Embodied Intelligent Mobile Operation Robots For Industrial Scenarios

Weifeng Zhao[1*]

[1] Foshan Rossum Robotics Co.,Ltd., Foshan, China, 528000

* 281606597@qq.com

https://doi.org/10.70695/IAAI202504A13

Abstract

In industrial scenarios with multiple workstations and processes, traditional rule-based control and small-scale policy networks are prone to performance degradation when tasks are expanded and equipment varies. Therefore, this study developed an embodied intelligent mobile operation platform. This platform unifies the observation -action interface of the mobile chassis, robotic arm, and multimodal sensors. An industrial-grade large-scale model was also designed, integrating vision, point cloud, force sensing, and language commands, and improved through pre-training and command fine-tuning. Experiments based on a real-world workshop task library show that the model's accuracy in multi-task operations in the source domain is close to that of manual operation. Furthermore, it demonstrates high adaptability to unfamiliar task synthesis and workstation layouts in zero-sample and few-sample scenarios. Ablation and engineering case studies further validate the benefits of multimodal fusion, hierarchical motion, and generalization improvement in terms of cycle time, human time, and safety, demonstrating its replicable engineering application value.

Keywords Embodied Intelligence; Mobile Manipulation Robot; Large Industrial Model; Multimodal Perception; Command Fine-tuning; Generalization Ability

1 Introduction

Embodied intelligence and large model integration help industrial robots evolve from "programmable" to "autonomous decision-making". Fan et al. applied large language models to manufacturing scenarios, which improved cross-workstation planning capabilities, but generalization capabilities were limited [1]. Cong and Mo reviewed multimodal embodied models, pointing out the importance of collaborative perception and the difficulties of spatiotemporal alignment [2]. Lee emphasized the coupling relationship between body and environment, but lacked industrial-level verification data [3]. Ren et al. explored the potential of basic models in the field of flexible manufacturing, but the engineering paradigm was not clear [4]. Xu et al. proposed the concept of human-centered embodied intelligence, focusing on the shortcomings in its robustness and interpretability [5]. Zhao and Yuan pointed out the gaps in the existing evaluation system [6]. Dong et al. demonstrated the advantages of graph understanding, but there was little immediate feedback [7]. Lisondra et al. verified the cross-task potential, but focused on service scenarios [8]. Bu et al.'s Agibot World Colosseo suffers from a lack of security constraints [9]. Song et al. explored its prospects in power grid operation and maintenance, but most of their work remained at the proof-of-concept level [10].

Existing research has laid a certain foundation in models, platforms, and applications, but a large-scale model system is still lacking. This system should be oriented towards industrial sites, possessing integrated "mobility + operation" characteristics, and having strong safety constraints. Based on this, this research, relying on an industrial -grade embodied platform and a multimodal task library, attempts to develop a large-scale mobile operation model. This model needs to possess immediacy, robustness, and cross-workstation generalization capabilities, and will provide a reproducible engineering verification approach through system experiments and an indicator system.

2 Overall Design of Industrial Embossed Mobile Operation Platform and Large Model

2.1 Industrial Embossed Mobile Operation Robot System Architecture

To support industrial- grade large-scale models, the platform adopts a configuration of mobile chassis + 6/7-DOF robotic arm + replaceable end effector + multi-sensor array: the chassis undertakes cross-workstation movement tasks, the robotic arm performs precise operations, the end effector completes tasks such as gripping, plugging or tightening, and RGB-D, 3D LiDAR, IMU and six-dimensional torque are connected to the Edge GPU and Safety PLC via EtherCAT and industrial Ethernet, thus creating a three-layer architecture of control-perception-safety (as shown in Figure 1).

To characterize the feasible operational domain, the end-effector workspace is modeled as follows:

$$W = \left\{ x \in ò^{3} | \ \| x - x_c \|_2 \le R, x = f(q) \right\} \qquad (1)$$

Here, W represents the set of reachable workspaces, $x \in R^3$ is the position vector of the end effector in the workpiece coordinate system, $x_c \in R^3$ i.e., the position of the geometric center of the target workstation, $R > 0$, this value is selected as the maximum working radius considering joint limits and safety distance factors, $q \in R^n$ is the angle vector of the robotic arm joint, $f(q)$ and is the positive kinematic mapping. $\|\cdot\|_2$ is the Euclidean 2 norm. Tab.1 specifies key parameters such as degrees of freedom, load, repeatability, frame rate, and bandwidth, thereby providing boundary conditions for subsequent perception modeling and large model inference.

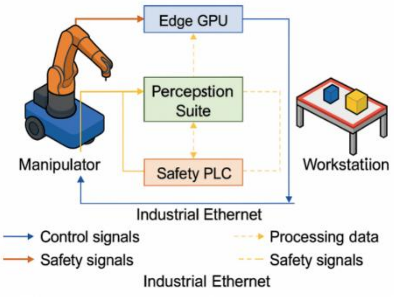

Fig. 1. Schematic diagram of overall system architecture and typical workstation layout

Table 1. Robot body and sensor key parameters

Module	Type	DoF	Payload capacity (kg)	Position/measurement precision	Rate (Hz / fps)	Interface	Bandwidth (Mbps)
Manipulator	6-axis industrial arm	6	10.0	±0.05 mm (repeatability)	250 Hz (servo loop)	EtherCAT	100
Mobile base	Indoor omni base	3	200.0	±10 mm (localization error)	50 Hz (odometry)	CAN + Industrial Ethernet	100
RGB-D camera	Industrial RGB-D cam	0	0.5	±2 mm @ 1 m depth	30 fps	GigE Vision	1000
3D LiDAR	32-line 3D LiDAR	0	1.0	±30 mm (typical range accuracy)	10 Hz (full scan)	UDP over Ethernet	1000
Force/Torque sensor	6-axis F/T sensor	6	0.6	0.1 N (force resolution), 0.005 Nm	1000 Hz	EtherCAT	10
IMU	Industrial IMU	3	0.2	0.05° (gyro bias stability)	400 Hz	RS-485 / CAN	10

2.2 Multimodal Perception and Action Space Modeling

The aforementioned platform needs to extract multi-source sensor data and low-level control interfaces into an "observation-operation" interface that can be directly used by large models, thereby achieving broad applicability across workstations and tasks. Definition of multimodal observation at discrete time t:

$$o_t = \{I_t, P_t, F_t, s_t\}, \quad z_t = \phi(o_t) = \Phi(\phi_I(I_t), \phi_P(P_t), \phi_F(F_t), \phi_s(s_t)) \quad (2)$$

Here, o_t the multimodal observation set representing time t I_t includes calibrated images or RGB-D frames, P_t point clouds generated by radar or depth maps, F_t wrist force/torque vectors, s_t chassis and joint states including position, velocity, and odometry information, $z_t \in R^d$ fused temporal features, $\phi_I, \phi_P, \phi_F, \phi_s$ and various modal encoding functions, including CNN/ ViT, point cloud networks, temporal force encoders, and MLPs, as well as $\Phi(\cdot)$ cross-modal fusion operators such as multi-head attention, which complete feature-level alignment.

The action adopts a hierarchical structure, interfaces with industrial control systems, and provides composable operation primitives for large models:

$$\pi(a_t | s_t, g) = \pi_{\text{low}}(a_t^{\text{low}} | z_t, u_t), \quad u_t \sim \pi_{\text{mid}}(u_t | z_t, g) \quad (3)$$

Among them, $\pi(a_t | s_t, g)$ represents the overall strategy distribution, a_t which is the actual operation command sent to the controller, a_t^{low} is the low-level continuous control quantity, that is, the joint or Cartesian velocity, u_t is the mid-level skill primitive, such as pick, insert, etc., s_t is the state vector, which has o_t the same source as, g is the integration of high-level task objectives or language instructions, π_{mid} is the skill selection strategy network, π_{low} and is the continuous control strategy network.

2.3 Large-scale Structural Design of Embodied Intelligent Mobile Operation Robot

The first line For embodied intelligent mobile operation in industrial scenarios. Multimodal observations are encoded as visual, point cloud, force /state, and language tokens, respectively. After alignment by a cross-attention fusion module, they are input into a temporal transformer to generate skill sequences and underlying action distributions. Fig. 2 illustrates the end-to-end "perception-semantics-action" process. The training process employs a unified multi-task loss to integrate supervision, behavior cloning, and contrastive learning.

$$L_{total} = \lambda_{sup} L_{sup} + \lambda_{bc} L_{bc} + \lambda_{ctr} L_{ctr} \quad (4)$$

Among them, L_{total} represents the total training loss, L_{sup} which is the supervision loss obtained when labeling actions or intermediate variables, L_{bc} the behavior cloning loss for the taught trajectory, L_{ctr} and the contrastive learning loss across tasks and workstations. The purpose is to improve the generalization ability of the representation. Moreover, $\lambda_{sup}, \lambda_{bc}, \lambda_{ctr} > 0$ the weights of these losses are determined by the parameter tuning of the validation set.

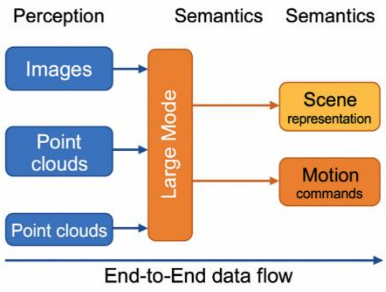

Fig. 2. Large model network structure and data flow diagram of "perception-semantics-action"

2.4 Industrial Deployment and Online Inference Framework

To ensure stable operation of large-scale models on the production line, an integrated framework of "offline training—online inference—online fine-tuning" is constructed. Offline, pre-training and instruction fine-tuning are completed on the cluster using multi-workstation data, and versioned releases are implemented. Online, observations o_t and instructions are received by edge GPUs/IPCs g , outputting actions and mapping them to skills and underlying instructions. Online fine-tuning transmits failures and new operating conditions back for incremental training, achieving continuous generalization across workstations and devices. The large-scale model output also incorporates confidence gating and a safety shell adjudication.

$$\gamma_t = \max_{a_t} p\left(a_t | s_t\right), \quad u_t = \begin{cases} a_t^{LLM}, & \gamma_t \geq \tau \\ a_t^{\text{safe}}, & \gamma_t < \tau \end{cases} \tag{5}$$

Wherein, $p\left(a_t | s_t\right)$ represents s_t the probability distribution of actions of the large model under the state; $\gamma_t \in [0,1]$ is the maximum action confidence; $\tau \in (0,1)$ is the threshold specified according to process risk; a_t^{LLM} is the candidate action provided by the large model; a_t^{safe} is the cautious safety action generated by the traditional u_t planner or Safety PLC; and is the action to be executed last.

3 Data Engineering and Training Methods in Industrial Scenarios

3.1 Industrial Task Library Construction and Data Acquisition Process

The embodied platform and the "perception-action" interface, an industrial task library covering typical processes is constructed. Tasks are defined as "movement+operation," encompassing handling, precision assembly, bolt tightening, insertion, and visual inspection. Each trajectory is broken down into navigation and operation segments, with key states such as approach, alignment, contact, and exit labeled. Data comes from simulation, teaching, and historical logs: digital twins generate baseline trajectories with multiple postures, and examples are collected from actual production lines through force control teaching and teleoperation, with richness expanded using playback of operation logs. Samples are uniformly stored as follows $\left(o_{1:T}, a_{1:T}, g, y\right)$: $o_{1:T}$ [data structure g not specified in the original text]. [data structure not specified y in the original text]. Table 2 provides statistical analysis of the scale, trajectory length, $a_{1:T}$ and duration, thus providing a quantitative basis for training and generalization experiments.

Table 2. Statistics of industrial task library

Task type	Episodes	Avg. length (steps)	Avg. duration (s)
Material handling	800	90	28.5
Precision assembly	600	120	42.3
Bolt fastening	450	110	35.7
Plug insertion	500	95	31.2
Visual inspection	700	80	26.9

3.2 Multimodal Data Preprocessing and Spatiotemporal Alignment

First, hardware timestamps and PTP clocks are used to I_t, P_t, F_t, s_t synchronize the execution time and adjust all sequences to the same frequency. Then, the images, point clouds, and force perceptions are reduced to the workpiece coordinate system through the extrinsic calibration matrix. Outliers can be removed by amplitude thresholding and continuity detection. For missing segments, interpolation and local downsampling are used to coordinate the temporal density.

Normalization and alignment errors use a unified loss metric:

$$L_{\text{align}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M} \| \frac{x_t^{(m)} - \mu^{(m)}}{\sigma^{(m)}} - \frac{x_{t+\delta_t^{(m)}}^{(m)} - \mu^{(m)}}{\sigma^{(m)}} \|_2^2 \tag{6}$$

In Equation (6), L_{align} represents the total alignment loss, T is the number of time steps contained in a single trajectory, M is the number of modalities (i.e., image, point cloud, force perception, state, etc.), is $x_t^{(m)}$ the original observation vector of $\mu^{(m)}$ the modality m at time, t and are $\sigma^{(m)}$ the mean and standard deviation of the modality in the task library, $\delta_t^{(m)}$ respectively, is m the time offset estimate of $\|\cdot\|_2$ the modality m at time, t and refers to the Euclidean 2 norm. By minimizing Lalign, the time axes of each modality can be aligned in the normalized space, thereby ensuring z_t the temporal consistency of the fused features in Section 2.2 and providing a reliable time axis alignment benchmark for the generation of success/failure and error labels.

3.3 Pre-training and Instruction Fine-tuning Strategies

After completing spatiotemporal alignment, offline pre-training and instruction fine-tuning are performed around the large model structure described in section 2.3. Pre-training includes behavior cloning to $o_{1:T}$ regress expert actions $a_{1:T}$ and inverse dynamics prediction to (s_t, s_{t+1}) enhance dynamic consistency by inferring intermediate actions. Instruction-tuning is then added, binding trajectories to natural language or template instructions and encoding them as conditional vectors, aligned with the high-level objective g, to improve task composability and interpretability. The overall training loss is expressed as:

$$L_{pre} = \lambda_{bc}L_{bc} + \lambda_{inv}L_{inv} + \lambda_{lang}L_{lang} \tag{7}$$

In equation (7), L_{pre} represents the overall loss in the pre-training phase, L_{bc} is the behavior cloning loss term, which aims to fit the demonstrated action sequence, L_{inv} is the inverse dynamics loss term, which restricts state transitions to meet action requirements, L_{lang} is the policy loss term with linguistic conditions, which is used to make the instruction semantics and action distribution match, and $\lambda_{bc}, \lambda_{inv}, \lambda_{lang} > 0$ are the weight coefficients of the corresponding loss terms.

3.4 Strategies and Theoretical Analysis for Enhancing Generalization Ability

Scene randomization, perturbation of workstations, lighting, and noise generates multi-view samples; multi-task hybrid training proportionally fuses handling and assembly to avoid overfitting; cross-workstation resampling supplements unseen combinations. Fig. 3 shows the convergence trend of overall loss, source/target domain success rate, and poor performance. From the perspective of domain adaptation, the strategy can be explained using the classical generalization error upper bound:

$$\varepsilon_T(h) \le \varepsilon_S(h) + d_{\mathrm{H}\Delta\mathrm{H}}(S,T) + \lambda \tag{8}$$

In equation (8), $\varepsilon_T(h)$ represents the error of the hypothesis h in the target domain, T $\varepsilon_S(h)$ represents the error in the source domain, S $d_{\mathrm{H}\Delta\mathrm{H}}(S,T)$ represents H the domain difference measure for the hypothesis set, and λ represents the non-reducible error of the hypothesis that is optimal in both domains. By using scenario randomization and cross-workstation resampling, the upper bound of can be clearly reduced within the empirical range $d_{\mathrm{H}\Delta\mathrm{H}}(S,T)$. Moreover, multi-task hybrid training and instruction fine-tuning are beneficial to reducing $\varepsilon_S(h)$, and together $\varepsilon_T(h)$ they tighten the upper bound of, thus providing a theoretical basis for subsequent cross-workstation experiments.

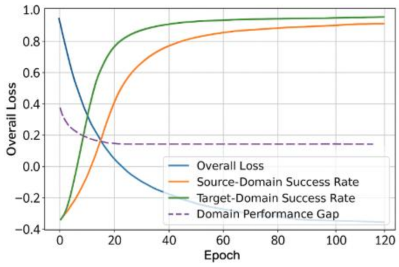

Fig. 3. Convergence curves of loss and generalization index during training

4 Experimental Design, Results, and Generalization Ability Assessment

4.1 Experimental Scenario and Evaluation Index System

Based on the platform and task library established in Chapters 2 and 3, six "movement-operation" consideration scenarios (S1-S6) are formed. Through multiple iterations of actual or semi-physical simulation episodes, multimodal observations, control commands, and safety shell trigger information are recorded, and abnormal samples are eliminated according to the cycle time. Evaluation metrics include success rate, end-point error, cycle time, and energy consumption. For assembly scenarios, alignment error and contact peak are also added, while for transport scenarios, path length and collision count are added. After calculating the mean and variance for each scenario, Table 3 shows that assembly and insertion constraints are more stringent, while handling and inspection are more affected by energy consumption.

Table 3. Statistics of scenarios and key metrics

Scenario ID	Task type	Episodes	Success rate (%)	Mean pose error (mm)	Mean cycle time (s)	Mean energy (Wh / task)
S1	Material handling	300	95.1	3.2	22.8	28.4
S2	Precision assembly	240	93.4	1.9	35.6	32.1
S3	Visual inspection	260	97.2	4.5	20.3	24.7
S4	Bolt fastening	220	90.8	2.6	33.1	30.9
S5	Plug insertion	230	89.7	2.2	31.4	29.8
S6	Cross-station move	210	94.5	5.1	18.9	26.0

4.2 Basic Performance and Comparative Experiment Results

The source domain scenarios S1–S3 of Table 3, the embodied large model is compared with three types of baselines: the training split used by traditional planning combined with PID/impedance control, small-scale single-task policy networks, and manual remote operation is the same. The large model performs joint pre-training and instruction fine-tuning according to Section 3.3, while the RL/BC baselines only implement improvements in their respective tasks. During the testing phase, 500 episodes were run under a unified safety shell, and the task indicators were statistically analyzed (Fig. 4). The results show that the large model is close to manual operation in terms of success rate, and outperforms traditional planning and small-scale network policies in terms of cycle time and energy consumption. The numerical and attitude error variances in Table 4 further show that EI-MO still maintains the low error and low fluctuation characteristics necessary for precision assembly under multi-task sharing conditions.

Table 4. Performance comparison on source-domain benchmark

Method	Success rate (%)	Mean pose error (mm)	Error var (mm^2)	Cycle time (s)	Energy / task (Wh)
EI-MO Large Model (Base)	96.2	1.8	0.9	24.5	31.2
EI-MO Edge Model	93.5	2.3	1.4	23.8	28.9
Classical planning + PID	87.1	3.9	3.2	29.7	37.5
SmallNet RL policy	82.4	4.5	4.1	27.9	34.8
BC per-task network	84.7	4.1	3.6	26.8	33.9
Human teleoperation	97.5	1.5	0.8	26.1	40.3

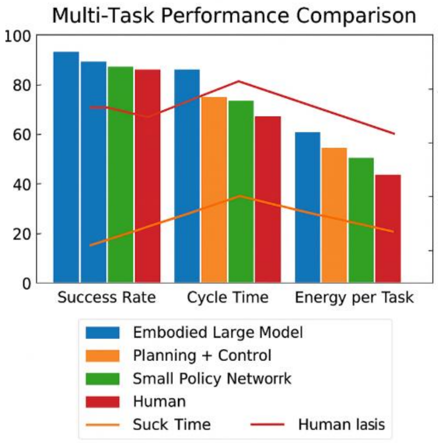

Fig. 4. Performance comparison of different methods on multiple tasks

4.3 Cross-task, Cross-workstation, and Cross-device Generalization Experiments

To verify the generalization capability for "industrial scenarios," three target domains were constructed outside the source domains S1–S3: the first is "unseen task synthesis," where the equipment remains unchanged but the operation sequence is rearranged; the second is "unseen workstations," where the workstation layout and fixture positions are changed; and the third is "unseen equipment," where the robotic arm, fixture, or mobile chassis is replaced at the same workstation. Zero-shot testing and few-shot adaptation were conducted for each setting. The former uses a small amount of data from the target domain to perform short-round fine-tuning, thus aligning with the actual production line debugging costs.

Table 5. Cross-task/cross-station generalization results

Source → Target	Setting	Target task type	Zero-shot succ (%)	Few-shot succ. (%)	Perf. drop vs. source (%)
S1 → S4 (buffer → assembly)	Layout only	Bolt fastening	78.3	90.6	6.1
S2 → S5 (asmA →plug)	Task change	Plug insertion	74.9	88.2	7.5
S3 → S2 (insp → asmA)	Task change	Precision assembly	70.4	86.7	8.9
S1 → S6 (buffer → transfer)	Device diff	New mobile base	80.1	91.3	5.4
S2 → S4 (asmA → asmB)	Layout+tool	New gripper	76.5	89.8	6.8
S5 → S4 (plug → bolt)	Task + layout	Bolt fastening	72.2	87.5	7.9

Fig. 5 visually shows that the success rate of the target domain rapidly approaches that of the source domain in the few-shot scenario, while a stable deviation still exists in the zero-shot scenario. Table 5 further lists the success rates of zero-shot and few-shot and the relative performance degradation by combination of "source domain → target domain". The results show that the multimodal embodied large model can recover to a level close to that of the source domain after a small amount of retraining at the new workstation and on the new equipment.

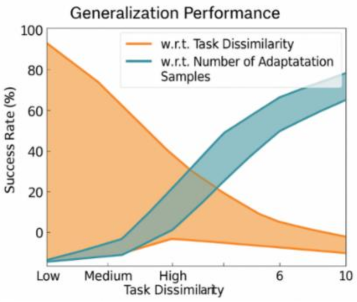

Fig. 5. Curve of generalization performance as a function of task variability or number of adapted samples

4.4 Case Analysis of Ablation Experiments and Engineering Applications

After completing the basic and generalization evaluations, this section analyzes the actual contribution of the model structure to the ablation process through ablation experiments and engineering case studies. While maintaining the data and training settings, we gradually removed multimodal fusion, linguistic conditions, hierarchical actions, and generalization enhancement strategies, and then conducted S1-S3 tests. The results show that the lack of multimodal fusion and hierarchical actions significantly reduces the success rate and increases the cycle time. Removing linguistic conditions primarily affects cross-task stability. A complete EI-MO model was deployed on a real assembly line, accessed via a confidence-gated safety architecture, and compared with the pre-deployment KPIs. Table 6 shows the cycle time, reduction rate of manual labor hours, and number of safety events per 100 hours, demonstrating that even without slowing down the cycle time, the model can significantly reduce human intervention and safety risks.

Table 6. Ablation study and engineering KPI summary

Config / Case	Fusion (0/1)	Language (0/1)	Hierarchy (0/1)	Gen. strategy (0/1)	Task succ. (%)	Cycle time (s)	Human work reduction (%)	Safety incidents /100h
Full model (all modules)	1	1	1	1	96.2	24.5	0.0	0.3
Without multimodal fusion	0	1	1	1	88.7	26.9	0.0	0.5
Without language conditioning	1	0	1	1	91.4	25.8	0.0	0.4
Without hierarchical control	1	1	0	1	89.1	27.3	0.0	0.6
Assembly line A (before deploy)	0	0	0	0	92.0	28.0	0.0	0.7
Assembly line A (after deploy)	1	1	1	1	96.8	27.5	24.3	0.3

5 Conclusion and Outlook

This research focuses on the design and generalization capabilities of a large-scale model for embodied intelligent mobile manipulators in industrial scenarios. It integrates hardware platforms, task libraries, multimodal large-scale models, and deployment frameworks. Experiments show that the unified model can cover various tasks such as handling, assembly, and inspection, exhibiting advantages over traditional planning and small-scale policy networks in terms of success rate, posture accuracy, and energy consumption. Furthermore, it can reproduce near-source domain performance with only a small number of samples in cross-task, cross-workstation, and cross-device scenarios. Ablation analysis indicates that multimodal fusion and hierarchical actions are crucial for improving robustness and cycle time. Instruction fine-tuning and generalization enhancement effectively improve domain offset. Assembly line case studies further confirm that this approach can reduce manual labor and mitigate safety risks. Future research will extend to collaborative operations and cross-factory migration, and further refine safety and interpretability mechanisms.

Acknowledgement

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Fan, H., Liu, X., Fuh, J. Y. H., et al. (2025). Embodied intelligence in manufacturing: Leveraging large language models for autonomous industrial robotics. Journal of Intelligent Manufacturing, 36(2), 1141-1157.
2. Cong, Y., & Mo, H. (2025). An overview of robot embodied intelligence based on multimodal models: Tasks, models, and system schemes. International Journal of Intelligent Systems, 2025(1), 5124400.
3. Lee, D. (2025). Design consideration of autonomous robots based on embodied intelligence. Clinical Research and Clinical Trials, 12(3).
4. Ren, L., Dong, J., Liu, S., et al. (2024). Embodied intelligence toward future smart manufacturing in the era of AI foundation model. IEEE/ASME Transactions on Mechatronics.
5. Xu, J., Sun, Q., Han, Q. L., et al. (2025). When embodied AI meets Industry 5.0: Human-centered smart manufacturing. IEEE/CAA Journal of Automatica Sinica, 12(3), 485-501.
6. Zhao, W., & Yuan, Y. (2025). Development of intelligent robots in the wave of embodied intelligence. National Science Review, 12(7), nwaf159.
7. Dong, W., Li, S., & Zheng, P. (2025). Toward embodied intelligence-enabled human–robot symbiotic manufacturing: A large language model-based perspective. Journal of Computing and Information Science in Engineering, 25(5), 050801.
8. Lisondra, M., Benhabib, B., & Nejat, G. (2025). Embodied AI with foundation models for mobile service robots: A systematic review. arXiv. https://doi.org/10.48550/arXiv.2505.20503
9. Bu, Q., Cai, J., Chen, L., et al. (2025). Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. arXiv. https://doi.org/10.48550/arXiv.2503.06669
10. Song, R., Majeed, A. P. P. A., Wang, L., et al. (2024). Potential implementation of embodied intelligence technology in the power grid: A review. In International Conference on Intelligent Manufacturing and Robotics (pp. 799-814). Springer Nature Singapore.

Biographies

1. **Weifeng Zhao** Bachelor, currently serves as General Manager and Technical Director of Foshan Rossum Robotics Co.,Ltd. with main research directions including artificial intelligence, robots and automatic control.

面向工業場景的具身智能移動操作機器人大模型設計與泛化能力研究

趙偉峰[1]

[1]佛山隆深機器人有限公司，佛山，中國，528000

摘要：針對多工位、多工藝的工業場景，傳統規則控制以及小規模策略網絡在面對任務擴展和設備差異時，容易出現性能下降的情況。於是，研究創建了具身智能移動操作平台，該平台統一了移動底盤、機械臂以及多模態傳感的觀測—動作接口，還設計出一種工業級大模型，此模型整合了視覺、點雲、力覺和語言指令，並藉由預訓練和指令微調進行改進。

以真實車間任務庫為依據開展的實驗顯示，該模型在源域多任務中的表現與人工操作精度較為接近，且在零樣本及少樣本情境下，對未見過的任務合成和工位佈局具有較高的適應能力。消融實驗與工程案例進一步驗證了多模態融合、分層動作及泛化改進在節拍、人工工時和安全方面的效益，體現出其具可複製性的工程應用價值。

關鍵詞：具身智能；移動操作機器人；工業大模型；多模態感知；指令微調；泛化能力

1. 趙偉峰，信息管理与信息系统（大数据）專業學士，現任佛山隆深机器人有限公司总经理、技术总监，主要研究方向包括人工智能、机器人和自动化控制。