Front. Eng. Manag. 2025, 12(1): 177–200
https://doi.org/10.1007/s42524-025-4136-9

REVIEW ARTICLE

Junming FAN, Yue YIN, Tian WANG, Wenhang DONG, Pai ZHENG, Lihui WANG

Vision-language model-based human-robot collaboration for smart manufacturing: A state-of-the-art survey

© The Author(s) 2024. This article is published with open access at link.springer.com and journal. hep.com.cn

Abstract human–robot collaboration (HRC) is set to transform the manufacturing paradigm by leveraging the strengths of human flexibility and robot precision. The recent breakthrough of Large Language Models (LLMs) and Vision-Language Models (VLMs) has motivated the preliminary explorations and adoptions of these models in the smart manufacturing field. However, despite the considerable amount of effort, existing research mainly focused on individual components without a comprehensive perspective to address the full potential of VLMs, especially for HRC in smart manufacturing scenarios. To fill the gap, this work offers a systematic review of the latest advancements and applications of VLMs in HRC for smart manufacturing, which covers the fundamental architectures and pretraining methodologies of LLMs and VLMs, their applications in robotic task planning, navigation, and manipulation, and role in enhancing human–robot skill transfer through multimodal data integration. Lastly, the paper discusses current limitations and future research directions in VLM-based HRC, highlighting the trend in fully realizing the potential of these technologies for smart manufacturing.

Keywords vision-language models, large language models, human–robot collaboration, smart manufacturing

Received Jul. 27, 2024; revised Nov. 13, 2024; accepted Nov. 15, 2024

Junming FAN, Yue YIN, Tian WANG, Wenhang DONG, Pai ZHENG (✉)
Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China
E-mail: pai.zheng@polyu.edu.hk

Lihui WANG (✉)
Department of Production Engineering, KTH Royal Institute of Technology, Stockholm, Sweden
E-mail: lihuiw@kth.se

This work was mainly supported by the funding support from the Research Institute for Advanced Manufacturing (RIAM) of The Hong Kong Polytechnic University (1-CDJT); the Intra-Faculty Interdisciplinary Project 2023/24 (1-WZ4N), by the Research Committee of The Hong Kong Polytechnic University; the State Key Laboratory of Intelligent Manufacturing Equipment and Technology, Huazhong University of Science and Technology (IMETKF2024010); Guangdong–Hong Kong Technology Cooperation Funding Scheme (GHX/075/22GD); Innovation and Technology Commission (ITC); the COMAC International Collaborative Research Project (COMAC-SFGS-2023-3148); and the General Research Fund from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. PolyU15210222 and PolyU15206723); Open access funding provided by the Hong Kong Polytechnic University.

1 Introduction

human–robot collaboration (HRC) has been regarded as a promising pathway to revolutionise the manufacturing sector by leveraging the complementary strengths of humans and robots (Matheson et al., 2019). This synergy aims to enhance productivity, adaptability, and efficiency, marking a significant paradigm shift in smart manufacturing (Wang et al., 2019). The recent astonishing breakthroughs in the Artificial Intelligence (AI) field, including computer vision and natural language processing, have exhibited huge potential to drive this transformation by endowing robots with multimodal perception and understanding capabilities, enabling more sophisticated and seamless collaborations between humans and robots (Fan et al., 2022; Wang et al., 2024c).

As the most recent advancement of AI, Large Language Models, such as GPT-3 (Brown et al., 2020) and GPT-4 (Achiam et al., 2023), have demonstrated exceptional capabilities in natural language processing, enabling them to exhibit human-like comprehension and conversational abilities. However, standard LLMs are inherently limited to processing textual information, which restricts their applicability in scenarios such as HRC that require visual context. In response to this limitation, Vision-Language Models (VLMs) have been developed to integrate visual and textual data (Zhou et al., 2022), thereby enhancing the robot's ability to interpret and interact with its environment. Prevalent examples of

VLMs include CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), which have shown promise in tasks such as image captioning, visual question answering, and multimodal reasoning. The significant advancements of VLMs have inspired the initial adoptions in HRC scenarios to enhance robotic intelligence and human–robot communication flexibility (Fan and Zheng, 2024; Park et al., 2024). However, existing research endeavors are rather scattered in different applications and perspectives, resulting in a lack of a comprehensive investigation of the potential of VLMs in HRC scenarios.

This paper aims to bridge this gap by providing a systematic review of the latest advancements and applications of VLMs in HRC. The overview of the structure of this survey is illustrated in Fig. 1. The exploration begins with the fundamental architectures and pretraining methodologies of LLMs and VLMs in Section 3, in which we briefly introduce the intricacies of transformer architectures, the mechanisms of pretraining on large-scale data sets, and the subsequent fine-tuning processes that tailor these models to specific applications. Next, the practical applications of VLMs in robotic task planning, navigation, and manipulation are examined in Section 4. These capabilities are essential for robots to function effectively in dynamic manufacturing settings where tasks are varied and environments are constantly changing. On top of the basic functionalities, it is imperative to equip robots with advanced skill acquisition ability in order to better adapt to the futuristic flexible manufacturing environments. Therefore, the role of VLMs in enhancing human–robot skill transfer is also explored in Section 5. Compared to traditional robot skill acquisition methods that often involve extensive programming, VLMs can significantly streamline this process by using visual and linguistic inputs to facilitate more intuitive and effective human–robot teaching. Lastly, the current challenges, that prevent the immediate deployment of VLMs in manufacturing scenes, and potentid future directions toward unlocking the full potential of VLMs in smart manufacturing are discussed in Section 6.

2 Literature review process

The literature review process for this paper follows a systematic manner to ensure a comprehensive and unbiased overview of the latest advancements and applications of VLMs in HRC. As depicted in Fig. 2, an initial search was employed to identify relevant literature from various academic databases, including Web of Science, Scopus, and IEEE Xplore. The search was conducted using a combination of keywords related to the core topics: "human–robot" and "vision language," covering the time span 2020–2024. The search yielded 63 items from Web of Science, 113 related documents from Scopus, and 89 from IEEE Xplore (July 15, 2024).

The literature selection process was conducted in two phases: *initial screening* and *detailed reviewing*, to ensure that the most relevant and high-quality studies were included. First, an initial filtering process was conducted, during which only journal and conference papers in English were included. Papers that were obviously beyond the scope of this survey were excluded based on their titles, keywords, and abstracts, resulting in 59 papers. Subsequently, an in-depth review process was implemented to further identify inadequate items and categorise suitable papers. Considering the results from the aforementioned databases are quite limited, supplementary relevant papers were added from other less rigorous search engines such as Google Scholar and arXiv. Notably, although arXiv papers are not normally accepted as trustworthy paper sources, a considerable amount of

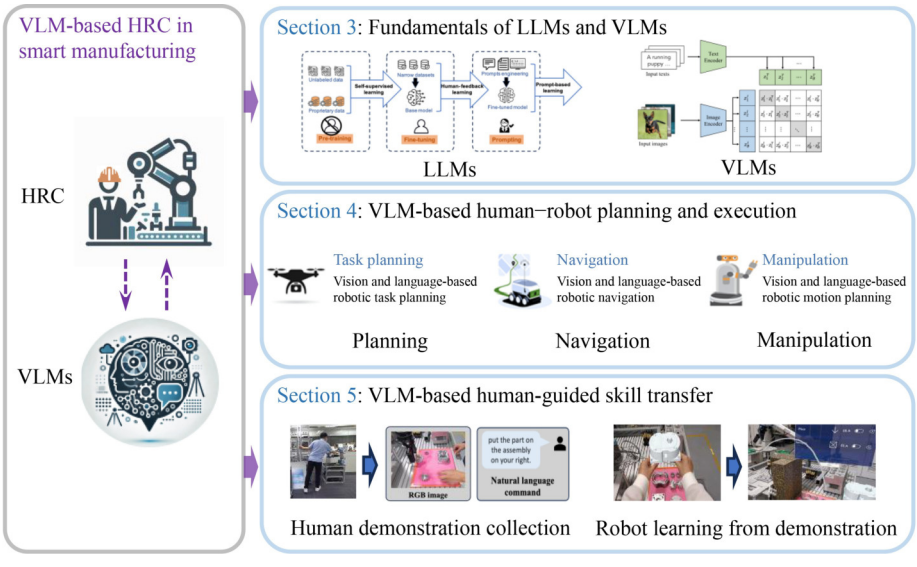

Fig. 1 Overview of the survey of VLM-based HRC for smart manufacturing.

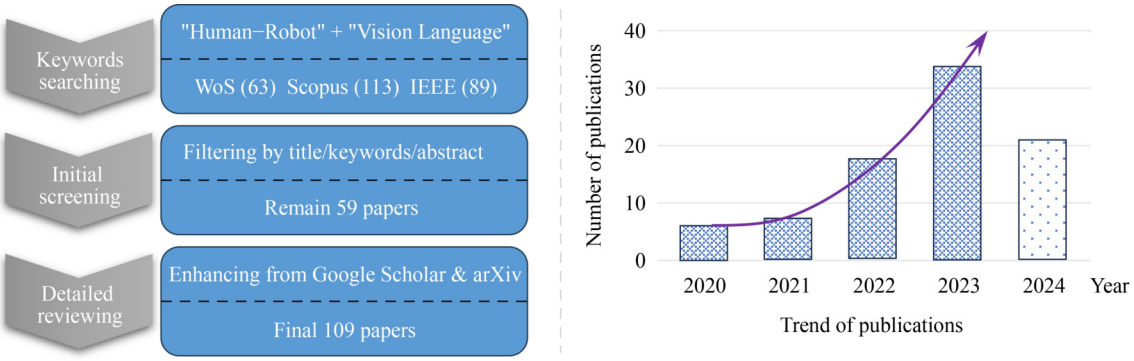

Fig. 2 Systematic literature review process and the trend of related publications.

state-of-the-art works related to LLMs and VLMs have not yet made it to formal publications and can only be found on arXiv, thereby leading to the frequent inclusion of arXiv papers in this review. Finally, 109 papers have been selected as the basis of this survey, which are further described in Section 3.5.

3 Revisiting LLMs and VLMs

In recent years, significant advancements have been made in the development of large language models (LLMs) (Zhao et al., 2023a; Chang et al., 2024). By scaling the size of data and models, these LLMs have exhibited exceptional emergent abilities, including instruction following (Peng et al., 2023), in-context learning (ICL) (Brown et al., 2020), and chain of thought (CoT) reasoning (Wei et al., 2022). Despite their impressive zero-shot and few-shot performance on various natural language processing (NLP) tasks, LLMs are intrinsically limited in their ability to interpret visual information, as they can only process discrete text. To overcome this limitation, researchers have developed vision-language models (VLMs) (Zhang et al., 2024a). VLM is designed to learn rich vision-language correlation from large-scale image-text pairs, thereby enhancing its capacity for comprehensive and accurate understanding and reasoning. The typical architectures of LLMs and VLMs are depicted in Fig. 3. This section provides a brief introduction to the fundamental concepts and development status of LLMs and VLMs, respectively.

3.1 Fundamentals of LLMs

LLMs are normally constructed on the transformer architecture (Vaswani et al., 2017), which uses a specifically designed neural network and the multi-head attention mechanism to understand context and meaning in text. The core mathematical operation in the self-attention mechanism is the computation of attention scores, which are derived from the query (Q), key (K), and value (V) vectors, as formulated in the following equation:

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where d_k represents the dimensionality of the key vectors. The softmax function normalizes the attention scores. The multi-head attention mechanism is then built upon several parallel self-attention heads:

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{head}_1,\ldots,\text{head}_h)W_O, \tag{2}$$

$$\text{head}_i=\text{Attention}(QW_{Q_i},KW_{K_i},VW_{V_i}), \tag{3}$$

where W_{Q_i}, W_{K_i}, W_{V_i} are learned projection matrices for each head, and W_O is the output projection matrix. The two main components of Transformers are encoders and decoders. Encoders extract and comprehend relevant information from input text using self-attention, while decoders generate translated text using the embeddings from the encoder (Fu et al., 2023). The encoder block combines multi-head self-attention and feed-forward neural networks. The output of each encoder layer can be

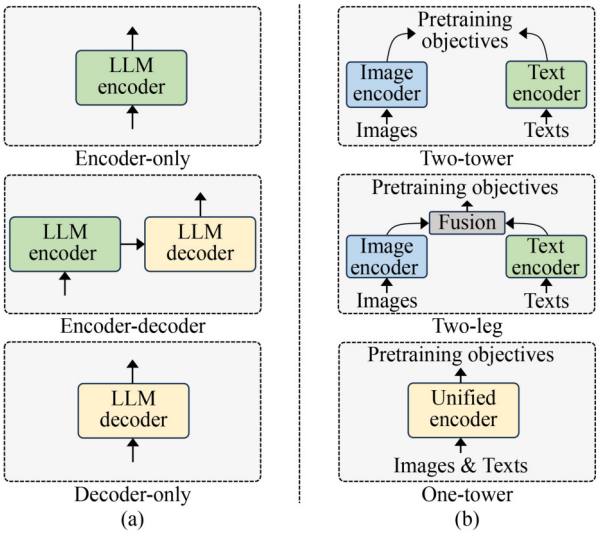

Fig. 3 Typical architectures of LLMs and VLMs: (a) Three types of LLMs; (b) The pretraining architectures of VLMs. (Zhang et al., 2024a).

represented as:

$$\text{EncoderOutput} = \text{LayerNorm}(X + \text{FFN}(\text{MultiHead}(X))), \tag{4}$$

where X is the input, FFN is the feed-forward network, and LayerNorm is the layer normalization operation. In contrast, the decoder block has an additional cross-attention mechanism that attends to the encoder's output:

$$\text{DecoderOutput} = \text{LayerNorm}(X + \text{FFN}(\text{MultiHead}(X, \text{EncoderOutput}))). \tag{5}$$

Based on the variation of the encoder-decoder structure, LLMs can be categorised into three types: Encoder-only, Decoder-only, and Encoder-Decoder.

• **Encoder-only models:** These models consist solely of encoders. While focusing on feature encoding for a better understanding of the text information, they cannot directly generate textual output, making them suitable for tasks like text categorisation (Kenton and Toutanova, 2019).

• **Encoder-decoder models:** This type of model incorporates both encoder and decoder components. They encode the input into feature information and pass it to the decoder, which then generates output according to the sequence. This structure effectively manages the connection between input and output sequences, making it suitable for translation and text summarizing tasks (Lewis et al., 2020).

• **Decoder-only models:** These models only have decoder components. They use the encoder to generate a corresponding sequence from the input encoding, focusing on generating or predicting output from a series of inputs (Wang et al., 2022a). They specialize in generation tasks, such as question answering, and represent the dominant architecture today.

Table 1 summarizes some well-known LLMs that fall into these categories. The following content of this section will briefly cover the pretraining, fine-tuning, and prompting technologies of LLMs.

3.1.1 Pretraining

Pretraining serves as the initial phase in which an LLM is subjected to training on an extensive collection of textual data in an unsupervised manner. This critical phase facilitates the development of fundamental linguistic capabilities and representational skills within the model. Utilizing the scalable features of the Transformer architecture, which incorporates self-attention mechanisms, the BERT model (Kenton and Toutanova, 2019) was developed. This framework advances the training of bidirectional language models using meticulously crafted tasks on expansive, unlabeled corpora. The word representations generated through this pretraining process are context-sensitive and highly effective with versatile semantic features. This innovative approach has motivated a considerable body of subsequent research, giving rise to the "pretraining and fine-tuning" paradigm. This paradigm has guided extensive investigations such as

Table 1 Mainstream LLMs

Type	Name	Year	Company	Open Source
Encoder-Only	BERT (Kenton et al., 2019)	2018	Google	Open
	RoBERTa (Liu et al., 2019)	2019	Meta	Open
	ERNIE (Zhang et al., 2019)	2019	Baidu	Open
	DeBERTa (He et al., 2020)	2020	Microsoft	Open
Encoder-Decoder	BART (Lewis et al., 2020)	2019	Meta	Open
	T5 (Raffel et al., 2020)	2019	Google	Open
	ChatGLM (GLM et al., 2024)	2023	Tsinghua University	Open
	FlanUL2 (Tay et al., 2023)	2023	Google	Open
Decoder-Only	GPT-1 (Radford et al., 2018)	2018	OpenAI	Open
	GPT-2 (Radford et al., 2019)	2019	OpenAI	Open
	GPT-3 (Brown et al., 2020)	2020	OpenAI	Close
	ERNIE3.0 (Sun et al., 2021)	2021	Baidu	Close
	LaMDA (Thoppilan et al., 2022)	2021	Google	Close
	PaLM (Chowdhery et al., 2023)	2022	Google	Close
	LLaMA (Touvron et al., 2023)	2023	Meta	Open
	Gemini (Team et al., 2023)	2023	Google	Open
	GPT-4 (Achiam et al., 2023)	2023	OpenAI	Close
	Claude (Anthropic 2023)	2023	Anthropic	Close

GPT-2 (Radford et al., 2019) and BART (Lewis et al., 2020), and the implementation of refined pretraining methodologies (Sanh et al., 2022). In practice, this paradigm typically involves adjusting the pretrained LLM through fine-tuning to meet the demands of specific downstream applications.

3.1.2 Fine-tuning and alignment

After pretraining, LLMs are fine-tuned to specialize in specific tasks or to align with human values and intents. Fine-tuning involves training the pretrained model on a smaller, task-specific data set, often using supervised learning techniques. Typical fine-tuning approaches are summarized as follows:

• **Transfer Learning:** Pretrained LLMs exhibit commendable performance across a spectrum of tasks. Nonetheless, for enhanced task-specific performance, these models need to undergo fine-tuning with task-specific data, a process referred to as transfer learning (Raffel et al., 2020).

• **Instruction-tuning:** To ensure models respond effectively to user queries, pretrained models are fine-tuned using instruction-formatted data, which includes natural language directives coupled with relevant input-output pairs. This approach not only enhances the model's ability to generalize across new scenarios, but also boosts its performance on specific tasks. Detailed methodologies for creating and varying instructional data are outlined in the literature (Chung et al., 2024).

• **Alignment-tuning:** LLMs can sometimes produce inaccurate or harmful content. To address this, alignment-tuning is conducted using human feedback to adjust outputs and discourage undesirable responses. This process ensures models align with ethical standards, characterized by the "HHH" criteria: helpful, honest, and harmless (Askell et al., 2021). Techniques such as reinforcement learning with human feedback (RLHF) (Ziegler et al., 2019) are employed, where models are refined through reward modeling (RM) and reinforcement learning (RL) to meet these criteria.

3.1.3 Prompting

After an LLM has been thoroughly trained and fine-tuned, the prompting technique is employed to elicit responses from the LLM. LLMs can be prompted in various configurations; some setups allow the model to adapt to instructions without further fine-tuning, while others require fine-tuning on data sets that incorporate diverse prompting styles (Kim et al., 2023b). Below is a brief overview of several commonly used prompt setups:

• **Zero-shot prompting:** The model is given a task without any examples, relying solely on its pretrained knowledge to generate a response (Kojima et al., 2022).

• **In-context learning:** Also known as few-shot learning, this method provides the model with a few examples within the prompt to guide its response (Dong et al., 2022).

• **Chain-of-Thought:** A prompting technique where the model is guided to generate step-by-step reasoning or explanations for its answers, improving performance on complex reasoning tasks (Wei et al., 2022).

3.2 Fundamentals of VLMs

The success of large-scale models in the NLP field has inspired the computer vision community to borrow text information to enhance visual recognition, leading to the thriving of VLMs. A VLM typically consists of two parallel encoders: one for processing visual data (such as images) and one for textual data (such as descriptions or instructions). These encoders transform the inputs into high-dimensional embeddings, which are then aligned or fused in a shared feature space, allowing the model to jointly interpret and reason about both visual and language inputs. The development of VLMs has attracted considerable attention as vision and language are the two most semantic-rich information sources. VLM is crucial for cross-modal applications like image captioning and visual question answering and plays an even more significant role in human–robot collaboration in smart manufacturing environments, where multimodal communications between human operators and robots are indispensable.

The most essential step of VLMs is pretraining (Radford et al., 2021), which allows models to learn the correlation between images and texts by employing a combination of visual and textual encoders. Typically, VLMs use separate encoders for each modality (vision and text), enabling the extraction of meaningful features from both input types. Subsequently, pretraining objectives guide the model in learning the relationships between these visual and textual elements. The rest of this section will introduce the three aspects of VLMs: vision-language encoding, vision-language correlation, and pretraining, of which the related works are summarized in Table 2.

3.2.1 Vision and language encoding

VLMs employ deep neural networks to extract features from the image-text pairs in a data set. The network normally consists of an image encoder and a text encoder, which respectively encode the image and text of an input data pair into visual and textual embeddings. This section outlines the architectures of the deep neural networks commonly used in VLM for image and text encoding.

• **Visual Encoder:** 1) CNN-based: Various Convolutional Networks, such as ResNet (He et al., 2016) and EfficientNet (Tan and Le, 2019), have been developed to

Table 2 Summarization of VLM pretraining works

Pretraining Aspect	Category	Name	Reference
Vision-Language Encoding	Visual Encoder	ResNet	He et al. (2016)
	Visual Encoder	EfficientNet	Tan and Le (2019)
	Textual Encoder	ViT	Dosovitskiy et al. (2020)
	Visual-Textual Encoder	CLIP	Radford et al. (2021)
Vision-Language Correlation	Contrastive Objectives	SimCLR	Chen et al. (2020)
	Generative Objectives	MoCo	He et al. (2020)
	Generative Objectives	Coca	Yu et al. (2022)
	Generative Objectives	Flava	Singh et al. (2022)
	Alignment Objectives	FIBER	Dou et al. (2022)
	Alignment Objectives	DetCLIP	Yao et al. (2022)
Pretraining Architecture	Two-Tower	CLIP	Radford et al. (2021)
	Two-Tower	ALIGN	Jia et al. (2021)
	Two-Leg	Coca	Yu et al. (2022)
	Two-Leg	Flava	Singh et al. (2022)
	One-Tower	CLIPPO	Tschannen et al. (2022)
	One-Tower	OneR	Jang et al. (2023)

enhance image feature learning (Radford et al., 2021). ResNet, which is particularly prevalent in VLMs, incorporates skip connections across convolutional blocks to prevent gradient issues and support deeper network architectures. 2) Transformer-based: Recent studies have extensively explored the application of Transformers in VLMs, especially ViT (Dosovitskiy et al., 2020), which is a prototypical Transformer architecture for image feature learning. It processes images by dividing them into fixed-size patches, which are linearly projected and position-embedded before encoding.

• **Textual Encoder:** Transformers and their variants have become fundamental for text feature encoding. The standard Transformer features an encoder-decoder framework. Studies in VLMs, such as CLIP (Radford et al., 2021), typically employ this standard architecture with slight adaptations.

3.2.2 Vision and language correlation

The core of VLMs is the understanding of the correlation between paired vision-language data. Regarding this issue, various pretraining objectives have been designed to enhance the learning of vision-language features:

• **Contrastive Objectives:** Contrastive objectives enable VLMs to acquire discriminative representations by attracting paired samples while pushing away unpaired ones within the feature space. In the case of VLMs, image-text contrastive learning is typically leveraged, which is achieved by minimising the symmetric image-text infoNCE loss (Chen et al., 2020).

• **Generative Objectives:** This type of objective learns the vision-language correlation feature by training networks to generate image/text data via image generation (He et al., 2021), language generation (Yu et al., 2022), or cross-modal generation (Singh et al., 2022).

• **Alignment Objectives:** Alignment objectives are designed to align the image-text features via global image-text matching (Dou et al., 2022) or region-word matching (Yao et al., 2022) on embedding space.

3.2.3 Pretraining architecture

The architecture for VLM pretraining mainly symbolises how the vision and language processing branches and embeddings are interconnected and communicated. The most widely adopted frameworks in VLMs include two-tower, two-leg and one-tower pretraining frameworks.

Specifically, the two-tower framework is commonly utilized in VLMs (Radford et al., 2021; Jia et al., 2021) employing separate encoders to process input images and texts. In a variation, the two-leg framework (Yu et al., 2022; Singh et al., 2022) includes extra multimodal fusion layers, facilitating interaction between image and text features. By contrast, one-tower VLMs (Tschannen et al., 2022; Jang et al., 2023) integrate vision and language processing into a single encoder, promoting more efficient communication across different data modalities.

While the pretraining and deployment of VLMs have witnessed significant progress in recent years, the application of current VLMs in real-life manufacturing scenarios still faces challenges such as high computational demands, the scarcity of high-quality data sets, and

latency issues, which comprise the robustness and reliability of VLMs in handling variations in real-world production (Challenge 6.1).

4 VLM-based human–robot interactive task planning and execution

The remarkable capabilities of LLMs and VLMs have attracted the attention of the scientific community, prompting extensive exploration into their potential applications in robotic interactive task planning, navigation, and manipulation, as shown in Fig. 4. The integration of LLMs and VLMs in these areas has shown promising results, demonstrating their potential to revolutionise how robots perform complex tasks. Moreover, their application extends to industrial collaborative robots and other intricate scenarios, showcasing their versatility and adaptability in enhancing robotic functions and efficiency. These advancements are paving the way for more sophisticated and intelligent robotic systems capable of operating in diverse and dynamic environments. This section summarizes the most representative work of VLMs and LLMs in task planning, navigation and manipulation, including the foundation models used, success rates and applications.

4.1 Vision and language task planning

Traditional task-planning methods often rely on predefined rules and logical reasoning. However, these methods are powerless when faced with a dynamic and complex real world. The effectiveness and capability of recent VLMs have attracted the attention of relevant researchers to explore vision and language task planning. As shown in Fig. 5, vision and language task planning refers to the ability of a robot to complete task planning based on its visual perception of the environment and its understanding of the target task. LLMs and VLMs have strong logical reasoning and visual perception capabilities, which enable robots to complete task planning with a full understanding of the task scenario and human instructions, and

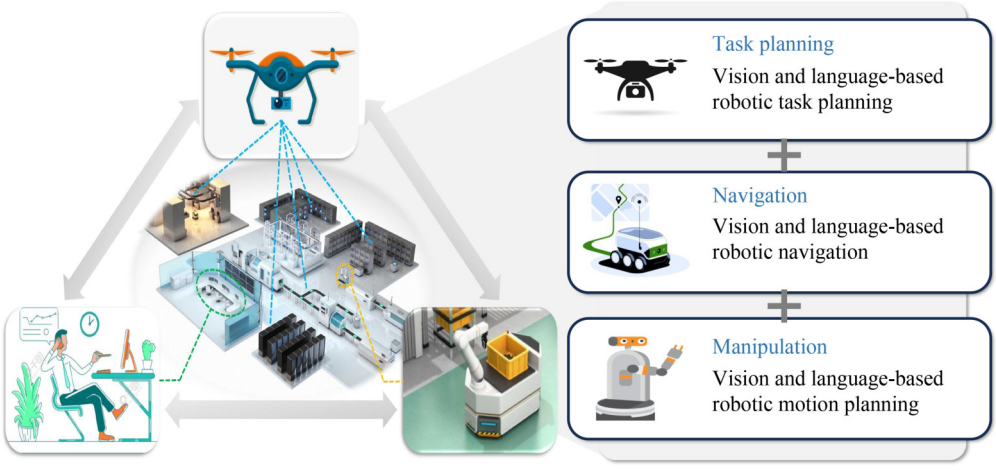

Fig. 4 VLM-based human–robot interactive task planning, navigation, and manipulation.

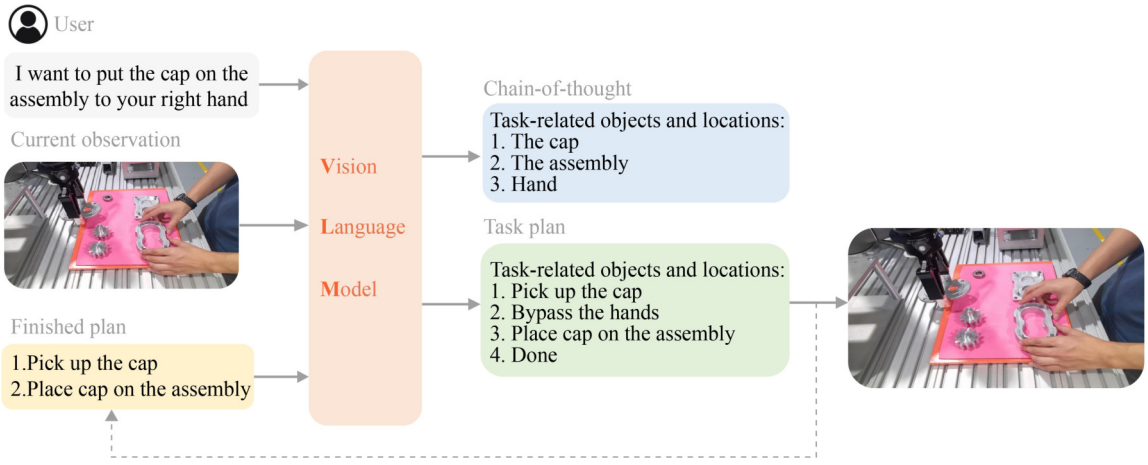

Fig. 5 Vision and language task planning adapted with permission from Hu et al. (2023).

promote effective communication and collaboration between humans and robots. Table 3 lists the representative results of the recent application of VLMs or LLMs to robot task planning.

4.1.1 Task understanding and decomposition

Task understanding and decomposition is the first step in task planning, which involves extracting task objectives from natural language descriptions and decomposing complex tasks into a series of manageable subtasks. VLMs play a crucial role in this process as they can extract rich semantic information from text and images. For example, Song et al. (2023) combined the user's natural language instructions with the environment information, and realized the understanding of tasks such as navigation and manipulation based on LLMs and object detectors. Zheng et al. (2024) adopted BERT and ResNet to parse real scenes, and prompt LLM to decompose the overall task into subtasks. Zhao et al. (2023b) studied the use of ViLD to perceive visual information to generate scene descriptions, and used text descriptions as clues to inform LLM to achieve adaptive robot grasping task planning. The shortcomings of these studies lie in the fact that spatial-based task understanding has not been achieved. Future research could focus on enhancing the processing of environmental spatial information to achieve a more comprehensive task understanding and planning.

4.1.2 Multimodal task information fusion

The fusion and alignment of the multimodal task information is essential for the VLM to successfully comprehend and break down the overall human–robot interactive task. The core of multimodal information fusion lies in how to capture the associations between different modalities in a unified representation space, and leverage the complementary information to provide a more comprehensive semantic understanding for subsequent task planning. Fan and Zheng (2024) aimed at a human–robot collaborative assembly task, in which they leveraged the CLIP model to parse visual information and LLM to comprehend language instructions and generated a feasible robot action sequence accordingly. Similarly, Song et al. (2024) adopted YOLO and VLM to analyze and reason about current social interactions, and generated immediate optimal robot actions to guide the motion planner. In addition, Rana et al. (2023) generated semantic graphs based on three-dimensional scene graphs, retrieved task-related semantic subgraphs through LLM, and then performed subtask planning. Gu et al. (2023) converted the real scene graph into a 3D conceptual scene structure graph, and then converted it into a text description and provided it to LLM, and used LLM to complete task planning. The above studies focus on utilizing visual and semantic interactions. For robots, the tactile modality can make a contribution to understanding task requirements better, facilitating task planning.

4.1.3 Action sequence generation

After the task decomposition and multimodal information fusion, corresponding action sequences should be generated to fulfil the designated task. In this process, the vision and language model can generate reasonable action sequences directly from vision and language inputs through end-to-end VLMs (Hu et al., 2023; Zhang et al., 2024; Skreta et al., 2024). Action sequence generation

Table 3 VLMs/LLMs in task planning

Method	Model	Success rate	Application	Year	Ref.
Matcha	ViLD, GPT-3	90.57% (custom data set)	Table-top manipulation task	2023	Zhao et al. (2023b)
TaPA	GPT-3.5, CLIP Mask RCNN	61.11% (custom data set)	Indoor embodied task	2023	Wu et al. (2023)
SayPlan	GPT-4	86.6% (Gibson)	Indoor navigation	2023	Rana et al. (2023)
VILA	GPT-4V	84.4% (custom task)	Table-top manipulation	2023	Hu et al. (2023)
ConceptGraphs	GPT-4, CLIP, DINO	97% (Replica)	Indoor navigation	2023	Gu et al. (2023)
LLM-Planner	GPT-3, object detector	51% (ALFRED)	Indoor embodied task	2023	Song et al. (2023)
\	CLIP, GPT-4	93.30% (custom data set)	human–robot collaboration assembly	2024	Fan and Zheng (2024)
NaVid	BERT, EVA-CLIP, Vicuna-7B	92% (VLN-CE R2R)	Indoor navigation	2024	Zhang et al. (2024b)
GameVLM	YOLOWorld	83.30% (custom data set)	Table-top manipulation	2024	Mei et al. (2024)
\	PaLM-E, LAVA	92% (RT-1)	Table-top manipulation	2024	Du et al. (2023)
VLM-SocialNav	YOLO, GPT-4V	100% (SCAND)	Indoor navigation	2024	Song et al. (2024)
REPLAN	GPT-4V	86.25% (RC)	Table-top manipulation	2024	Skreta et al. (2024)
PaLM-E	PaLM	82.5% (OK-VQA)	Embodied task	2023	Driess et al. (2023)
RT-2	PaLM-E\PaLI-X	62% (custom data set)	Embodied task	2023	Zitkovich et al. (2023)
\	GPT-3.5, resnet-50, BERT	85% (custom data set)	Assembly task, pick and place.	2024	Zheng et al. (2024)

needs to consider the context of the task, the dynamic changes of the environment, and the path to achieving the task goals. In this way, the system is able to generate coherent and executable action steps, improving the efficiency and reliability of task execution. In addition, researchers from Google have conducted a series of work to explore the perception-action end-to-end robot model, including the development of a general and transferable multi-decision agent by mixing specific data into the input of the multimodal LLM (Driess et al., 2023), and reformulating the LLM to directly output robot action parameters in language format (Zitkovich et al., 2023), which requires a lot of computing resources because the novel LLM structure needs to be fully trained.

4.1.4 Long-horizon task planning

Long-horizon task planning refers to task planning over a longer time span. This type of planning involves multiple steps and stages, and requires consideration of the long-horizon goals of the task and the coordination of intermediate steps. Unlike short-term task planning, long-horizon task planning needs to deal with more uncertainty and complexity, and usually requires more advanced strategies and a more comprehensive environmental understanding. Wu et al. (2023) fine-tuned the LLaMA network through a triple of visual scenes, instructions, and corresponding plans, and used it as a fine-tuned task planner to achieve long-horizon planning for complex tasks. Mei et al. (2024) proposed a decision and expert agent system based on VLM, in which the decision agent is responsible for planning tasks, and the expert agent evaluates these task plans and resolves inconsistencies between different agents by introducing zero-sum game theory to determine the optimal solution. Du et al. (2023) took long-horizon task instructions and current image observations as input and output detailed multimodal (video and language) long-horizon video plans. The program scales with increasing computational budgets and can be synthesized across different robotics domains, from multi-object rearrangement to multi-camera dual-arm dexterous manipulation, providing better task planning results.

A key challenge for VLM-based task planning is its applicability in dynamic environments. Models such as TaPA and GameVLM have shown high success rates in long-term task planning by accurately encoding the recognition features of static scenes. Du et al. generated real-time long-term planning results based on current image observations, which reflects the trend of combining real-time scene parsing with dynamic planning. This capability is key to realizing industrial scene applications, because in industrial scenes, as production tasks proceed dynamically, robot planning needs to be constantly updated in real-time according to the scene. (Challenge 6.2).

4.2 Vision and language navigation

Vision and language navigation (VLN) refers to a robot's ability to complete navigation tasks based on visual perception of the environment and comprehension of human natural language commands. In this task, the agent is typically provided with the following information: 1) Environmental representation: the agent's visual perception in a 3D space, including images, point clouds, and other forms of environmental representation; 2) Natural language instructions: a text description of how the agent should move to reach the target location; 3) Initial position: the agent's starting point in the environment. The agent's task is to navigate the environment and reach the target location based on the natural language instructions. This process requires the agent to parse and comprehend each step of the instructions, match them with the visual perception of the environment, plan a path, and take actions to complete the navigation task as instructed. VLMs and LLMs play a crucial role in VLN tasks as they can comprehend vision and language cues to significantly enhance the agent's navigation performance in complex environments. Based on different navigation scenarios, we categorise these works into indoor navigation, outdoor navigation, and web navigation. The general approach of VLM-based VLN is shown in Fig. 6. VLMs are often employed as visual encoders to capture semantic information from visual observations, while

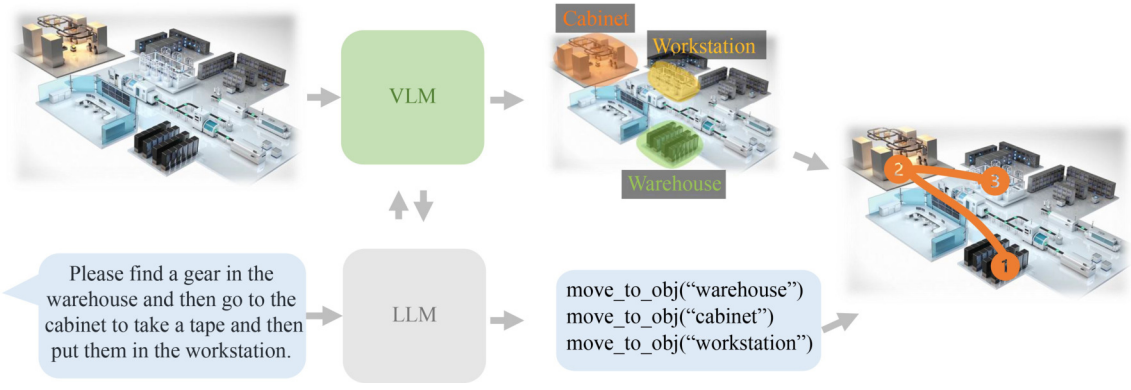

Fig. 6 General framework of vision and language navigation.

LLMs are frequently utilized to understand human language commands and subsequent reasoning processes. The recent representative works using VLMs or LLMs in navigation tasks are summarized in Table 4.

4.2.1 VLM-based navigation in an indoor environment

Indoor navigation is widely applied in the field of domestic robotics, with some work also involving human–robot collaboration in industrial settings. Indoor VLN can assist with assembly, help humans retrieve and place tools, manage logistics distribution within the factory premises, and inspect and maintain equipment.

In VLN, the most common tasks are simple instruction-following tasks, in which a robot or virtual agent receives detailed, step-by-step natural language instructions and navigates through the environment to a specified target location based on these instructions. These instructions are typically clear and specific, describing each step of the action path. Khandelwal et al. (2022) explored CLIP's capabilities in embodied AI, discovering that CLIP's feature representations encode these primitives more effectively than ImageNet-pretrained backbones. They proposed the EmbCLIP model, which achieved a 47% task success rate in the RoboTHOR OBJECTNAV Challenge 2021. Korekata et al. (2023) proposed a model called SheFU, consisting of switching image embedder and funnel transformer. This model could predict the target object and destination separately and achieve a task success rate of 83% in the ALFRED-fc data set. Hong et al. (2023) presented Ego2-Map based on the ViT-B/16 model from CLIP for VLN in continuous environments and could reach a 47% success rate in the R2R-CE data set.

Apart from simple instruction-following tasks, some works focus on remote embodied referring expression tasks, which means the instructor only gives high-level instructions, and the robot needs to figure out the clear destination by continuously asking questions according to the current state. Huang et al. (2023a) presented VLMaps based on LSeg and GPT-3.5, in which they utilized GPT-3.5 to generate Python code for robot navigation, and the method reached a 62% success rate in their custom data set. Qiao et al. (2023) utilized CLIP as a scene perceiver and GPT-2 as an in-context learning model to build the MiC model for remote embodied referring expression task, which achieved a task success rate of 55.74% in the REVERIE data set. Similarly, Gao et al. (2024) introduced

Table 4 VLMs/LLMs in navigation

Navigation environment	Method	VLM/LLM	Success rate	Application	Year	Ref.
Indoor	EmbCLIP	CLIP	47% (RoboTHOR OBJECTNAV Challenge 2021)	Domestic service robot	2022	Khandelwal et al. (2022)
	SHeFU	Switching Image Embedder, Funnel Transformer	83.1% (ALFRED-fc)	Domestic service robot	2023	Korekata et al. (2023)
	Ego2-Map	CLIP	47% (R2R-CE)	Domestic service robot	2023	Hong et al. (2023)
	VLMaps	LSeg, GPT-3.5	62% (custom dataset)	Domestic service robot	2023	Huang et al. (2023a)
	MiC	GPT-2, CLIP	55.74% (REVERIE)	Remote embodied referring expression, domestic service robot	2023	Qiao et al. (2023)
	CKR+	CLIP	23.13% (REVERIE)	Remote embodied referring expression, domestic service robot	2024	Gao et al. (2024)
	NavGPT	GPT-3.5, BLIP-2	34% (R2R)	Domestic service robot	2024	Zhou et al. (2024)
	LANA+	CLIP	70.1% (R2R)	Instruction following and generation, domestic service robot	2024	Wang et al. (2024d)
	ACK	CLIP	59.1% (REVERIE)	Remote embodied referring expression, domestic service robot	2024	Mohammadi et al. (2024)
	CONSOLE	ChatGPT, CLIP	72% (R2R)	Domestic service robot	2024	Lin et al. (2024)
	DISH	CLIP	44.3% (R2R)	Domestic service robot	2024	Wang et al. (2024a)
	A vision AI-based HRC assembly approach	GPT-4	\	Human–robot collaboration	2024	Liu et al. (2024)
	A vision and language cobot navigation approach	GPT-3.5	\	Human–robot collaboration	2024	Wang et al. (2024b)
Outdoor	LM-Nav	CLIP, GPT-3	80% (custom dataset)	Urban VLN	2022	Shah et al. (2022)
	VELMA	CLIP, GPT-4	23.1% (Map2seq)	Urban VLN	2024	Schumann et al. (2024)
	VLN-VIDEO	GPT-2	31.7% (Touchdown)	Urban VLN	2024	Li et al. (2024)
Website	WebVLN	GPT-3.5, BLIP-2	34.76% (WebVLN-v1)	Web Navigation and Question–Answering	2024	Chen et al. (2024)

CKR + based on CLIP, which could achieve a 23.13% success rate in the REVERIE data set. Wang et al. (2024d) devised a CLIP-based LANA + model which could mimic the human process of finding a path through iterative question-and-answer interactions and reached a success rate of 70.1% in the R2R data set. Mohammadi et al. (2024) introduced the ACK framework, which utilizes commonsense information structure as a spatio-temporal knowledge graph to enhance agent navigation. Within this framework, the CLIP model was employed to gather and prioritise the most relevant knowledge concerning the scene and identified objects. In the REVERIE data set, ACK obtained a 59.1% task success rate.

Many works also aim to explore the reasoning capabilities of LLMs and apply them to task planning in VLN. These research efforts aim to harness the sophisticated natural language understanding and generation abilities of LLMs to improve the decision-making processes involved in navigating complex environments. By integrating LLMs with visual information, researchers seek to develop advanced navigation systems capable of interpreting and responding to dynamic scenarios. Zhou et al. (2024) developed a purely LLM-based VLN system based on GPT-3.5 and BLIP-2. This work innovatively explored GPT's zero-shot reasoning capabilities in complex environments and achieved a 34% task success rate on the R2R data set without any training. Similarly, Lin et al. (2024) proposed CONSOLE based on ChatGPT and CLIP, and this model gained a 72% success rate in the R2R data set.

Furthermore, VLMs can be leveraged to formulate strategies for robot learning in VLN tasks. This involves using VLMs to generate and optimise action plans that guide robots through various navigation challenges. The integration of VLMs in robot learning not only enhances the robots' ability to understand and process visual and linguistic cues but also improves their adaptability and performance in real-world applications. Wang et al. (2024a) introduced an RL framework for discovering intrinsic subgoals via hierarchical (DISH) RL in which CLIP's image encoder was applied for visual feature extraction. This method overcame the label annotation problem in reinforcement learning and achieved a 44.3% task success rate in the R2R data set.

While models like EmbCLIP and SheFU demonstrate high task success rates by effectively encoding visual features and predicting target locations, approaches such as VLMaps and MiC emphasize interactive question-asking to improve task completion in more ambiguous scenarios. This highlights a common trend toward integrating visual perception with dynamic linguistic interaction, although the varying success rates underline the ongoing challenge of achieving consistency across different data sets and environments.

In the industrial sector, some works have also introduced LLM-based VLN for HRC tasks. Liu et al. (2024) utilized GPT-4 along with 3D object reconstruction and pose estimation methods in human–robot collaborative assembly tasks, while Wang et al. (2024b) built a cobot navigation framework using GPT-3.5 combined with other visual methods to assist operators in tool retrieval and placement in HRC. However, while these advancements showcase promising progress, the variability in success rates across different models and data sets highlights a critical challenge in performance consistency, indicating a need for improved generalisation and adaptability in dynamic real-world environments (Challenge 6.3).

4.2.2 VLM-based navigation in outdoor environment

Compared to indoor VLN, the outdoor VLN environment is typically more open and expansive, with fewer constraints on movement. Outdoor landscapes can include a variety of terrains, weather conditions, and lighting variations. The presence of dynamic elements such as vehicles and pedestrians adds to the complexity. The applications of outdoor VLN include autonomous driving, robotic delivery, smart city management, rescue operations, and environmental monitoring. In the industrial sector, outdoor VLN can also be employed for factory security patrols and logistics distribution between factory sites. Shah et al. (2022) presented a system called LM-Nav based on CLIP and GPT-3 for long-horizon navigation through outdoor and complex environments and achieved an 80% success rate in their custom data set. Similarly, Schumann et al. (2024) proposed VELMA, a model based on CLIP and GPT-4, which uses verbal descriptions of trajectories and visual environment observations as contextual prompts for the next action, and they achieved a task success rate of 23.10% in the Map2seq data set. Li et al. (2024) introduced VLN-VIDEO, utilizing urban driving videos combined with automatically generated navigation instructions and actions to enhance outdoor VLN performance. GPT-2 was used to filter out templates with low generation probability. They gained a 31.7% success rate in the Touchdown data set. Despite these advancements, the significant variation in success rates indicates that outdoor VLN systems must address the challenges of diverse environmental conditions and dynamic elements more effectively to achieve reliable performance in real-world applications (Challenge 6.3).

4.2.3 VLM-based navigation in web environment

Apart from VLN in real-world environments, Chen et al. (2024) proposed a VLN model for web environments. They utilized GPT-3.5 and BLIP-2 to construct a WebVLN model capable of answering questions based on web content and automatically navigating to the required web pages. They also created a WebVLN-v1 data set, where the WebVLN model achieved a task completion rate of 34.7%. This is a newly proposed task

that has not yet been explored in the industrial sector, but it has potential applications in industrial QA systems.

4.3 Vision and language manipulation

VLMs have also been frequently leveraged in robotic manipulation tasks to enable robots to perform physical tasks by parsing and understanding visual inputs (such as images or videos) and language inputs (such as instructions or descriptions). This type of interaction enables robots to execute more flexible and adaptive operations in complex environments. The methodology in manipulation is similar to navigation; it also requires environmental representation, natural language instructions, and the robot's initial state. VLMs and LLMs can accomplish joint reasoning for visual perception and language comprehension, while also facilitating further trajectory planning. Recent works about VLMs/LLMs in robotic manipulation are provided in Table 5.

For vision and language manipulation algorithms, current approaches primarily utilize VLMs for scene understanding and LLMs for natural language command comprehension, combined with planning algorithms to generate trajectory key points or dense waypoints for the end-effector, as illustrated in Fig. 7. These methods can be categorised into two types: one involves designing complex prompts for VLMs and LLMs to directly generate trajectories, referred to as *planning-based vision and language manipulation*, and the other employs VLMs and LLMs to assist in policy generation within robot learning to accomplish manipulation tasks, known as *learning-based vision and language manipulation*. Regarding application scenarios, most work applying VLMs/LLMs to robotic manipulation focuses on tabletop household tasks, with some studies extending their use to industrial settings.

4.3.1 Planning-based vision and language manipulation for tabletop household tasks

Planning-based vision and language manipulation is an approach in robotic manipulation where VLMs and

Table 5 VLMs/LLMs in manipulation

Method	VLM/LLM	Success rate	Application	Year	Ref.
CLIP-SemFeat	CLIP	58.8% (LVIS)	Scene rearrangement, tabletop household tasks	2022	Goodwin et al. (2022)
Act3D	CLIP ResNet50	83% (RLBENCH)	Tabletop household tasks	2023	Gervet et al. (2023)
CALAMARI	CLIP	84% (RLBENCH)	Contact-rich manipulation, tabletop household tasks	2023	Wi et al. (2023)
GNFactor	Stable diffusion, CLIP	31.7% (RLBENCH)	Tabletop household tasks	2023	Ze et al. (2023)
GVCCI	Grounding entity transformer (Get), Setting entity transformer (Set)	50.98% (VGPI)	Pick and place task, tabletop household tasks	2023	Kim et al. (2023a)
MOO	Owl-ViT	~50% (RT-1)	Tabletop household tasks	2023	Stone et al. (2023)
PaLM-E	PaLM-E	66.1% (OK-VQA)	Tabletop household tasks	2023	Driess et al. (2023)
SMS	ViLD, BLIP-2, SAM	41.7% (custom dataset)	Mechanical search, tabletop household tasks	2023	Sharma et al. (2023)
Voxposer	GPT-4, OWL-ViT, SAM	88% (RLBENCH)	Tabletop household tasks	2023	Huang et al. (2023b)
Vision-language guided robotic planning	CLIP, GPT-4	93.3% (custom dataset)	Human–robot collaboration	2024	Fan and Zheng (2024)

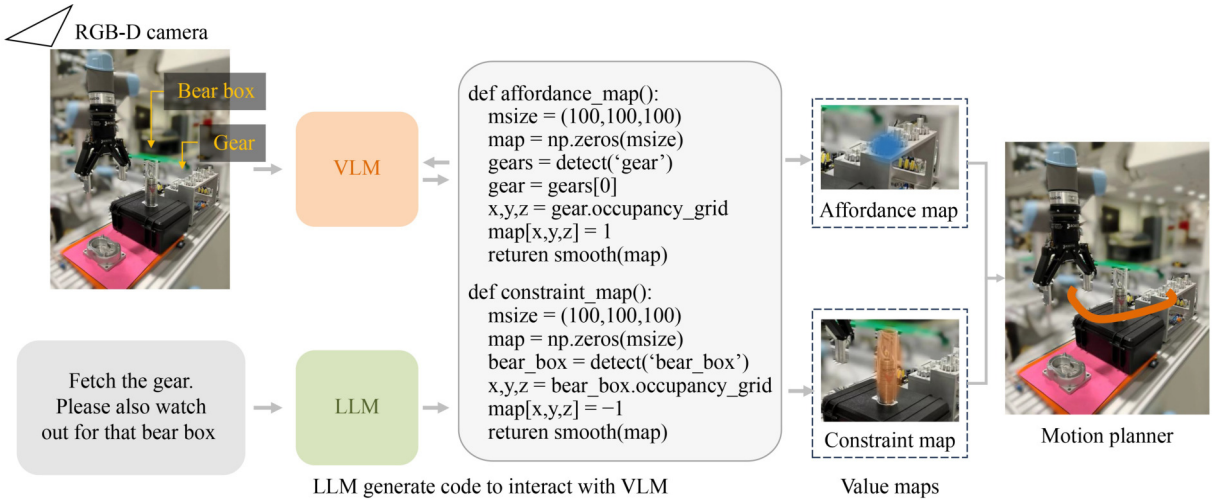

Fig. 7 General approach of vision and language manipulation adapted with permission from Huang et al. (2023b).

LLMs are utilized to generate precise movement trajectories for robotic tasks. This method involves the design of complex natural language prompts that instruct LLMs to produce detailed paths that the robot's end-effector should follow. Goodwin et al. (2022) proposed a novel method called CLIP-SemFeat for scene rearrangement in Tabletop household tasks by leveraging CLIP to solve cross-instance matching problems, and have reached a 58.5% task success rate in the LVIS data set. Driess et al. (2023) introduced a VLM for robot manipulation called PaLM-E, which can translate knowledge from the vision-language domain into concrete reasoning, enabling robots to plan in environments with complex dynamics and physical constraints and to answer questions about the observable world. PaLM-E could achieve a 66.1% task success rate in the OK-VQA data set. Sharma et al. (2023) presented SME consisting of ViLD, BLIP-2 and SAM for mechanic searching for tabletop household tasks, and could reach a 41.7% task success rate in the custom data set. Huang et al. (2023b) proposed a novel framework named Voxposer for model-based planning. This framework utilized ViLD for object detection, SAM for semantic segmentation, and GPT-4 for reasoning and planning. VoxPoser's planning precision reached the waypoint level of robotic arm movement trajectories, making manipulation tasks highly flexible. It achieved an 88% task success rate in the RLBENCH tasks. While these approaches demonstrate significant progress in household settings, the current methods often lack the precision required for complex industrial tasks, underscoring the need for further research to enhance motion planning precision and extend their applicability to industrial environments (Challenge 6.4).

4.3.2 Learning-based vision and language manipulation for tabletop household tasks

Learning-based vision and language manipulation is an approach where VLMs and LLMs are used to assist in generating policies for robot learning. Instead of directly generating trajectories, this method focuses on developing policies that enable robots to learn and perform manipulation tasks through experience and interaction with the environment. Gervet et al. (2023) proposed Act3D, utilizing CLIP ResNet50 and leveraging their shared vision-language feature space to interpret instructions and foundational references for learning-from-demonstration tasks. They achieved an average success rate of 83% in RLBENCH tasks. Wi et al. (2023) introduced CALAMARI for contact manipulation in household environments based on CLIP and obtained a success rate of 90% in the *wipe_desk* task, 84% in the *sweep_to_dustpan* task, and 60% in the *push_putton* task in the RLBENCH environment. Ze et al. (2023) presented a GNFactor model based on stable diffusion and CLIP for imitation learning in robot manipulation tasks and gained an average success rate of 31.7% in 10 tasks in RLBENCH. Kim et al. (2023a) proposed GVCCI consisting of Grounding Entity Transformer (Get) and Setting Entity Transformer (Set) for pick and place actions, and obtained a task success rate of 50.98% in Visual Grounding on the Pick-and-place Instruction (VGPI) data set. Stone et al. (2023) introduced MOO in which Owl-ViT is applied to object-identifying information from the language command and image. They have reached about a 50% task success rate in the RT-1 data set. This section focuses on the task of manipulation, with more detailed discussions on learning-based methods to be presented in Section 5. Despite notable advancements, the variability in task success rates across different models and tasks indicates that these learning-based approaches need further refinement to enhance adaptability and precision, particularly for more complex and varied manipulation tasks beyond household environments (Challenge 6.4).

4.3.3 Vision and language manipulation for industrial tasks

It is found that most works regarding vision-language robotic manipulation are based on RLBENCH and thus only consider household tasks. The investigations of VLM-based manipulation are still quite rare in the industrial sector, and current applications have been limited to simple planning-based methods. Fan and Zheng (2024) innovatively introduced a VLM-based manipulation method in HRC tasks. They proposed a vision-language-guided robotic planning approach for robotic action planning for HRC assembly, and have reached a 93.3% success rate in their custom industrial data set. In the industrial sector, VLM-based manipulation, in addition to completing HRC assembly tasks, has potential value in warehouse management, equipment maintenance, flexible production line adjustments, and item sorting, and is worth further exploration.

It is important to note that while planning-based methods like CLIP-SemFeat and Voxposer excel in task success rates for household tasks, they often struggle with the precision required for industrial applications. In contrast, learning-based approaches such as Act3D and CALAMARI show promise in adaptability through learning from interactions but display variability in success rates across different tasks. Identifying these trends and common challenges suggests a potential for integrating the strengths of both approaches to improve precision and adaptability, particularly in transitioning from household to industrial applications. This synthesis highlights a critical trend toward developing hybrid models that can bridge current gaps, offering a more robust solution across varied environments. The adaptation of VLM-based manipulation from household to industrial applications faces challenges such as the need for higher precision, robustness in diverse conditions, and integration

with existing industrial systems, indicating significant opportunities for future research to bridge these gaps and fully realize its potential in industrial settings (Challenge 6.4).

5 VLM-based human-guided skill transfer and robot learning

Another crucial challenge in human–robot collaborative manufacturing is to enable robots to effectively learn new skills and actions, especially in unstructured and dynamic environments where pre-programmed movements are insufficient. One potential solution is to allow robots to acquire new skills from human demonstrations (Yin et al., 2024). Learning from demonstration, or imitation learning, aims to enable robots to extract human behaviors and decision-making processes. This encompasses a range of competencies, including operational skills, sequence planning, adaptability to different scenarios, and even the ability to handle uncertainty and weigh pros and cons.

Compared to traditional algorithms, VLM-based skill transfer has attracted increasing attention due to its powerful in-context learning capabilities. VLM not only integrates the efficient feature representation capabilities of both vision and language, but also, due to its pre-training on large-scale data, possesses strong generalisation ability. This allows it to perform exceptionally well in new tasks and environments, reducing the need for extensive demonstration data. Hence, the advancement of VLMs has enabled robots to exhibit enhanced logical reasoning and compositional generalisation abilities in manufacturing tasks.

In this section, the latest progress in VLM-based human-guided skill transfer and robot learning are elaborated, as illustrated in Fig. 8. Our focus is primarily on literature that leverages both vision and language modalities to collect human demonstrations and facilitate skill learning. The main content is divided into two parts: 1) high-quality human demonstration gathering approaches and 2) VLM-based robot learning algorithms and architectures.

5.1 VLM-based human demonstration collection

Obtaining human demonstrations has long been a crucial research focus in enabling robots to acquire human skills. The integration of multiple modalities, such as vision and language, cannot only help robots better comprehend human demonstrations but also enable them to learn skills more robustly in partially observable environments. Figure 9 showcases the primary demonstration methods, highlighting the innovative approaches to human teaching enabled by vision language integration. As shown in Table 6, the main benefits of including language in the demonstration process have been summarized through related works. Specifically, compared to passive video observation, integrating the language modality can align more closely with human cognition, improve comprehension and generalisation capabilities, extract explicit rules, enhance human–robot interaction, and support the learning of more complex manipulation skills.

5.1.1 Align with human cognition

Humans typically learn new skills by combining the observation of demonstrated actions with listening to related verbal explanations. Thus, integrating video with textual or spoken information can more closely mimic human learning processes. Azagra et al. (2020) designed an incremental learning pipeline. Through natural user interactions (such as pointing, showing, and verbal descriptions), the robot utilizes an RGB camera to capture image data and combines skeleton detection and speech recognition technologies to incrementally learn and update object models. This enables robots to gradually learn to accurately identify and understand different objects in dynamic and diverse interactive environments, just like humans. Ding et al. (2023) proposed a system called Embodied Concept Learner (ECL), enabling robots to emulate human capabilities in learning visual concepts and understanding geometric mapping within an interactive 3D environment. The robotic agent, akin to a baby, can acquire long-term, interpretable, unsupervised semantic and depth learning through interactions with humans.

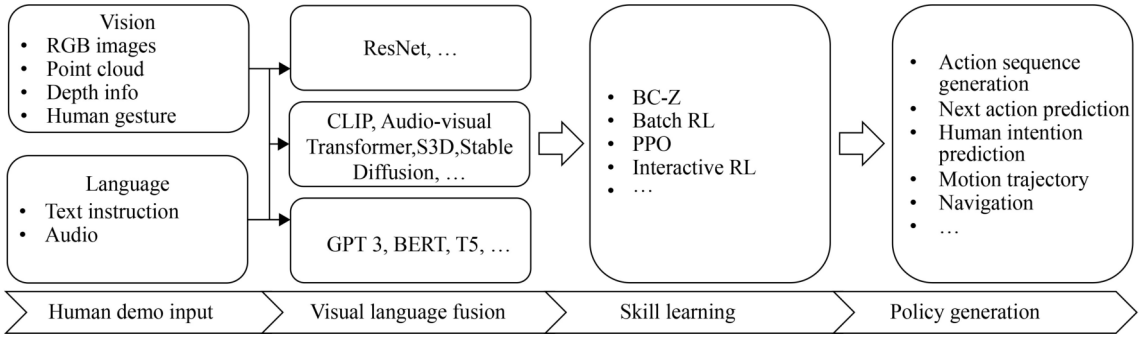

Fig. 8 VLM-based human-guided skill transfer and robot learning.

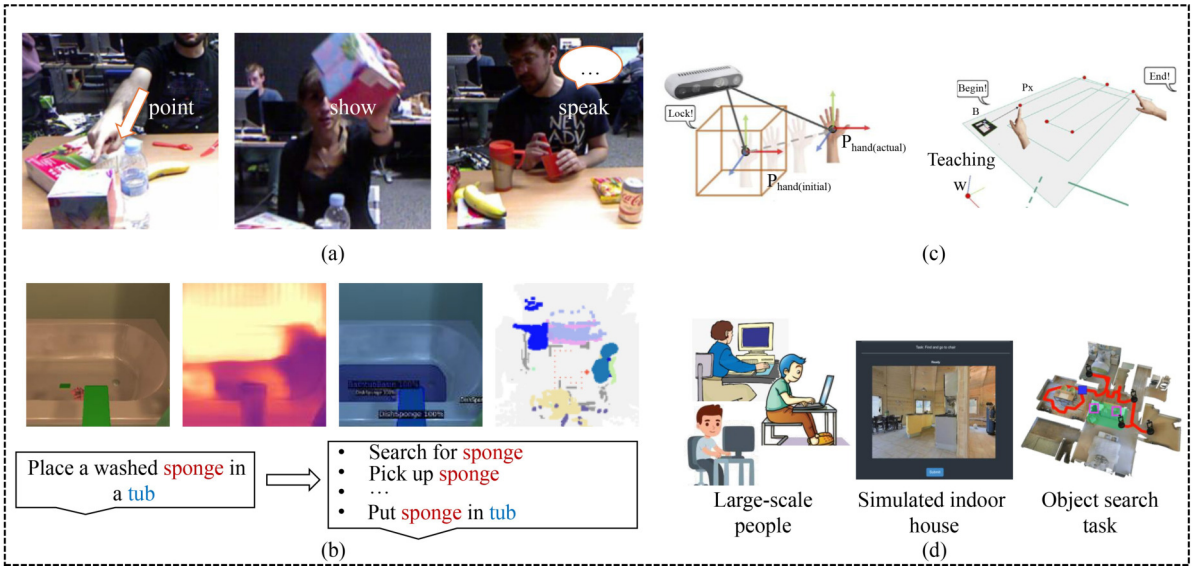

Fig. 9 VLM-based human demonstration collection: (a) Human-like teaching methods (Azagra et al., 2020), including point, show, and speak; (b) Vision-language fusion and goal parsing (Ding et al., 2023), to assist robot in learning concepts through rich information and self-supervised learning; (c) Gesture-based teaching and teleoperation methods (Halim et al., 2022), to provide natural interactions for direct programming of the robot; (d) human demonstration at scale (Ramrakhya et al., 2022), to learn the skill preferences of object navigation from different human agents.

Table 6 VLM-based human demonstration collection

Category	Description	Demonstration method	Tasks/Application	Ref.
Align more closely with human cognition	Incremental Learning of Object Models from interactive human demonstration	RGB camera, microphone	Recognise and understand object	Azagra et al. (2020)
	Self-supervised learning of concepts and mapping through instruction following in an interactive 3D environment	RGB observation, Language command	Routine manipulation tasks	Ding et al. (2023)
Improve generalisation capability	Improve robot command generation accuracy	Video, Key frame caption	Routine manipulation tasks	Yin and Zhang (2023)
Extract implicit rules	A virtual teleoperation infrastructure to collect large-scale demonstrations	RGBD camera, GPS+compass sensor, Task instruction	Navigation, pick and place tasks	Ramrakhya et al. (2022)
	Behaviour-informed state abstractions via language model queries to capture human task-relevant preferences	Video, language specification	Tabletop manipulation tasks	Peng et al. (2024)
Enhance human–robot interaction	Multimodal no-code robotic programming	RGB-D camera, Microphone	Line movement, zigzag movement, contour movement, etc.	Halim et al. (2022)
	Multimodal demonstration interface	3D camera, speech, hand guiding	Pick and place tasks	Lu et al. (2022)
Support more complex manipulation skill learning	Multimodal human demonstration collection system	Motion capture, depth camera, video recording, speech	Complex tool manipulation skills	Shukla et al. (2023)

5.1.2 Improve generalisation capability

In learning from human demonstration, purely video-based passive observation may lead robots to struggle to capture the underlying intentions and semantics of actions. In contrast, language information can provide crucial contextual details, aiding robots in better understanding the purpose and logic of demonstrations, thereby enhancing learning accuracy and generalisation capabilities. For instance, Yin and Zhang (2023) presented a framework that integrates key frame extraction with multimodal information (text caption) fusion, significantly improving the accuracy of robot command generation. Therefore, when integrated with an affordance detection network and a motion planner, this framework enables robots to effectively reproduce the tasks demonstrated.

5.1.3 Extract implicit rules

When robots learn from human demonstrations, they can acquire implicit rules and subtle habits that are difficult to explicitly abstract into clear guidelines, unlike autonomous learning methods such as reinforcement learning (RL). By integrating visual and linguistic inputs, robots are better equipped to uncover these latent policy patterns. Ramrakhya et al. (2022) discovered that imitation

learning (IL) enables robots to learn effective object navigation skills from humans—such as peeking into rooms, checking corners for small objects, and turning around to get a panoramic view. This surpasses the limitations of traditional reinforcement learning (RL) methods, which require intricate reward engineering to induce such behaviors. Peng et al. (2024) proposed a Preference-Conditioned Language-Guided Abstraction (PLGA) method, using language model (LM) queries to capture implicit human preferences in tasks. For instance, observing behavior changes like avoiding electronics during "throwing away a jar" allows the robot to infer and train strategies based on these abstracted preferences. Notably, robots may inadvertently replicate undesirable human biases when learning implicit rules, an issue that warrants careful consideration in future research.

5.1.4 Enhance human–robot interaction

Compared to passive video observation, leveraging robot teleoperation to gather demonstration data is a more accurate and efficient approach. This natural teaching method, which combines action demonstration and language interpretation, leads to more intuitive and effective human–robot interaction. By grounding the learning process in multimodal interactions, Halim et al. (2022) introduced a no-code robotic programming system designed for beginners. This approach utilizes a visual system that enables users to convey spatial information, including 3D points, lines, and trajectories, through hand and finger gestures. Additionally, a speech recognition system aids users in setting robot parameters and engaging with the robot's state machine. Lu et al. (2022) also proposed a multimodal demonstration system integrating natural language instruction, visual observation, and hand guiding to let robot learn task comprising goal concepts, task plans, and basic actions, which can be applied to pick-and-place tasks. However, some interaction approaches may introduce additional implementation costs and system complexity. It is essential to consider the scalability and practicality of interaction methods to ensure the applicability across broader real-world scenarios.

5.1.5 Support more complex manipulation skill learning

Acquiring complex manipulation skills often necessitates the use of multimodal information. Humans not only grasp the actions required for a skill, but also understand the various states, state transitions, and constraints associated with the task. Shukla et al. (2023) introduced a multimodal framework supporting data collection from various modalities, including speech, gestures, motion, video, and 3D depth data. This framework integrates visual and linguistic data to gather rich human demonstration data for learning intricate tool operation skills, such as granular media transport tasks. It is evident that, although visual and linguistic cues have been widely utilized to facilitate human–robot skill transfer, the integration of additional multimodal sensory data remains underexplored (Challenge 6.5).

5.2 VLM-based robot learning from human demonstrations

In this section, we delve into the details of the learning mechanisms employed by researchers utilizing the rich information from vision-language integrated human demonstration data. This exploration encompasses a spectrum of encoding, decoding, and learning strategies. The learning process is systematically divided into three key areas: vision-language fusion, the combination of imitation and reinforcement learning, and enhancements in adaptability. Figure 10 illustrates a typical application for each of these three subdomains. Table 7 provides a detailed account of the learning methods, training data sets, and evaluation metrics utilized in the reviewed articles.

5.2.1 Vision-language fusion and multimodal learning

Vision-language fusion and multimodal learning offer significant advantages in the field of robotic manipulation and human–robot interaction. By integrating visual and linguistic information, these approaches enable robots to understand and execute complex tasks with greater accuracy and flexibility. The fusion of multiple modalities allows for a richer and more comprehensive understanding of the environment, facilitating more robust decision-making processes.

For example, Shao et al. (2021) presented a new imitation learning framework combining natural language instructions and visual inputs for enhanced robot manipulation, offering improved generalisation, efficient learning, and handling of complex manipulations. Evaluation results show that the multi-task strategy performs well in terms of success rate and handling complex tasks. Wang et al. (2022b) put forward a novel Teaching-Learning-Prediction (TLP) framework that enables robots to predict human intentions in hand-over tasks using multimodal data, including natural language and wearable sensor inputs. Leveraging the extreme learning machine (ELM) algorithm, the TLP framework significantly improves prediction accuracy and stability compared to traditional methods, facilitating more efficient human–robot collaboration. Hori et al. (2023) put forward a multimodal imitation learning approach that leverages videos, audio, and text to generate robot action sequences using an audio-visual Transformer (AVTransformer). This method integrates dynamic motion primitives (DMPs) and style transfer

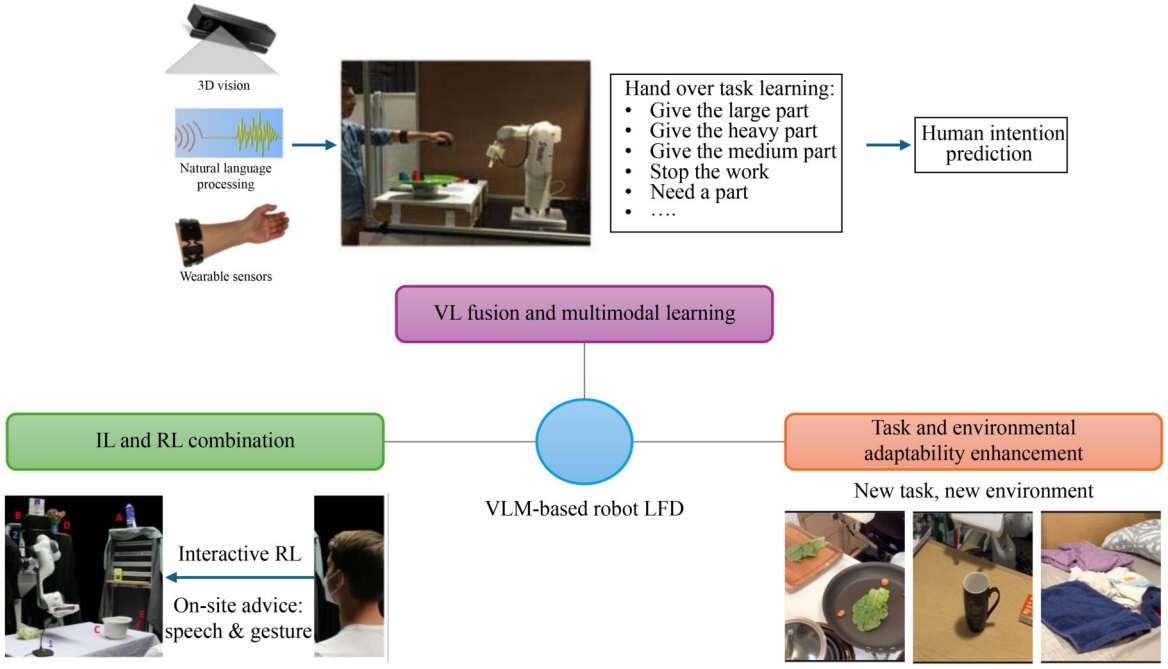

Fig. 10 VLM-based robot learning from human demonstrations, which is primarily divided into three sections: Vision language fusion and multimodal learning (Wang et al., 2022b), IL and RL combination (Trick et al., 2022), and task and environmental adaptability enhancement (Nair et al., 2022).

Table 7 VLM-based robot learning from human demonstrations

	Description	Method	Training data	Evaluation	Ref.
Multimodal data fusion	Learn manipulation concept	BERT, ResNet-18, DDPG, Batch RL	78 tasks from "Something-Something" dataset (video, task description)	Multi-task success rate 76.3%	Shao et al. (2021)
	Predicting human intentions	TLP model, Extreme Learning Machine	25000 Multimodal data from five participants	Prediction accuracy 98.5%, time efficiency improved	Wang et al. (2022b)
	Action sequence acquisition	AVTransformer, CLIP	Epic-Kitchen-100, YouCookII, QuerYD, and in-house instruction video datasets	Success rate 32 %, improves DMP sequence quality by 2.3 times	Hori et al. (2023)
	CSATO multi-task skill learning	Transformer architecture	RLBench including RGB image, depth map, instruction, action sequence	Success rates of 90.7% and 83.9% in single-task and multi-task scenarios	Han et al. (2024)
IL and RL combination	Zero-shot task generalisation	BC-Z	VR teleoperation data, human video, task language strings	24 unseen manipulation tasks with 44% success rate	Jang et al. (2022)
	Interactive reinforcement learning	MIA-IRL	Train the classifier with collected speech and gesture	Convergence time, success rate, and robustness are significantly higher.	Trick et al. (2022)
	RoboCLIP: one demo to learn policies	S3D VLM, RL (PPO)	HowTo100M to pretrain VLM, MetaWorld for simulation	2–3 times higher zero-shot performance than competing IL methods	Sontakke et al. (2024)
Adaptability enhancement	R3M: pretrained visual representation	ResNet-18, ResNet-34	Egocentric 4D	Average success rate 56%	Nair et al. (2022)
	Learning Object Spatial Relationship	Spatial probability models, SVM, GPT3	NUSCENES dataset and tabletop scenes	Accuracy 91.3%	Yu et al. (2023)
	GNFactor multi-Task skill learning	Stable Diffusion, CLIP	10 tasks in RLBench	Success rate 31.7%	Ze et al. (2023)
	User directed hierarchical learning	CLIP, PerAct, GPT-4, Bard	RLBench including RGB-D, text instruction, voxel representation	13% improvement in task success rates	Winge et al. (2024)
	SHOWTELL	GPT-4,ViLD, BLIP-v2	Text instruction, RGB data, hand detection, hand-object interaction	Success rate over 85%, out-perform GPT4-V	Murray et al. (2024)

learning to enhance performance. CSATO algorithm (Han et al., 2024) employed a visual-language fusion network, a token reduction network, and a Transformer decoder to model correlations among instructions, current, and historial visual observations, generating autoregressive action predictions.

5.2.2 Combination of IL and RL

The combination of imitation learning and reinforcement learning offers powerful synergy, significantly enhancing task generalisation and zero-shot learning capabilities in robotic systems. Imitation learning provides a strong foundation by allowing robots to quickly acquire complex behaviors from human demonstrations, effectively reducing the initial learning curve and enabling rapid adaptation to new tasks. Reinforcement learning complements this by refining and optimising the learned behaviors through trial and error, guided by reward signals. By integrating the strengths of both approaches, robots can achieve robust task generalisation, applying learned strategies to a wide array of scenarios with minimal retraining.

Jang et al. (2022) combined a multilingual sentence encoder for vision-based robotic manipulation. Using the BC-Z model, it trains on a large data set of human demonstrations, achieving success in zero-shot task generalisation and unseen task performance. Trick et al. (2022) introduced an interactive reinforcement learning (IRL) approach that leverages the Bayesian fusion of multimodal advice. The proposed method, MIA-IRL, enables a robot to learn the pancake-making task from various initial states by incorporating human-provided multimodal guidance, such as speech and gestures. Experimental results demonstrate that the MIA-IRL approach achieves faster convergence and greater robustness in this task compared to existing methods. Sontakke et al. (2024) presented RoboCLIP, an innovative imitation learning method that uses pretrained VLMs to generate reward functions from a single demonstration (video or text). It reduces the reliance on expert demonstrations and complex reward engineering, leading to superior zero-shot performance and efficient fine-tuning. However, the challenges of integrating these methods remain significant, including the complexity of designing effective reward functions for various tasks in reinforcement learning, as well as ensuring data and training environment consistency between reinforcement and imitation learning.

5.2.3 Task and environmental adaptability enhancement

Other applications also explore the potential of VLM in learning spatial relationships to further improve task and environmental adaptability of robots. Specifically, Nair et al. (2022) presented R3M, a visual representation model pretrained using a combination of time-contrastive learning, video-language alignment, and L1 sparsity penalties, which significantly enhances data-efficient imitation learning for robotic manipulation tasks. By leveraging the diverse human video data set Ego4D, R3M outperforms state-of-the-art visual representations like CLIP and MoCo, demonstrating superior performance in unseen environments and tasks. Yu et al. (2023) proposed a method enabling robots to learn and recognize 3D spatial relationships between objects. By leveraging image and language demonstrations, the method constructs new spatial relationship probability distribution models, thereby allowing robots to execute tasks more accurately in complex environments. Regarding visual representation and encoding, Ze et al. (2023) introduced GNFactor, a visual BC agent for multi-task robotic manipulation concentrating on improved generalisation abilities. It leverages the Stable Diffusion model to encode semantic information into 3D voxel representations, enabling visual reconstruction and language-conditioned action prediction. Winge et al. (2024) proposed a hierarchical robot learning framework, in which end users can provide text instructions related to observation (spatial scenario). Combined with the VLM Bard's capability to automatically decompose complex tasks into intermediate skills, robots can effectively learn to execute high-level tasks by understanding the environment and building from basic actions. A highly modular neural-symbolic framework is also introduced by Murray et al. (2024), for synthesizing robotic skills from visual demonstrations and natural language instructions. Current methodologies often assume an idealised training environment, where objects are expected to maintain reasonable initial poses and training steps are standardised. However, when these algorithms are applied in real-world scenarios, they face a much broader spectrum of uncertainties, rendering them insufficient without human oversight. This reliance on continuous human intervention for monitoring and adjustment severely limits their practical implementation in real-world applications (Challenge 6.6).

6 Challenges and future perspectives

Although preliminary explorations of the application of VLMs have demonstrated impressive capabilities, especially within the human–robot collaborative manufacturing area, many technical issues have not yet been well considered or addressed, which largely constrains the applicability of VLMs in practical real-life environments. In this section, some key challenges and corresponding future directions are discussed in the hope of motivating further research endeavors for VLM-based HRC systems.

6.1 Data and computation-efficient training and deployment of VLMs

The pretraining and deployment of VLMs in practical manufacturing scenes face significant challenges, primarily due to the high computational demands and extensive data requirements for training these models. Obtaining high-quality, annotated data sets in diverse manufacturing

environments is costly and time-consuming. Additionally, effective HRC applications require real-time processing, but VLMs often face latency issues. Meanwhile, in real-life production scenarios, VLMs must also be robust and reliable to handle variations and ambiguities effectively. Enhancing robustness through rigorous testing and incorporating fail-safe mechanisms is essential for reliable deployment. Potential pathways to addressing these issues include efficient training strategies, model optimisation techniques such as pruning and quantisation, and robust data collection methods. These approaches are essential for the practical and effective training and deployment of VLMs in HRC manufacturing applications.

6.2 Vision and language task planning in dynamic environments

One of the significant challenges in VLM-based task planning is the current focus on static scenes, which limits the applicability of these models in dynamic environments. Real-time task planning in dynamic scenes remains an unresolved issue. To address this, future research could explore the integration of Simultaneous Localization and Mapping (SLAM) technology. SLAM can provide real-time updates of the environment, enabling VLMs to adapt and plan tasks dynamically. This integration would allow robots to navigate and perform tasks in ever-changing environments, significantly enhancing their utility and robustness. Additionally, improving the computational efficiency of VLMs to handle real-time data and developing more sophisticated algorithms for dynamic task planning are crucial areas for future exploration.

6.3 Real-time 3D scene reconstruction and segmentation for vision and language navigation

Despite the significant advancements in VLN brought about by VLMs and LLMs, their application in industrial manufacturing remains limited. Current navigation planning relies on pre-established static maps, but in real-world scenarios, both humans and machines can be mobile. Therefore, real-time updates of 3D scene maps are crucial. However, most 3D reconstruction and segmentation techniques rely on RGB-D video frames for point cloud reconstruction and color rendering, which results in long inference times and significant delays in real-time reconstruction. This limitation restricts their application in real industrial settings. Achieving fast, low-latency real-time 3D reconstruction and segmentation is a critical research direction for the future. A potential solution involves combining large models with lightweight networks and dynamic tracking by human operators, enabling efficient, low-latency 3D scene updates and adaptability in industrial environments.

6.4 Motion planning with high precision for vision and language manipulation

VLM and LLM-based robotic manipulation is a highly active research topic in the fields of AI and robotics. However, the focus has predominantly been on household tasks, with limited work in industrial applications. One reason is that the planning precision of LLMs is not yet sufficient for industrial requirements. Currently, VLM/LLM-based robotic manipulation can only perform tasks such as “picking up a cup and pouring water,” but assembly tasks are much more complex and precise, such as aligning gears or installing screws. These tasks require motion planning with millimeter-level or even finer precision. VLMs and LLMs are known for their flexibility and generalisation, which inevitably leads to a decrease in precision for specific tasks. Enhancing the motion planning precision of VLM/LLM-based manipulation to achieve a balance of flexibility, generalisation, and accuracy is a promising research direction. To address the challenge of achieving high precision in motion planning for vision and language manipulation in industrial applications, integrating advanced sensor technologies and feedback control systems may offer a viable solution.

6.5 Additional modalities and complex instruction understanding

Integrating additional modalities into VLM-guided human–robot skill transfer could enhance a robot's contextual understanding and skill acquisition. While visual and linguistic cues furnish robots with spatial semantics, the incorporation of haptic feedback through advanced sensors is pivotal. This addition enables robots with precise force information, thereby augmenting their capacity to execute complex tasks that necessitate delicate force control. The multimodal approach not only enriches the sensory feedback loop for real-time action adjustment but also broadens the robot's interaction repertoire.

Furthermore, existing research often confines language instructions to a simplistic format, typically a “(verb) (noun)” structure. However, advancing toward more complex linguistic instructions is crucial for enhancing real-world applicability. The current limitations of VLMs in handling intricate or contextually dependent instructions, especially in scenarios with dynamic task or action sequences, require improvement. A key focus should be enhancing VLMs to comprehend multi-step logical instructions and clarify non-standard linguistic cues. For example, integrating advanced natural language processing techniques, such as context-aware transformers and hierarchical task analysis, can significantly improve the comprehension of complex instructions. This evolution is vital for enabling robots with the sophistication required to navigate and adapt effectively in diverse real-world scenarios.

6.6 Dynamic task adaptation and unsupervised evaluation

To facilitate the effective transfer of skills acquired in simulations to real-world applications, learning algorithms must possess robustness to real-world variability, encompassing factors such as fluctuating lighting and discrepancies in physical engine accuracy. However, current methods always rely on an assumption of an ideal training environment, with expectations of reasonable object original poses or a standardised training step. However, real-world deployment confronts these algorithms with a far greater range of unpredictability, which necessitates human operators for continuous monitoring and intervention. This significantly hinders real applications.

To overcome this, robots must evolve to autonomously adapt to dynamic environments and tasks, incorporating multimodal feedback for self-adjustment. This may require an advanced approach that integrates domain randomization and augmented reality for training, intentionally introducing a broad spectrum of real-world noise and variability within simulated environments. By doing so, skill transfer models can be equipped with enhanced generalisation abilities. Moreover, the implementation of unsupervised evaluation mechanisms can establish a robust self-assessment framework. Leveraging the observation of robot behavior or state changes could reduce constant human supervision when faced with crucial situations. Specifically, implementing a continuous learning mechanism allows robots to adapt and improve their performance autonomously over time based on new experiences and feedback from the environment.

7 Conclusions

This survey investigated recent advancements and applications of VLMs in HRC for smart manufacturing, highlighting their potentials and current limitations. Starting with an overview of the fundamental knowledge of LLMs and VLMs, it detailed how integrating visual and textual data enhances robot planning, execution, and learning capabilities. One can observe that initial explorations of VLMs in robotic task planning, navigation, and manipulation have shown promising improvements in flexibility and efficiency, playing a vital role for HRC in dynamic manufacturing environments. Meanwhile, VLMs have also demonstrated the ability to streamline robot skill learning by leveraging multimodal data integration. Despite these advancements, challenges such as real-time processing, computational demands, and handling dynamic environments are yet to be explored urgently and timely. To fully unlock the potential of VLMs in human-centric smart manufacturing, the following key perspectives can be considered in future research: 1) exploring the scalability of VLMs in highly variable and real-time HRC scenarios to improve their robustness and applicability, 2) developing more natural and intuitive human–robot interaction mechanisms that can enhance the collaboration efficiency and smoothness, particularly with advancements in large multimodal models, and 3) exploring methods to reduce the data and computational requirements of VLMs to make them more practical for large-scale industrial deployment. By addressing these areas, future research can build on the progress made so far and push the boundaries of VLM integration into real-world HRC applications.

Competing Interests The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, Almeida D, Altenschmidt J, Altman S, Anadkat S others (2023). Gpt-4 technical report. arXiv preprint arXiv:230308774

Anthropic (2023). The Claude 3 Model Family: Opus, Sonnet, Haiku.

Askell A, Bai Y, Chen A, Drain D, Ganguli D, Henighan T, Jones A, Joseph N, Mann B, DasSarma N others (2021). A general language assistant as a laboratory for alignment. arXiv preprint arXiv:211200861

Azagra P, Civera J, Murillo A C (2020). Incremental learning of object models from natural human–robot interactions. IEEE Transactions on Automation Science and Engineering, 17(4): 1883–1900

Brown T, Mann B, Ryder N, Subbiah M, Kaplan J D, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A others (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems 33:1877–1901

Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, Chen H, Yi X, Wang C, Wang Y, Ye W, Zhang Y, Chang Y, Yu PS, Yang Q, Xie X (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3): 1–45

Chen Q, Pitawela D, Zhao C, Zhou G, Chen H T, Wu Q (2024). WebVLN: Vision-and-language navigation on websites. In: Proceedings of the AAAI Conference on Artificial Intelligence, 38(2): 1165–1173

Chen T, Kornblith S, Norouzi M, Hinton G (2020). A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, 1597–1607

Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A others (2023). Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240): 1–113

Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, Li Y, Wang X, Dehghani M, Brahma S others (2024). Scaling instruction-finetuned language models. Journal of Machine Learning Research, 25(70): 1–53

Ding M, Xu Y, Chen Z, Cox D D, Luo P, Tenenbaum J B, Gan C (2023). Embodied concept learner: Self-supervised learning of concepts and mapping through Instruction Following. In: Conference on Robot Learning. PMLR, 1743–1754

Dong Q, Li L, Dai D, Zheng C, Wu Z, Chang B, Sun X, Xu J, Sui Z (2022). A survey on in-context learning. arXiv preprint arXiv:230100234

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations

Dou Z Y, Kamath A, Gan Z, Zhang P, Wang J, Li L, Liu Z, Liu C, LeCun Y, Peng N others (2022). Coarse-to-fine vision-language pre-training with fusion in the backbone. Advances in Neural Information Processing Systems 35: 32942–32956

Driess D, Xia F, Sajjadi M S, Lynch C, Chowdhery A, Ichter B, Wahid A, Tompson J, Vuong Q, Yu T others (2023). PaLM-E: An embodied multimodal language model. In: International Conference on Machine Learning. PMLR, 8469–8488

Du Y, Yang M, Florence P, Xia F, Wahid A, Ichter B, Sermanet P, Yu T, Abbeel P, Tenenbaum J B others (2023). Video language planning. arXiv preprint arXiv:231010625

Fan J, Zheng P (2024). A vision-language-guided robotic action planning approach for ambiguity mitigation in human–robot collaborative manufacturing. Journal of Manufacturing Systems, 74: 1009–1018

Fan J, Zheng P, Li S (2022). Vision-based holistic scene understanding towards proactive human–robot collaboration. Robotics and Computer-integrated Manufacturing, 75: 102304

Fu Z, Lam W, Yu Q, So A M C, Hu S, Liu Z, Collier N (2023). Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. arXiv preprint arXiv:230404052

Gao C, Liu S, Chen J, Wang L, Wu Q, Li B, Tian Q (2024). Room-object entity prompting and reasoning for embodied referring expression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(2): 994–1010

Gervet T, Xian Z, Gkanatsios N, Fragkiadaki K (2023). Act3D: 3D feature field transformers for multi-task robotic manipulation. In: Conference on Robot Learning. PMLR, 3949–3965

GLM T Zeng A, Xu B, Wang B, Zhang C, Yin D, Rojas D, Feng G, Zhao H, Lai H (2024). ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. arXiv preprint arXiv:240612793

Goodwin W, Vaze S, Havoutis I, Posner I (2022). Semantically grounded object matching for robust robotic scene rearrangement. In: 2022 International Conference on Robotics and Automation (ICRA). IEEE, Philadelphia, PA, USA, 11138–11144

Gu Q, Kuwajerwala A, Morin S, Jatavallabhula K M, Sen B, Agarwal A, Rivera C, Paul W, Ellis K, Chellappa R others (2023). Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. arXiv preprint arXiv:230916650

Halim J, Eichler P, Krusche S, Bdiwi M, Ihlenfeldt S (2022). No-code robotic programming for agile production: A new markerless-approach for multimodal natural interaction in a human-robot collaboration context. Frontiers in Robotics and AI, 9: 1001955

Han R, Liu N, Liu C, Gou T, Sun F (2024). Enhancing robot manipulation skill learning with multi-task capability based on transformer and token reduction. In: Cognitive Systems and Information Processing. Springer Nature Singapore, Singapore, 121–135

He K, Fan H, Wu Y, Xie S, Girshick R (2020). Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 9729–9738

He K, Zhang X, Ren S, Sun J (2016). Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp 770–778

He P, Liu X, Gao J, Chen W (2021). Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv: 200603654

Hong Y, Zhou Y, Zhang R, Dernoncourt F, Bui T, Gould S, Tan H (2023). Learning navigational visual representations with semantic map supervision. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Paris, France, pp 3032–3044

Hori C, Peng P, Harwath D, Liu X, Ota K, Jain S, Corcodel R, Jha D, Romeres D, Le Roux J (2023). Style-transfer based speech and audio-visual scene understanding for robot action sequence acquisition from videos. arXiv preprint arXiv: 230615644

Hu Y, Lin F, Zhang T, Yi L, Gao Y (2023) Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. arXiv preprint arXiv: 231117842

Huang C, Mees O, Zeng A, Burgard W (2023a). Visual language maps for robot navigation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 10608–10615

Huang W, Wang C, Zhang R, Li Y, Wu J, Fei-Fei L (2023b). Voxposer: Composable 3D value maps for robotic manipulation with language models. In: Conference on Robot Learning. PMLR, 540–562

Jang E, Irpan A, Khansari M, Kappler D, Ebert F, Lynch C, Levine S, Finn C (2022). Bc-z: Zero-shot task generalization with robotic imitation learning. In: Conference on Robot Learning. PMLR, 991–1002

Jang J, Kong C, Jeon D, Kim S, Kwak N (2023). Unifying vision-language representation space with single-tower transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. 980–988

Jia C, Yang Y, Xia Y, Chen Y T, Parekh Z, Pham H, Le Q, Sung Y H, Li Z, Duerig T (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. PMLR, 4904–4916

Kenton J D M W C, Toutanova L K (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp 4171–4186

Khandelwal A, Weihs L, Mottaghi R, Kembhavi A (2022). Simple but effective: Clip embeddings for embodied AI. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, New Orleans, LA, USA, 14809–14818

Kim J, Kang G C, Kim J, Shin S, Zhang B T (2023a). GVCCI: Lifelong learning of visual grounding for language-guided robotic manipulation. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, Detroit, MI, USA, 952–959

Kim S, Joo S J, Kim D, Jang J, Ye S, Shin J, Seo M (2023b). The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In: The 2023 Conference on Empirical Methods in Natural Language Processing. 12685–12708

Kojima T, Gu S, Reid M, Matsuo Y, Iwasawa Y (2022). Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems. 22199–22213

Korekata R, Kambara M, Yoshida Y, Ishikawa S, Kawasaki Y, Takahashi M, Sugiura K (2023). Switching head-tail funnel UNITER for dual referring expression comprehension with fetch-and-carry tasks. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, Detroit, MI, USA, 3865–3872

Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 7871–7880

Li J, Padmakumar A, Sukhatme G, Bansal M (2024). Vln-video: Utilizing driving videos for outdoor vision-and-language navigation. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17): 18517–18526

Lin B, Nie Y, Wei Z, Zhu Y, Xu H, Ma S, Liu J, Liang X (2024). Correctable landmark discovery via large models for vision-language navigation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12): 1–14

Liu S, Zhang J, Wang L, Gao R X (2024). Vision AI-based human-robot collaborative assembly driven by autonomous robots. CIRP Annals, 73(1): 13–16

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692

Lu S, Berger J, Schilp J (2022). System of robot learning from multimodal demonstration and natural language instruction. Procedia CIRP, 107: 914–919

Matheson E, Minto R, Zampieri E G, Faccio M, Rosati G (2019). Human–robot collaboration in manufacturing applications: A review. Robotics, 8(4): 100

Mei A, Wang J, Zhu G N, Gan Z (2024). GameVLM: A decision-making framework for robotic task planning based on visual language models and zero-sum games. arXiv preprint arXiv:24051375

Mohammadi B, Hong Y, Qi Y, Wu Q, Pan S, Shi J Q (2024). Augmented commonsense knowledge for remote object grounding. Proceedings of the AAAI Conference on Artificial Intelligence, 38(5): 4269–4277

Murray M, Gupta A, Cakmak M (2024). Teaching robots with show and tell: Using foundation models to synthesize robot policies from language and visual demonstration. In: 8th Annual Conference on Robot Learning

Nair S, Rajeswaran A, Kumar V, Finn C, Gupta A (2022). R3M: A universal visual representation for robot manipulation. In: Conference on Robot Learning. PMLR, 892–909

Park S, Menassa C C, Kamat V R (2024). Integrating large language models with multimodal virtual reality interfaces to support collaborative human-robot construction work. arXiv preprint arXiv:240403498

Peng A, Bobu A, Li B Z, Sumers T R, Sucholutsky I, Kumar N, Griffiths T L, Shah J A (2024). Preference-conditioned language-guided abstraction. In: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. ACM, Boulder CO USA, 572–581

Peng B, Li C, He P, Galley M, Gao J (2023). Instruction tuning with gpt-4. arXiv preprint arXiv:230403277

Qiao Y, Qi Y, Yu Z, Liu J, Wu Q (2023). March in chat: interactive prompting for remote embodied referring expression. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Paris, France, 15712–15721

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J others (2021). Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, 8748–8763

Radford A, Narasimhan K, Salimans T, Sutskever I (2018). Improving language understanding by generative pre-training. OpenAI blog

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, others (2019). Language models are unsupervised multitask learners. OpenAI blog 1(8):9

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu P J (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140): 1–67

Ramrakhya R, Undersander E, Batra D, Das A (2022). Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 5173–5183

Rana K, Haviland J, Garg S, Abou-Chakra J, Reid I, Suenderhauf N (2023). Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In: 7th Annual Conference on Robot Learning. pp 23–72

Sanh V, Webson A, Raffel C, Bach S H, Sutawika L, Alyafeai Z, Chaffin A, Stiegler A, Le Scao T, Raja A others (2022). Multitask prompted training enables zero-shot task generalization. In: International Conference on Learning Representations

Schumann R, Zhu W, Feng W, Fu T J, Riezler S, Wang W Y (2024). VELMA: Verbalization embodiment of LLM agents for vision and language navigation in street view. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17): 18924–18933

Shah D, Osinski B, Ichter B, Levine S (2022). Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In: Conference on Robot Learning. PMLR, 492–504

Shao L, Migimatsu T, Zhang Q, Yang K, Bohg J (2021). Concept2Robot: Learning manipulation concepts from instructions and human demonstrations. International Journal of Robotics Research, 40(12-14): 1419–1434

Sharma S, Huang H, Shivakumar K, Chen L Y, Hoque R, Ichter B, Goldberg K (2023). Semantic mechanical search with large vision and language models. In: Conference on Robot Learning. PMLR,

971–1005

Shukla R, Manyar O M, Ranparia D, Gupta S K (2023). A framework for improving information content of human demonstrations for enabling robots to acquire complex tool manipulation skills. In: 2023 32nd IEEE International Conference on Robot and Human Interactive Communication. IEEE, Busan, Korea, Republic of, 2273–2280

Singh A, Hu R, Goswami V, Couairon G, Galuba W, Rohrbach M, Kiela D (2022). Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 15638–15650

Skreta M, Zhou Z, Yuan J L, Darvish K, Aspuru-Guzik A, Garg A (2024). Replan: Robotic replanning with perception and language models. arXiv preprint arXiv:240104157

Song C H, Wu J, Washington C, Sadler B M, Chao W L, Su Y (2023). Llm-planner: Few-shot grounded planning for embodied agents with large language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2998–3009

Song D, Liang J, Payandeh A, Xiao X, Manocha D (2024). Socially aware robot navigation through scoring using vision-language models. arXiv preprint arXiv:240400210

Sontakke S A, Zhang J, Arnold S M R, Pertsch K, Bıyık E, Sadigh D, Finn C, Itti L (2024). Roboclip: One demonstration is enough to learn robot policies. In: Proceedings of the 37th International Conference on Neural Information Processing Systems, 55681-55693

Stone A, Xiao T, Lu Y, Gopalakrishnan K, Lee K H, Vuong Q, Wohlhart P, Kirmani S, Zitkovich B, Xia F, Finn C, Hausman K (2023). Open-world object manipulation using pre-trained vision-language models. In: Conference on Robot Learning. PMLR, 3397–3417

Sun Y, Wang S, Feng S, Ding S, Pang C, Shang J others (2021). Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:210702137

Tan M, Le Q (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, 6105–6114

Tay Y, Dehghani M, Tran V, Garcia X, Wei J, Wang X, Chung H W, Bahri D, Schuster T, Zheng S, Zhou D, Houlsby N, Metzler D (2023). UL2: Unifying Language Learning Paradigms. In: The Eleventh International Conference on Learning Representations

Team G, Anil R, Borgeaud S, Wu Y, Alayrac J B, Yu J others (2023). Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:231211805

Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A, Cheng H T others (2022). Lamda: Language models for dialog applications. arXiv preprint arXiv:220108239

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T others (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971

Trick S, Herbert F, Rothkopf C A, Koert D (2022). Interactive reinforcement learning with Bayesian fusion of multimodal advice. IEEE Robotics and Automation Letters, 7(3): 7558–7565

Tschannen M, Mustafa B, Houlsby N (2022). Image-and-language understanding from pixels only. arXiv preprint arXiv:221208045

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N Kaiser Lukasz, Polosukhin I (2017). Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. 6000–6010

Wang J, Wang T, Xu L, He Z, Sun C (2024a). Discovering intrinsic subgoals for vision-and-language navigation via hierarchical reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems, 1–13

Wang L, Gao R, Váncza J, Krüger J, Wang X V, Makris S, Chryssolouris G (2019). Symbiotic human-robot collaborative assembly. CIRP Annals, 68(2): 701–726

Wang T, Fan J, Zheng P (2024b). An LLM-based vision and language cobot navigation approach for Human-centric Smart Manufacturing. Journal of Manufacturing Systems, 75: 299–305

Wang T, Roberts A, Hesslow D, Le Scao T, Chung H W, Beltagy I, Launay J, Raffel C (2022a). What language model architecture and pretraining objective works best for zero-shot generalization? In: International Conference on Machine Learning. PMLR, 22964–22984

Wang T, Zheng P, Li S, Wang L (2024c). Multimodal human–robot interaction for human-centric smart manufacturing: A survey. Advanced Intelligent Systems, 6(3): 2300359

Wang W, Li R, Chen Y, Sun Y, Jia Y (2022b). Predicting human intentions in human–robot hand-over tasks through multimodal learning. IEEE Transactions on Automation Science and Engineering, 19(3): 2339–2353

Wang X, Wang W, Shao J, Yang Y (2024d). Learning to follow and generate instructions for language-capable navigation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(5): 3334–3350

Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le Q V, Zhou D others (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35: 24824–24837

Wi Y, Mark V der M, Pete F, Zeng A, Fazeli N (2023). CALAMARI: Contact-aware and language conditioned spatial action mapping for contact-rich manipulation. In: Conference on Robot Learning. PMLR, 2753–2771

Winge C, Imdieke A, Aldeeb B, Kang D, Desingh A (2024). Talk through it: End user directed manipulation learning. IEEE Robotics and Automation Letters, 9(9): 8051–8058

Wu Z, Wang Z, Xu X, Lu J, Yan H (2023). Embodied task planning with large language models. arXiv preprint arXiv:230701848

Yao L, Han J, Wen Y, Liang X, Xu D, Zhang W, Li Z, Xu C, Xu H (2022). Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. Advances in Neural Information Processing Systems, 35: 9125–9138

Yin C, Zhang Q (2023). A multi-modal framework for robots to learn manipulation tasks from human demonstrations. Journal of Intelligent & Robotic Systems, 107(4): 56

Yin Y, Zheng P, Li C, Wan K (2024). Enhancing human-guided robotic assembly: AR-assisted DT for skill-based and low-code programming. Journal of Manufacturing Systems, 74: 676–689

Yu J, Wang Z, Vasudevan V, Yeung L, Seyedhosseini M, Wu Y (2022). Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:220501917

Yu T, Zhou Z, Chen Y, Xiong R (2023). Learning object spatial rela-

tionship from demonstration. In: 2023 2nd International Conference on Machine Learning, Cloud Computing and Intelligent Mining. 370–376

Ze Y, Yan G, Wu Y H, Macaluso A, Ge Y, Ye J, Hansen N, Li L E, Wang X (2023). GNFactor: Multi-task real robot learning with generalizable neural feature fields. In: Conference on Robot Learning. PMLR, 284–301

Zhang J, Huang J, Jin S, Lu S (2024a). Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8): 5625–5644

Zhang J, Wang K, Xu R, Zhou G, Hong Y, Fang X, Wu Q, Zhang Z, Wang H (2024b). NaVid: Video-based VLM plans the next step for vision-and-language navigation. arXiv preprint arXiv:240215852

Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q (2019). ERNIE: Enhanced Language Representation with Informative Entities. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 1441–1451

Zhao W X, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z others (2023a). A survey of large language models. arXiv preprint arXiv:230318223

Zhao X, Li M, Weber C, Hafez M B, Wermter S (2023b). Chat with the environment: Interactive multimodal perception using large language models. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 3590–3596

Zheng P, Li C, Fan J, Wang L (2024). A vision-language-guided and deep reinforcement learning-enabled approach for unstructured human-robot collaborative manufacturing task fulfilment. CIRP Annals, 73(1): 341–344

Zhou G, Hong Y, Wu Q (2024). Navgpt: Explicit reasoning in vision-and-language navigation with large language models. Proceedings of the AAAI Conference on Artificial Intelligence, 38(7): 7641–7649

Zhou K, Yang J, Loy C C, Liu Z (2022). Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9): 2337–2348

Ziegler D M, Stiennon N, Wu J, Brown T B, Radford A, Amodei D, Christiano P, Irving G (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:190908593

Zitkovich B, Yu T, Xu S, Xu P, Xiao T, Xia F, Wu J, Wohlhart P, Welker S, Wahid A others (2023). Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. PMLR, 2165–2183