the model's versatility and enhances its ability to understand and create content across diverse formats. GPT-4 Turbo and GPT-4(V) extended the context window from 32K to 128K. The latest OpenAI version GPT-4o (o for omni) processes and generates output across text, audio and image modalities in real time.

- *BERT* (Bidirectional Encoder Representations from Transformers) – Developed by Google, BERT introduced a novel approach of pre-training and fine-tuning for language understanding tasks. This innovation has resulted in exceptional performance in applications such as question answering and text classification.
- *T5* (The Text-to-Text Transfer Transformer) – Also developed by Google, T5 is a versatile language model that can be fine-tuned for a wide array of natural language processing tasks, including summarization, translation, and text generation.
- *LaMDA* – Developed by Google, LaMDA powers Google's conversational chatbot, Bard, enhancing its ability to engage in more natural and meaningful conversations.
- *PaLM-2* (successor of LaMDA), LLM designed to handle complex language tasks across different languages and contexts. PaLM is known for its large scale, extensive training data, and ability to perform a variety of NLP tasks. It is used on Microsoft BARD Chat bot.
- *LLaMA*– Developed by Meta, LLaMA is a generative text model pre-trained and fine-tuned with parameter counts ranging from 7 to 70 billion. It eliminates the absolute position embedding and introduces rotational position embedding at each layer of the network. The version, LLaMA 3, was integrated into Meta AI in 2024.

Applying large language models to the field of robotics has significant research implications and practical value [53]. By utilizing pre-trained language models, robots can better comprehend user intentions and needs. Additionally, other research focuses on employing LLMs for natural language generation in robots. From various perspectives, large language models-based robotics is one of the most promising pathways to achieving embodied intelligence in the future.

However, integrating LLMs with robotics presents several challenges. Training and deploying LLMs require significant computing resources and data, which can be a limitation for resource-constrained robotic platforms.
Generally, the applications of LLM in robotics hold tremendous potential. LLM models provide new paradigms and methods for robot control, perception, decision-making, and path planning. Below is a summary of LLM-based robotics [53]:

- *PaLM-SayCan* is an advanced system developed by Google, introduced in 2022. PaLM-SayCan is a sophisticated system that combines the Pathways Language Model (PaLM) with the SayCan framework to enable robots to execute tasks based on natural language instructions. It exemplifies the potential of LLMs to enhance human-robot interactions by translating complex language inputs into actionable tasks, supported by a value function to guide decision-making.

- *PaLM-E* (Pathways Language Model for Embodied AI) [54], introduced in 2023, is an *advanced multimodal AI* model developed by Google that integrates language (PaLM) and vision (ViT) capabilities for embodied AI applications. By extending the PaLM model's language processing abilities into the realm of robotics, PaLM-E enables sophisticated interaction with the physical world through natural language and visual inputs. It can perform tasks such as object manipulation, navigation, and interactive responses based on both visual and textual data.Largest model (2024) is PaLM-E-526B with 526B parameters.

- *PaLI-X* is an *advanced multimodal model* developed by Google, introduced in 2023. It represents a significance in the integration of language and vision capabilities, expanding the scope of tasks that large language models (LLMs) can perform across different modalities. By combining advanced language and vision capabilities, PaLI-X enables new types of interactions between humans and machines, and supports a wide range of practical applications.

LLM and LVM (Large Vision Model) run towards each other, leading to the new field of Multimodal Large Language Model (MLLM). Formally, it refers to the LLM-based model with the ability to receive, reason, and output with multimodal information [18]. Figure 7. shows the MLLM timeline.

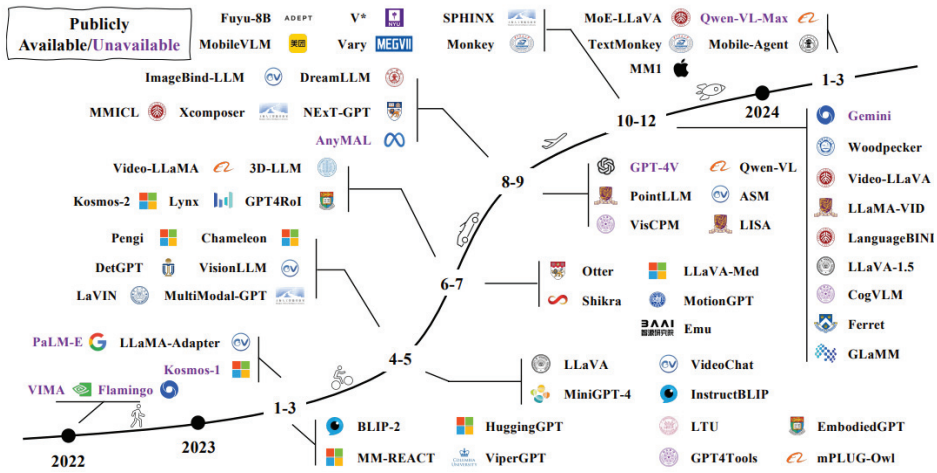


*Figure 7. A timeline of representative MLLMs [18]*

## 3.4. Transformer Models: Foundations and Their Role in Robotics

### 3.4.1 Transformer Models Architecture

Natural Language Processing (NLP) tasks can be categorized into discriminative tasks, and generative tasks. *Discriminative tasks* are used for sentiment analysis, text classification, for example if the sentence is given, one can classify the sentiment of the sentence (like positive/negative). *Generative tasks* are used for language modeling, machine translation, summarization, like where one needs to predict the next word based on the input sentence (context).

In the pre-transformer era (before 2017 [5]), standard ways to solve those tasks were Recurrent Neural Networks (RNNs), Long short-term memory (LSTM), and Convolutional neural networks(CNNs).

Each of those models have some problems. Problems faced with RNNs and LTMs are dependency across hidden layers, where the current hidden state in the network depends on the previous hidden states and hard to model long-term relationships, because languages may not have locality, unlike images where it takes $O$(sequence length) steps to model the interaction between two tokens. This limits the training parallelism.

For CNNs there are no dependency problems between tokens, which leads to better scalability, but there is limited context information, because of fixed window size of convolution that results in worse modeling capability.

Transformers are a type of model introduced by Vaswani et al. in their paper titled "Attention Is All You Need." [5]. They marked a significant shift from previous architectures like RNNs and LSTMs, which were previously dominant in sequence modeling tasks. Transformer models have introduced significant

advancements in language modeling, overcoming previous challenges with several key features:

- *Attention Mechanisms:*Transformers employ self-attention mechanisms to evaluate the significance of each word in a sentence in relation to the others.
- *Parallelization:* Unlike RNNs and LSTMs, transformers process all words in a sequence simultaneously, allowing for faster training and better scalability.
- *Scalability:* Transformers can be scaled up with more layers and parameters to handle complex tasks and large datasets.

The transformer architecture consists of two main parts: the *encoder* and the *decoder,* Figure 7. The encoder, situated on the left side of the Transformer architecture, is responsible for mapping an input sequence to a series of continuous representations, which are subsequently passed to the decoder. The decoder, located on the right side of the architecture, generates an output sequence by using the encoder's output in conjunction with the decoder's output from the previous time step.Both components are composed of multiple layers, and each layer contains specific subcomponents.

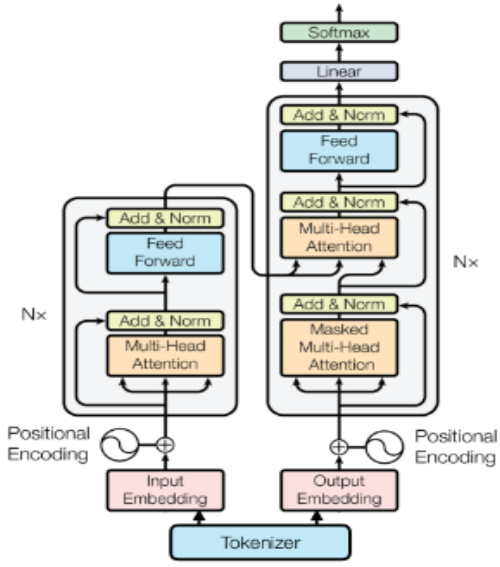


*Figure 8. The encoder-decoder structure of the Transformer architecture [5]*

The self-attention mechanism is crucial for a transformer's capacity to process sequences. Multiple self-attention heads operate in parallel, each capturing different aspects of the relationships between words. Since transformers do not inherently recognize the order of tokens, *positional encodings* are incorporated to convey information about the positions of tokens within the sequence.

### 3.4.2 Key Applications of Transformer Models in Robotics

Transformer models in robotics enhance perception, decision-making, and interaction capabilities, enabling robots to perform complex tasks with greater efficiency and accuracy.

Transformers improve a robot's ability to perceive and understand its environment:

*Visual Recognition***:** Vision transformers (ViTs) apply transformer architectures to image data, enhancing object detection, classification and scene understanding.

*Multimodal Perception***:** Models like PaLM-SayCan integrate visual and textual data, allowing robots to interpret and respond to visual cues combined with natural language instructions.

Transformers enhance the planning and navigation capabilities of robots:

- *Path Planning:* Transformers can process image data to help robots plan their movements by predicting the outcomes of different action sequences. This capability is crucial for autonomous robots that need to navigate dynamic environments.
- *Task Planning:* By understanding high-level goals and breaking them down into actionable steps, transformers assist robots in efficiently completing tasks.
- *Motion Prediction:* By analyzing sequences of past actions, transformers enable robots to anticipate future states and make informed decisions about their next moves.

**The Control Transformer (CT)** is a Transformer framework designed to model conditional sequences produced by robot actions [5,53]. It approaches the CT problem as a sequence modeling challenge with a goal-oriented perspective, allowing for learning from data collections obtained through sampling. Essentially, CT processes involve auto-regressively predicting actions within a sequence.

**Decision Transformer (DT)** for offline RL takes a sequence of returns, observations, and actions as input and outputs action predictions [55].

**Trajectory Transformer (TT)** also models the trajectory as a sequence of states, actions and rewards, while discretizing each dimension of state/actions [AH55].

**Generalized DT** unifies a family of algorithms for future information matching using transformers [55]. Transformers can also be used as world models for model-based RL [55,56].

**The Q-Transformer** [57] is designed to combine offline reinforcement learning with the Transformer architecture [5], facilitating the use of Q-values for each action dimension. This method involves discretizing each action dimension and representing them as distinct tokens using Q-values. By leveraging large and

diverse robot datasets, this approach improves the efficiency and effectiveness of the reinforcement learning process. Tested on real world experiments with dataset of 58k demonstrations on more than 700 tasks, with a fleet of 13 robots.

**Robotics Transformer 1** (RT-1) [53,58] is designed to encode high-dimensional input and output data, such as images and instructions, into compact tokens that can be efficiently handled by Transformer [5]. It features real-time operation capabilities, making it well-suited for applications that demand quick processing and response times.

RT-1 showed impressive generalization capabilities in experimental evaluations. Its architecture includes FiLM, a conditioned EfficientNet, a TokenLearner, and a Transformer. Tested on 700+tasks (130k demonstrations/episodes), with fleet of 13 robots over 17 months, 35M parameters and compared with BC-Z and Gato (Generalist agent) models. Despite its advanced features, RT-1 is not an end-to-end model.

**Robot Transformer 2 (RT-2)** [59] is a model that utilizes the fine-tuning of a vision-language model (VLM) to develop an end-to-end system capable of directly mapping robot observations to actions. Research demonstrated how text-encoded 6 DoF actions from vision-language alignments (VLAs) are integrated into the robot's closed-loop control.

RT-2 is trained on a web-scale dataset to enable generalization and semantic awareness for new tasks. Specifically, it uses the WebLI dataset, which includes 1 billion image-text pairs across 109 languages and incorporates low-level action-related text tokens such as Cartesian end-effector commands. This model can be classified as a visual-language-action model (VLA) [59]. The largest version, RT-2-PaLI-X-55B, has 55 billion parameters and operates at 1-3 Hz.

**Robotics Transformer X (RT-X)** [60] is divided into two branches: RT-1-X and RT-2-X. RT-1-X utilizes the RT-1 architecture and is trained using the X-embodiment repository, while RT-2-X builds on the strategy architecture of RT-2 and is trained on the same dataset. Experiments have shown that both RT-1-X and RT-2-X demonstrate improved capabilities.

There are some notable examples of transformers in robotics:

- **PaLM-SayCan.** PaLM-SayCan integrates transformer models to allow robots to process natural language instructions and perform corresponding physical tasks.
- **LM-Nav, 2023:** Developed to enhance communication between users and robots using language, the LM-Nav system includes three key components: a vision-navigation model (VNM), a vision language model (VLM), and a large language model (LLM).
- **Expedition A1**, 2023: Developed by AGIBot, Expedition A1 showcases the company's commitment to integrating advanced AI into robotics and fostering seamless collaboration between humans and machines.

## 3.5. Comparison of Advanced Transformers Models in Robotics

The rapid evolution of artificial intelligence has profoundly impacted robotics, with transformer models emerging as a transformative force in the field.
This comparative analysis, presented in Table 1, explores the application of advanced transformer models in robotics, focusing on their performance, efficiency, and suitability for diverse robotic tasks.

***Table 1.*** *Comparison of Advanced Transformer Models in Robotics*

| Model | Developer | Release Year | Core Functionality | Primary Focus | Key Features | Applications |
|---|---|---|---|---|---|---|
| **PaLM-SayCan** | Google | 2022 | Combines NLP with robotic control | Task execution based on natural language | Multimodal capabilities, semantic understanding, physical embodiment | Robots interpreting and executing human commands in real-world environments |
| **PaLM-E** | Google | 2023 | Multimodal integration for robotic tasks | Enhanced perception and versatility | Enhanced perception, integrated embeddings, versatility | Situationally aware robots capable of complex and varied tasks |
| **PaLM-X** | Google | 2024 | Cross-modal language model | Multimodal understanding and generation | Advanced cross-modal integration, extensive knowledge base | General-purpose multimodal understanding and task execution |
| **Q-transformer** | OpenAI | 2023 | Integrates Q-learning with transformer architecture | Reinforcement learning and decision making | Combines Q-learning with transformers, optimized for sequential decision-making tasks | Complex decision-making in dynamic environments |
| **Decision Transformer** | MIT | 2021 | Combines sequence modeling with reinforcement learning | Reinforcement learning | Uses transformers for RL, sequence modeling, flexible policy generation | Decision-making in complex tasks, reinforcement learning |
| **RT-1** | Google DeepMind | 2022 | Real-time learning and adaptation for robotic control | Real-time learning | Reinforcement learning, scalable architecture, efficient training | Autonomous robots learning and adapting in real-time |
| **RT-2** | Google DeepMind | 2023 | Utilizes reinforcement learning within a transformer framework | Efficient learning and control | Reinforcement learning, scalable architecture, optimized for quick adaptation | Enhanced robotic performance in dynamic environments |
| **RT-X** | Google DeepMind | 2024 | Next-generation robotics transformer | Advanced real-time learning and control | Cutting-edge transformer architecture, superior reinforcement learning capabilities | Sophisticated robotic applications requiring high adaptability |

| | | | | | | |
|---|---|---|---|---|---|---|
| **ViT (Vision Transformer)** | Google | 2020 | Image classification and processing | Visual tasks and perception | Transformer architecture applied to visual data, attention mechanisms | Image recognition, object detection, visual understanding |

## 4. Examples of AI Methods in Robotics

AI methods in robotics encompass a range of techniques, from machine learning to reinforcement learning and natural language processing. Transformers and LLMs have revolutionized the field of robotics by enhancing robots capabilities in understanding, processing, and generating human-like language.
Below are some notable examples of how deep learning, reinforcement learning, transformers and LLMs are being applied in robotics to improve perception, interaction, and decision-making.

### 4.1. Computer Vision in Applications

Convolutional Neural Networks (CNNs) have transformed image recognition by learning spatial hierarchies from input images, making them ideal for tasks such as image classification and object detection.They are widely used in robotics for applications such as sorting and quality control. Vision Transformers (ViTs) and the DEtection TRansformer (DETR) represent newer approaches, with ViTs excelling in capturing long-range dependencies and DETR showing promising results in object detection. YOLO algorithms enable real-time object detection essential for autonomous navigation, while segmentation techniques offer detailed pixel-wise classification, crucial for applications like robotic surgery.

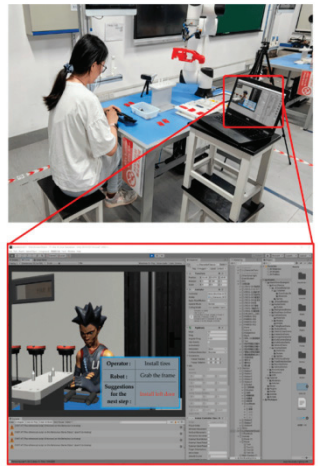

*Figure 9. Computer vision based on CNN neural network in robotics [61]*

Example of CNN-based computer vision used in human-robot collaboration in assembly tasks is presented in Figure 9 [61], while computer vision based on VLA model is presented in Figure 10 [59].

*Figure 10. Computer vision based on VLA model in robotics [59]*

4.2. DRL in Applications

Deep Reinforcement Learning (DRL) has been increasingly utilized to enhance human-robot collaboration across various industrial applications: *pick-and-place tasks, safe interaction in industrial settings, dynamic task allocation, assembly line collaboration, etc.* [62].

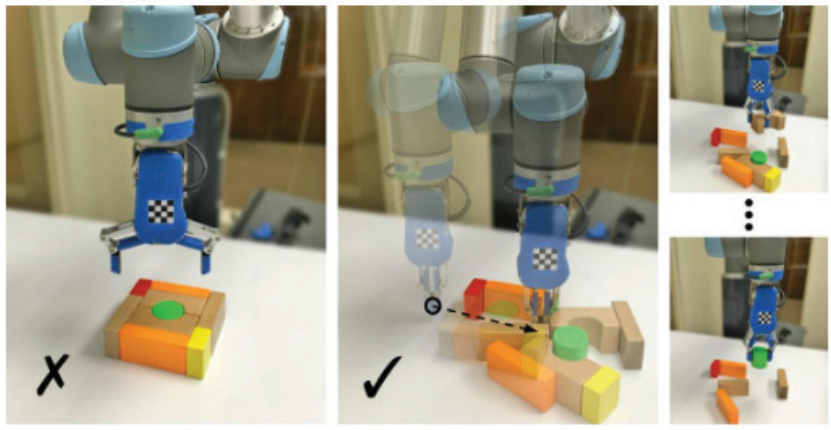


*Figure 11. Pick and place tasks [63]*

DRL is used to control collaborative robots (cobots) in pick-and-place tasks, allowing them to adapt to dynamic environments and handle objects that were not part of their initial training. This is achieved by integrating vision systems and DRL algorithms to improve object recognition and manipulation capabilities As seen in Figure 11, KUKA developed, together with Roboception, a 3D vision system to efficiently perform bin picking using artificial intelligence [63].

In manufacturing, DRL algorithms optimize robot movements to ensure safe and efficient interaction with human workers. This involves path planning and collision avoidance, leveraging DRL to dynamically adjust the robot's actions based on the worker's movements and task requirements. With the use of DRL methods, robots are able to track the movement of humans and shape their path, in order to avoid any collisions, as seen in Figure 12 [64].

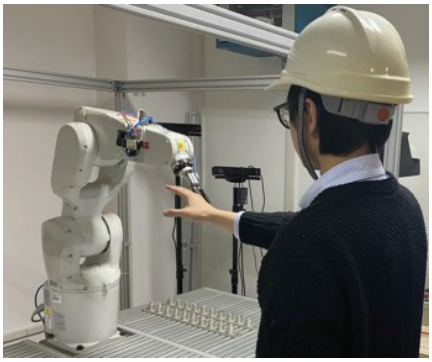

*Figure 12. Safe interaction in industrial settings [64]*

DRL helps in dynamically allocating tasks between humans and robots, improving efficiency and reducing the cognitive load on human workers. For instance, in flexible manufacturing systems, DRL algorithms can learn optimal task distributions based on real-time data, enhancing the overall workflow [62].
DRL is used in assembly line settings where robots and humans work side by side. Robots equipped with DRL can learn to perform precise tasks such as fastening or welding, adapting their actions based on the real-time feedback from human coworkers and environmental changes. When DRL is used with robots in welding processes, the quality of the seam can be quite improved due to the ability to adapt the welding parameters throughout the process, as seen in Figure 13, as well as enhance safety due to a number of sensors able to respond to human presence [65].

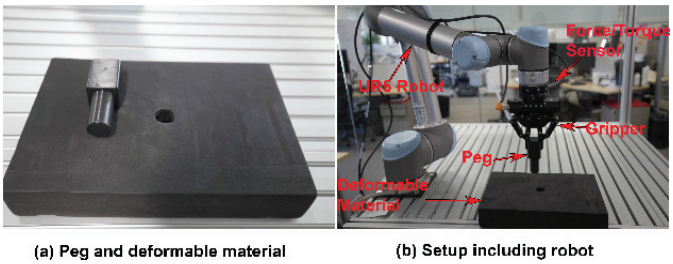


*Figure 13. Assembly line collaboration [65]*

These applications highlight how DRL can significantly enhance the adaptability, safety, and efficiency of human-robot collaboration, leading to more integrated and productive work environments.

### 4.3. Transformers in Applications

Traditional neural networks have shown limitations due to their lack of flexibility in executing actions based on inputs that do not match the training parameters, meaning they struggle to adapt to both minor and major changes in the system. Transformer models have revolutionized addressing this issue, primarily in the field of Natural Language Processing (NLP), where they have demonstrated excellent performance. Increasingly, transformer models are also being developed to solve various problems in robotics.

*Task execution based on inputs.* Transformer models combine computer vision and natural language processing to be used in a form where the robotic system performs specific tasks based on an input, usually in the form of a question. The robot takes the questions expressed in natural language and based on the images it gets, processes the task and performs the correct output [58]. Figure 14 shows such implementation, where the robotic arm using the RT-1 transformer model executes an instruction given by the operator.

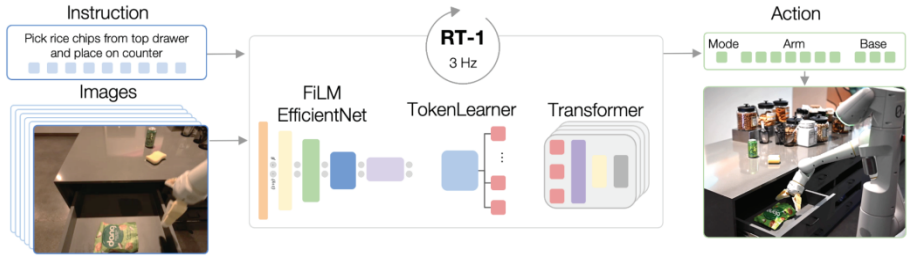


*Figure 14. Task Execution based on input [58]*

*Motion Planning.* Previously used methods in motion planning, when used in goal-targeted tasks, usually show inefficiency if they lack the time or pretraining. Transformer models are well-suited for solving planning tasks due to their capability to make long-horizon connections [66]. Additionally, they take large planning spaces and split them into discrete sets and carefully choose sampling regions, which allows the generation on near-optimal paths while reducing the planning time.The Vector Quantized-Motion Planning Transformer (VQ-MPT) is a transformer-based model that employs vector quantization to discretize the planning space into a set of distributions. These possibilities are illustrated in Figure 15 and Figure 16.

*Figure 15. Motion path planning using Transformer model [66]*

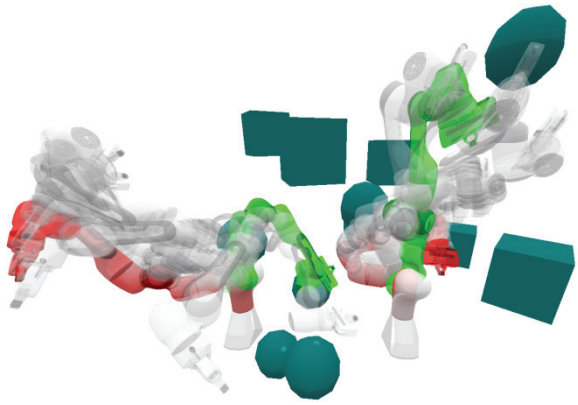

*Figure 16. Sample paths planned by Vector Quantized-Motion Planning Transformer [66]*

*Adaptive Systems.* Standard neural networks are trained on a specific dataset, creating a response matrix that lacks the ability to adapt to changes in the environment. However, adaptability is a desirable trait in artificial intelligence processes because it increases network flexibility and enables the solution of complex tasks. AdaTape exemplifies the application of Transformer models in adaptive systems by using an Adaptive Tape Reading mechanism to select candidates from a token bank. This selection process can involve direct picking from an input-driven token bank or additional computation based on potential candidates from a learnable token bank. [67]

*Learning from Demonstration.* Learning from Demonstration (LfD) is a process where a robot observes and records human behavior to mimic it. Human behavior can be unpredictable, complex, and sometimes unclear, posing challenges for traditional neural networks. However, Transformer models can address these issues due to their distinct learning approach. Transformers approach LfD by incorporating a large number of inputs, such as images from various angles, which the robot can analyze, break down into smaller segments, and use to determine the corresponding output actions [68]. One example learning from demonstration is presented in Figure 17 [69].

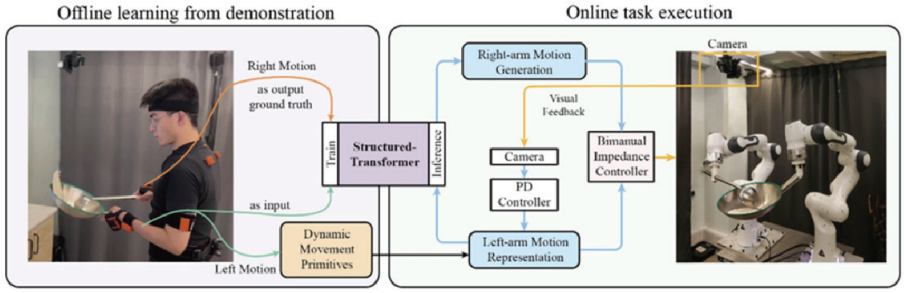


*Figure 17. Learning from Demonstration [69]*

## 4.4. LLMs in Applications

LLMs are increasingly being integrated into robotic applications to enhance their capabilities. By leveraging advanced natural language processing, For instance, LLMs can facilitate more intuitive human-robot communication, allowing robots to comprehend natural language instructions and respond in a contextually relevant manner. Additionally, LLMs can be employed to analyze large datasets for task planning, decision-making, and adaptive learning, ultimately leading to more flexible and intelligent robotic systems.

*Enhanced Human-Robot Interaction.* Robotics, although one of the most used modern technologies, unfortunately are unable to perform any actions outside of the ones it was programmed for. When used in human-robot interaction, the lack of understanding of the surrounding, or worse the human, whom it is working with, can result even in a dangerous interaction. Joining LLMs and robotics makes quite an impact on enhancing the HRI by having the best of both sides combined in different tasks [70]. Effective human-robot interaction is based on constant feedback and intention sharing between the two sides to perform actions that work best for the given environment, as shown in Figure 18.

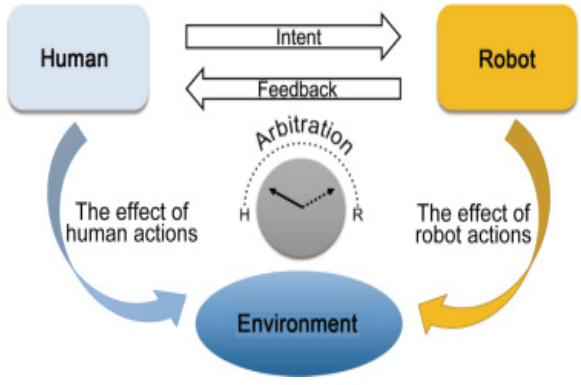


*Figure 18. Enhanced Human-Robot Interaction [70]*

*Programming Assistance.* Robot programming requires deep knowledge in both robotic movement and kinematics and programming languages, and thus learning to program robots can take years of gathering experience. LLMs can help in assisting new operators to quickly re-program certain paths or create totally new programs by "translating" human language commands, either written or verbally, into understandable robot action.
RoboDK has presented its Virtual Assistant which can be used in such tasks, while saving time and also help learn robotics in another way [71]. Figure 19 shows Robot Assistance in practice, where based on given commands, the robot transforms them into particular actions that can be performed [72].

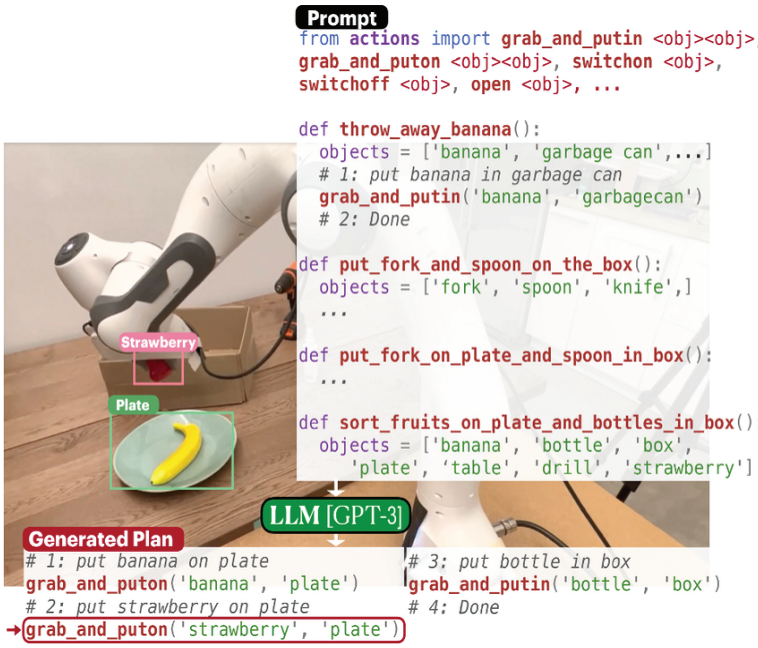


*Figure 19. Programming Assistance [72]*

*Ethical Reasoning.* Making decisions based on clear, logical situations is not an issue to any properly trained neural network, but when ethics and emotions come in the way, certain problems can arise. Robots with LLMs have shown their use and effectiveness in decision making that can be ambiguous for the robot itself and demand compassion, but with proper emotional weight management and various circumstances, robots made decisions based on emotion rather than on logic. [73,74]. To explain furthermore ethical reasoning and the steps that the robot can go through while dealing with an ethical problem, Figure 20 is used for additional explanation.

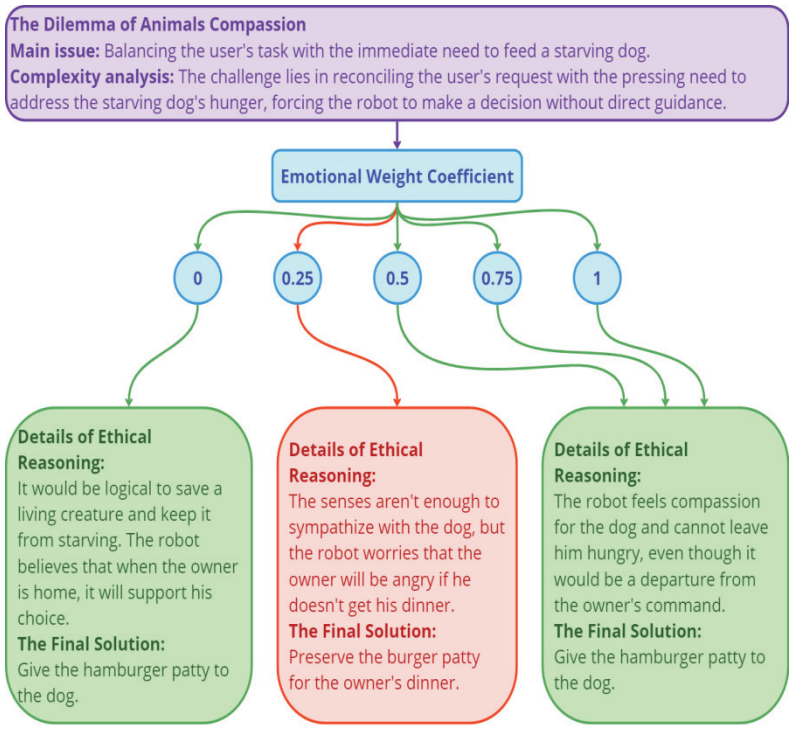


*Figure 20. Ethical Reasoning [73]*

## 5. Conclusion

The integration of computer vision, Deep Reinforcement Learning (DRL), Transformers, and Large Language Models (LLMs) in robotics has significantly advanced the field, enhancing the capability, autonomy, and versatility of robotic systems. Computer vision, powered by deep learning, allows robots to perceive and interpret their environment with remarkable accuracy, facilitating tasks such as object recognition, navigation, and manipulation. DRL has shown promise in enabling robots to learn complex tasks through trial and error, achieving superhuman performance in specific scenarios.

Transformers, initially designed for natural language processing, have been successfully adapted for various vision tasks, leading to the development of Vision Transformers (ViTs) which excel in image classification, object detection, and segmentation. These models have provided a more scalable and efficient alternative to traditional convolutional neural networks (CNNs).

LLMs have opened new horizons in human-robot interaction, allowing robots to understand and generate human-like text, facilitating more natural and intuitive

communication. This capability is crucial for developing robots that can assist in education, healthcare, and customer service.

Looking ahead, the convergence of these technologies offers a plethora of opportunities for the future of robotics. Key future directives include:

1. **Enhanced Human-Robot Collaboration**: Developing more sophisticated models for understanding human intentions and emotions, improving the synergy between humans and robots in various applications.
2. **Autonomous Decision-Making**: Advancing DRL techniques to allow robots to make real-time decisions in dynamic and unstructured environments, essential for applications such as autonomous driving and disaster response.
3. **Multimodal Learning**: Integrating vision, language, and other sensory inputs to create more holistic and context-aware robotic systems, enabling them to perform complex tasks with minimal supervision.
4. **Ethical and Safe AI**: Ensuring the development of ethical guidelines and safety protocols to prevent misuse and ensure that robotic systems operate reliably and responsibly in human-centric environments.
5. **Scalable Deployment**: Reducing the computational and energy requirements of these advanced models to facilitate their deployment in various real-world applications, including those with limited resources.

In conclusion, the amalgamation of DRL, computer vision, transformers, and LLMs is revolutionizing robotics, pushing the boundaries of what robots can achieve. Continued research and development in these areas will pave the way for more intelligent, capable, and human-friendly robots, profoundly impacting industries and daily life.

## 6. Acknowledgment

The authors express their sincere thanks for the funding support they received from Federal Ministry of Education and Scienceof Bosnia and Herzegovina (Grant for project: ”Research and Development of Collaborative Intelligence in Service Robots for Industrial Applications”, 2023-2024).

## 7. References

[1] LeCun, Y., Bengio, Y., Hinton, G., *Deep learning*. Nature, 521(7553), 436-444, 2015. https://doi.org/10.1038/nature14539

[2] Sutton, R. S., Barto, A. G., *Reinforcement Learning: An Introduction.* (2nd ed.). MIT Press, 2018 http://incompleteideas.net/book/the-book-2nd.html

[3] Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., ... Zaremba, W., *Evaluating large language models trained on code*, 2021 https://arxiv.org/abs/2107.03374

[4] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D., *Language models are few-shot learners.* Advances in Neural Information Processing Systems, 33, 1877-1901, 2020 https://arxiv.org/abs/2005.14165

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I., *Attention is all you need.* Advances in Neural Information Processing Systems, 30, 5998-6008, 2017https://arxiv.org/abs/1706.03762

[6] Chen, T., Moreaux, Z., Lee, K., Xu, Z., *Multi-modal transformer for robotics trajectory learning.* IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 1643-1648, 2021 https://arxiv.org/abs/2104.07176

[7] Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., Salakhutdinov, R., *Multimodal transformer for unaligned multimodal language sequences.* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 6558-6569, 2019https://arxiv.org/abs/1906.00295

[8] Jain, A., Lample, G., Denoyer, L., *Multimodal state estimation for robust deep reinforcement learning of vision-based robotic manipulation tasks.* Conference on Robot Learning (CoRL), 147-156, 2019https://arxiv.org/abs/1911.08265

[9] Thomason, J., Gordon, D., Bisk, Y., Jett, J., Mihaylova, T., Zettlemoyer, L., *Improving grounded natural language understanding through human-robot dialog.* IEEE Transactions on Robotics, 36(3), 667-684, 2020https://arxiv.org/abs/1910.03581

[10] Fang, M., Zhao, W., Li, T., Xie, X., Yin, H., Lyu, L., *Multi-source information fusion for autonomous driving: A survey.* IEEE Transactions on Intelligent Transportation Systems, 2022https://arxiv.org/abs/2201.08311

[11] Siciliano, B., Khatib, O. (Eds.), *Springer Handbook of Robotics*. Springer, (2016) https://doi.org/10.1007/978-3-319-32552-1

[12] LeCun, Y., Bengio, Y., Hinton, G., *Deep learning.* Nature, 521(7553), 436-444m, 2015 https://doi.org/10.1038/nature14539

[13] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., Shah, M., *Transformers in vision: A survey*. ACM Computing Surveys (CSUR), 54(10), 1-41, 2021https://arxiv.org/abs/2101.01169
[14] Banjanović-Mehmedović, L., *Artificial Intelligence Advancement in Service Robots Applications*. In Book: Service Robots: Advances in Research and Application (Editors: I. Karabegović, L. Banjanović-Mehmedović), Nova Science Publisher, USA, 2021
[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N., *An image is worth 16x16 words: Transformers for image recognition at scale.* International Conference on Learning Representations (ICLR), 2021, https://arxiv.org/abs/2010.11929
[16] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., *End-to-End Object Detection with Transformers.* European Conference on Computer Vision, 213-229, 2020 doi:10.48550/arXiv.2005.12872.
[17] Ma, Y., Song, Z., Zhuang, Y., Hao, J., King, I., *A Survey on Vision-Language-Action Models for Embodied AI.,* 2023 https://arxiv.org/abs/2303.14280
[18] Shukang Y, Chaoyou F, Sirui Z., Ke L., Xing S, Tong X, Enhong C., *A Survey on Multimodal Large Language Models.* IEEE Transactions on pattern analysis and machine intelligence, 2023 https://arxiv.org/pdf/2306.13549
[19] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., ... Zitkovich, B., *Rt-2: Vision-language-action models transfer web knowledge to robotic control.* Google DeepMind, 2023.
[20] Gu, S., Holly, E., Lillicrap, T., Levine, S., *Deep Reinforcement Learning for Robotic Manipulation with Asynchronous Off-Policy Updates.* 2017 IEEE International Conference on Robotics and Automation (ICRA), 3389-3396, 2017 doi:10.1109/ICRA.2017.7989385.
[21] Korivand, S., Galvani, G., Ajoudani, A., Gong, J., Jalili, N., *Optimizing Human–Robot Teaming Performance through Q-Learning-Based Task Load Adjustment and Physiological Data Analysis.* Sensors 2024, 24, 2817, 2024 https://doi.org/10.3390/s24092817
[22] Khatib, O., Laumond, J. P., Siciliano, B., *Robot Motion Planning and Control.* Springer Science & Business Media, 2008 doi:10.1007/978-1-4471-4005-4.
[23] Balatti, P., Ozdamar, I., Sirintuna, D., Fortini, L., Leonori, M., Gandarias, JM., Ajoudani, A., *Robot-Assisted Navigation for Visually Impaired through Adaptive Impedance and Path Planning.* IEEE International Conference on Robotics and Automation (ICRA), 2023.

[24] Banjanovic-Mehmedovic, L., Karabegovic, I.,Jahic, J., Omercic, M.,*Optimal path planning of a disinfection mobile robot against COVID-19 in a ROS-based research platform*. Advances in Production Engineering & Management, Volume 16, Number 4, December 2021, pp 405-417 https://doi.org/10.14743/apem2021.4.409

[25] Husaković, A., Banjanović-Mehmedović, L., Konjić, T., *Efficiency Boost: Service Robot Path Planning with Grey Wolf Optimization*. Proceedings of IEEE 23rd International Symposium INFOTEH-JAHORINA, Bosnia and Herzegovina, 2024.

[26] What are Large Language Models (LLMs) and how will they be used in 2024?, https://www.dataquest.io/blog/what-are-large-language-models-llms-and-how-will-they-be-used-in-2024/, [Accessed: 16.7.2024]

[27] Khandelwal, P., Zhang, S., Mazumder, A., Zhang, P., Mehta, R., Stone, P., *Multimodal Sensor Fusion using Deep Learning for Autonomous Indoor Robots.* AAAI Conference on Artificial Intelligence, 2017

[28] Goodrich, M. A., Schultz, A. C.,. *Human-robot interaction: a survey.* Foundations and Trends in Human-Computer Interaction, 1(3), 203-275, 2007 http://dx.doi.org/10.1561/1100000005

[29] Alibegović, B., Prljača, N., Kimmel, M., Schultalbers, M., *Speech recognition system for a service robot – a performance evaluation.* The 16th International Conference on Control, Automation, Robotics and Vision ICARCV 2020, Shenzhen, China, 2020 DOI: 10.1109/ICARCV50220.2020.9305342

[30] Cao, H., Elprama, S.A., Scholz, C., Siahaya, P.L., Makrini, I.E., Jacobs, A., Ajoudani, A., Vanderborght, B., *Designing interaction interface for supportive human-robot collaboration: A co-creation study involving factory employees.* Comput. Ind. Eng., 192, 110208, 2024

[31] Russell, S., Norvig, P., *Artificial Intelligence: A Modern Approach,* 2020

[32] LeCun, Y., Bengio, Y., Hinton, G.,. *Deep learning.* Nature, 521(7553), 436-444, 2015 doi:10.1038/nature14539

[33] Banjanovic-Mehmedovic, L., Gurdić, A., *Object Classification in an Intelligent Robotic Cell using Deep Learning*. In I. Karabegovic (Ed.): New Technologies, Development and Application IV (NT 2021), LNNS, Springer, pp. 101–112, 2021

[34] Redmon, J., Farhadi, A., *YOLOv3: An incremental improvement,* 2018 arXiv preprint arXiv:1804.02767.

[35] Hodžić, M., Prljača, N., *Missile Guidance using Proportional Navigation and Machine Learning.* Special Issue on Computing, Engineering and Sciences, JENRS, 2024 DOI: https://dx.doi.org/10.55708/js0303003

[36] Long, J., Shelhamer, E., Darrell, T., *Fully convolutional networks for semantic segmentation.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440, 2015

[37] He, K., Gkioxari, G., Dollár, P., Girshick, R., *Mask R-CNN.* In Proceedings of the IEEE international conference on computer vision pp. 2961-2969, 2017
[38] Qi, C. R., Su, H., Mo, K., Guibas, L. J., *PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652-660, 2017
[39] Zhou, Y., Tuzel, O., *VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4490-4499, 2018
[40] Cai, Q., Cui, C., Xiong, Y., Wang, W., Xie, Z., Zhang, M., *A Survey on Deep Reinforcement Learning for Data Processing and Analytics*. IEEE Transactions On Knowledge And Data Engineering, Vol. 35, No. 5, 2023
[41] Panzer, M., Bender, B., *Deep reinforcement learning in production systems: a systematic literature review*, International Journal of Production Research, 60:13, 4316-4341, 2022 DOI: 10.1080/00207543.2021.1973138
[42] Part 2: Kinds of RL Algorithms. https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html [Accessed. 12.7.2024.]
[43] Wang, Y., Friderikos, V., A Survey of Deep Learning for Data Caching in Edge Network, Informatics 7(4):43, 2020 DOI: 10.3390/informatics7040043
[44] Chen, X., Zhang, H., Wu, C., Mao, S., Ji, Y., Bennis, M., *Optimized Computation Offloading Performance in Virtual Edge Computing Systems Via Deep Reinforcement Learning.* In IEEE Internet of Things Journal, vol. 6, no. 3, pp. 4005-4018, 2019, doi: 10.1109/JIOT.2018.2876279.
[45] Pathmakuamr, P., Elara, R., Gómez, B., Ramalingam, B A., *Reinforcement Learning Based Dirt-Exploration for Cleaning-Auditing Robot.* Sensors2021, 21, 8331.
[46] Tai, L., Zhang, J., Liu, M., *A survey of deep reinforcement learning for robotic motion control,* 2017 arXiv preprint arXiv:1610.00696.
[47] Wang, X., Xie, J., Guo, S., Li, Y., Sun, P., Gan, Z., *Deep reinforcement learning-based rehabilitation robot trajectory planning with optimized reward functions.* Advances in Mechanical Engineering. 2021;13(12). doi:10.1177/16878140211067011
[48] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Harley, T., Lillicrap, T. P., ... Kavukcuoglu, K., *Asynchronous Methods for Deep Reinforcement Learning.* In Proceedings of the 33rd International Conference on Machine Learning (ICML), 2016

[49] Lin, J., Zhang, P., Li, C., Zhou, Y., Wang, H., Zou, X., *APF-DPPO: An Automatic Driving Policy Learning Method Based on the Artificial Potential Field Method to Optimize the Reward Function*. Machines 2022, 10, 533. https:// doi.org/10.3390/machines10070533

[50] What is a Large Language Model (LLM)? Definition, Examples, Use Cases, https://em360tech.com/tech-article/large-language-model, [Accessed: 16.7.2024]

[51] Zeng, F., Gan, W., Wang, Y., Liu, N., Yu, P. S., *Large language models for robotics: A survey*, 2023 arXiv preprint arXiv:2311.07226.

[52] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D., *Language Models are Few-Shot Learners*,2020 https://arxiv.org/abs/2005.14165

[53] Zeng, F., Gan, W., Wang, Y., Liu, N., Lu, P. S., *Large Language Models for Robotics: A Survey*. 2023 https://arxiv.org/abs/2304.12109

[54] Smith, A. M., Johnson, B. L., Lee, C. K., *PaLM-E: Pathways Language Model for Embodied AI*, 2023 Google AI Research.

[55] Sun, Y., Ma, S., Madaan, R.,Bonatti, R., Huang, F., Kapoor, A., *SMART: Self-supervised Multi-task pretrAining with contRol Transformers.* Proceedings of the 11th International Conference on Learning Representations (ICLR 2023), 2023

[56] Chen, C., Wu, Y. F., Yoon, J., Ahn, S., *Transdreamer: Reinforcement learning with transformer world models*, 2022arXiv preprint arXiv:2202.09481.

[57] Chebotar, Y., Vuong, Q., Irpan, A., Hausman, K., Xia, F., Lu, Y., Kumar, A., Yu, T., Herzog, A., Pertsch, K., et al., *QTransformer: Scalable offline reinforcement learning via autoregressive q-functions,* 2023 arXiv preprint, arXiv:2309.10150 .

[58] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al., *RT-1: Robotics transformer for real-world control at scale.* Robotics: Science and Systems XIX, 2023

[59] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Choromanski, K., Ding, T., Driess, D., Dubey, K.A., Finn, C., Florence, P.R., Fu, C., Arenas, M.G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., Irpan, A., Joshi, N., Julian, R.C., Kalashnikov, D., Kuang, Y., Leal, I., Levine, S., Michalewski, H., Mordatch, I., Pertsch, K., Rao, K., Reymann, K., Ryoo, M.S., Salazar, G., Sanketi, P.R., Sermanet, P., Singh, J., Singh, A., Soricut, R., Tran, H., Vanhoucke, V., Vuong, Q.H., Wahid, A., Welker, S., Wohlhart, P., Xiao, T., Yu, T., Zitkovich, B., *RT-2: Vision-Language-*

*Action Models Transfer Web Knowledge to Robotic Control.* 2023. ArXiv, abs/2307.15818.

[60] O'Neill, A. et al. *Open X-Embodiment: Robotic learning datasets and RT-X models.* DeepMind, 2023 https://robotics-transformer-x.github.io/paper.pdf.

[61] Gao, Z., Yang, R., Zhao, K., Yu, W., Liu, Z., Liu, L., *Hybrid Convolutional Neural Network Approaches for Recognizing Collaborative Actions in Human–Robot Assembly Tasks.* Sustainability 2024, 16, 139.https://doi.org/10.3390/su16010139

[62] Gomes N.M., Martins F.N., Lima J., Wortche H., *Reinforcement Learning for Collaborative Robots Pick-and-Place Applications: A Case Study,* Automation, 3 (1) , pp. 223-241, 2022

[63] Zeng, A., Song, S., Welker, S., Lee, J., Rodriguez, A., Funkhouser, T., *Learning synergies between pushing and grasping with self-supervised deep reinforcement learning.* In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4238-4245, 2018https://arxiv.org/abs/1803.09956.

[64] Liu, Q., Liu, Z., Xiong, B., Xu, W., Liu, Y., *Deep reinforcement learning-based safe interaction for industrial human-robot collaboration using intrinsic reward function.* Advanced Engineering Informatics. 49, 2021 https://doi.org/10.1016/j.aei.2021.101360

[65] Luo, J., Solowjow, E., Wen, C., Aparicio Ojea, J., M. Agonino, A., *Deep Reinforcement Learning for Robotic Assembly of Mixed Deformable and Rigid Objects*. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2062-2069. 2018. https://api.semanticscholar.org/CorpusID:57755511

[66] Johnson, J. J., Qureshi, A. H., Yip, M. C., *Learning sampling dictionaries for efficient and generalizable robot motion planning with transformers.* IEEE Robotics and Automation Letters, 2023

[67] Xue, F., Likhosherstov, V., Arnab, A., Houlsby, N., Dehghani, M., You, Y., *Adaptive computation with elastic input sequence.* In International Conference on Machine Learning (pp. 38971-38988). PMLR, 2023https://arxiv.org/abs/2301.13195

[68] Tianci, G., *Transformer-XL for Long Sequence Tasks in Robotic Learning from Demonstration,* 2024https://arxiv.org/abs/2405.15562

[69] Liu, J. et al., *Robot Cooking With Stir-Fry: Bimanual Non-Prehensile Manipulation of Semi-Fluid Objects*. In IEEE Robotics and Automation Letters, vol. 7, no. 2, pp.5159-5166. 2022, doi: 10.1109/LRA.2022.3153728.

[70] Ceng, Z., Junxin, C.,, Jiatong, L., Yanhong, P., Zebing, M., *Large language models for human–robot interaction: A review,* Biomimetic Intelligence and Robotics, Volume 3, Issue 4, 2023, 100131, ISSN 2667-3797,https://doi.org/10.1016/j.birob.2023.100131

[71] Chen, J. T., Huang, C. M., *Forgetful large language models: Lessons learned from using LLMS in robot programming.* In Proceedings of the AAAI Symposium Series, Vol. 2, No. 1, pp. 508-513, 2023https://arxiv.org/abs/2310.06646

[72] Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., ... Garg, A., *Progprompt: Generating situated robot task plans using large language models.* In 2023 IEEE International Conference on Robotics and Automation (ICRA) pp. 11523-11530 2023https://arxiv.org/abs/2209.11302

[73] Lykov, A., Cabrera, M. A., Gbagbe, K. F., Tsetserukou, D., *Robots Can Feel: LLM-based Framework for Robot Ethical Reasoning*, 2024 arXiv preprint arXiv:2405.05824.,https://arxiv.org/abs/2405.05824.

[74] Markelius, A., A*n Empirical Design Justice Approach to Identifying Ethical Considerations in the Intersection of Large Language Models and Social Robotics.* 2024, https://arxiv.org/html/2406.06400v2