2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MIMO: A Multimodal Imitation Learning Framework for Mobile Manipulation with Exoskeleton-VR Teleoperation

Jie Mei, Xinkai Wu, Yue Zhang, Tao Song, Zhongxia Xiong*

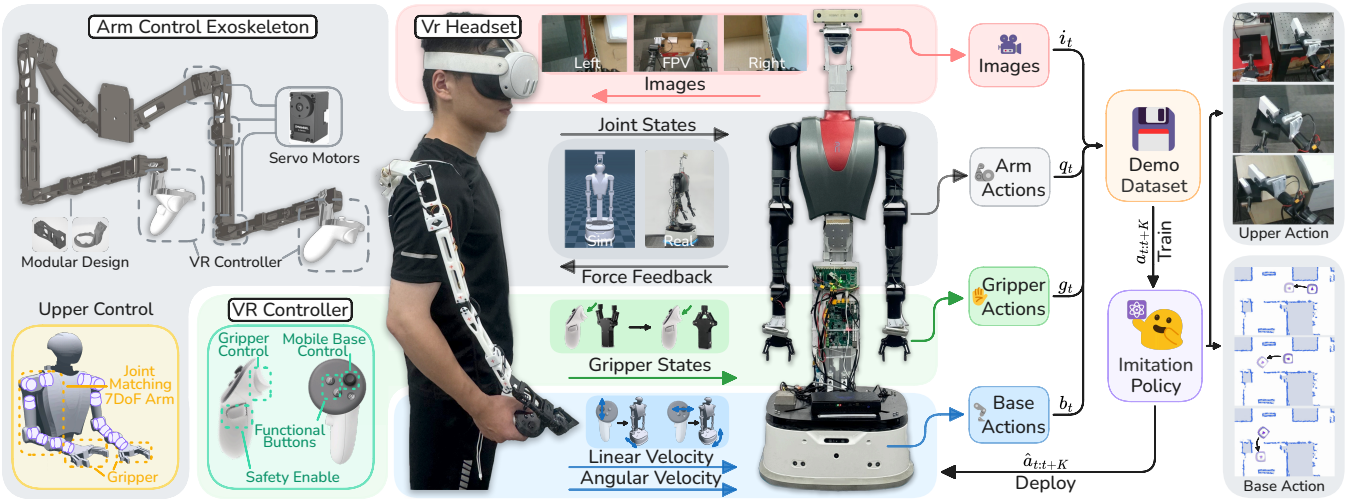

Fig. 1: Overview of the MIMO framework. The integrated exoskeleton-VR teleoperation system enables single-operator whole-body control of the robot with simple force feedback, collecting multimodal data (including multi-view images, arm actions, gripper actions, and base actions) to form a demonstration dataset. This dataset is used to train an imitation policy, which is then deployed to generate sequential actions for the robot's whole-body mobile manipulation tasks.

***Abstract*— In whole-body mobile manipulation, existing teleoperation systems often suffer from high complexity and cost, while imitation learning approaches are frequently limited by insufficient modeling of long-horizon action sequences and inadequate fusion of multi-receptive-field visual features. These constraints significantly hinder the collection of high-quality demonstration data and the effective transfer of complex robotic skills. To address these challenges, this paper proposes an integrated exoskeleton-VR teleoperation system that enables single-operator whole-body control of mobile manipulators with basic force feedback, substantially reducing the cost of data collection while improving demonstration quality. Furthermore, we introduce MIMO, an encoder–decoder imitation learning framework, which incorporates an Efficient Context Modeling Network (ECM-Net) based on linear-complexity temporal modeling to mitigate error accumulation in long-horizon tasks, and a Multi-Receptive Field Fusion Network (MRF-Net) that employs dual-path attention to achieve precise alignment between multi-scale visual cues and motion phases. Real-world experiments on a mobile manipulator demonstrate that MIMO consistently outperforms state-of-the-art baselines across multiple whole-body mobile manipulation tasks, confirming its effectiveness in long-horizon, fine-grained robotic control.**

I. Introduction

Whole-Body mobile manipulation entails coordinated control of a robot's arms and base, which is essential for executing complex tasks in unstructured environments [1]–[3]. Imitation Learning offers a viable approach to acquiring such skills from expert demonstrations [4]–[7]. However, the performance of this paradigm depends on both the collection of high-quality demonstration data and long-horizon temporal reasoning and multimodal alignment algorithms. Enhancing its performance faces the following challenges:

Inadequate Teleoperation for Whole-Body Data Collection. Current teleoperation systems fail to balance intuitiveness, cost, and whole-body control. Low-cost heterogeneous devices often incur kinematic inaccuracies and lack force feedback, compromising demonstration quality [4], [8]. Although high-fidelity isomorphic manipulators allow precise control, they are immobile, costly, and often require multiple operators, impairing demonstration consistency [9], [10]. A low-cost, force-feedback-enabled, whole-body teleoperation system operable by a single user remains an unmet need.

Deficiencies in Long-Horizon Modeling and Visual-Motor Alignment. Algorithms struggle with long-term temporal dependencies and precision required in whole-body tasks. Behavioral Cloning (BC) methods suffer from covariate shift and error accumulation over long sequences [11], [12]. While Transformer-based methods excel at capturing long-range contextual dependencies, their quadratic computational complexity inhibits real-time deployment on robotic platforms [13]–[15]. Crucially, a receptive field gap exists between global navigation cues and local hand-eye

This research was supported by National Key R&D Program of China (2023YFC3805400). The authors are members of the School of Transportation Science and Engineering, Beihang University, Beijing, 100191, China. *(e-mail: xiongzhongxia@buaa.edu.cn).*

coordination. Most methods homogenize multi-scale visual features without explicit mechanisms to align "macro-micro" visual semantics with corresponding motion phases, resulting in incoherent planning and execution [4], [10].

To address these issues, this paper proposes MIMO, an end-to-end framework from exoskeleton-VR teleoperation to multimodal imitation which is shown in Fig. 1. The core contributions of this paper are as follows:

- An Integrated Teleoperation Platform combining an exoskeleton with VR controllers for intuitive single-operator whole-body control, lowering data collection costs and improving safety via simple force feedback.
- A Co-Designed Imitation Learning Algorithm MIMO, which introduces an efficient encoder (ECM-Net) for linear-complexity temporal modeling and a multi-receptive field decoder (MRF-Net) for enhanced cross-modal coordination.
- Extensive Real-World Validation on a mobile manipulator, showing MIMO's superior success rates over state-of-the-art methods in multiple complex tasks.

The remainder of this paper is organized as follows. Section II discusses the related work. Section III details the integrated exoskeleton-VR teleoperation system and the design of MIMO. Section IV introduces the experimental setup and results analysis. Section V concludes the paper.

II. Related Work

A. Whole-Body Teleoperation Systems

Teleoperation systems enable robots to perceive and interact with the environment through remote control interfaces and have become a key bridge for data collection and skill transfer in robot imitation learning. Early solutions typically used two identical robotic arms to construct master-slave control systems [9], [16]. While these achieved high-precision operation mapping, they suffered from high equipment costs and difficulty in scaling to large mobile robot platforms. To reduce costs, researchers introduced pose capture solutions based on lightweight devices such as JoyCon [17], vision cameras [18], [19] or VR [20]–[23], generating robot joint control commands through inverse kinematics solving. However, limited by sensor accuracy and kinematic solution uncertainties, these methods fail to ensure reliable end-effector pose accuracy. Furthermore, the absence of force feedback causes non-intuitive operation, adjustment delays, and degraded quality of demonstration data.

In recent years, teleoperation devices based on isomorphic exoskeletons have improved control intuitiveness and accuracy while effectively reducing hardware costs by matching human and robot joint structures [8], [10], [24]. However, these systems are typically designed only for the robotic arm and cannot achieve unified control of the whole robot body, still relying on other input devices or multiple operators, limiting their application in complex tasks. Addressing these issues, this paper proposes a novel whole-body teleoperation device that integrates an isomorphic exoskeleton arm and VR controllers, providing a first-person perspective and simple force feedback, enabling intuitive and precise whole-body control of the robot by a single operator.

B. Whole-Body Imitation Learning

Imitation learning is a core paradigm for robots to acquire complex manipulation skills from expert demonstrations. As its foundational form, Behavioral Cloning achieves imitation via a direct observation-to-action mapping [25]. To boost model performance, early research focused on enhancing temporal modeling: RNNs, LSTMs and related architectures were introduced to capture dynamic dependencies in action sequences [12], [26], [27]. A parallel line of work improved policy robustness through training objective regularization, including adversarial loss and uncertainty modeling [28], [29]. While these advances perform well in simple manipulation tasks, they fail to overcome error accumulation caused by covariate shift in long-horizon, high-precision tasks, limiting their stability in complex scenarios [30].

As task complexity increased, the research focus gradually shifted to the generalization and multi-task transfer capabilities of imitation learning. Multi-task learning enhanced the model's adaptability to new goals and scenarios by sharing feature extraction structures and incorporating language instructions as conditional signals [31]. Few-shot learning, leveraging meta-learning mechanisms, enabled models to quickly generalize to new tasks from a small number of demonstrations [32]–[34]. However, when faced with complex operations requiring whole-body coordination and high precision, existing methods still exhibit significant shortcomings. On one hand, the lack of differential modeling between arm and base actions limits cross-modal representation alignment capability. On the other hand, the error accumulation problem in long action sequence prediction remains unresolved, restricting the policy's generalization and reliability in real-world environments.

To tackle these challenges, this paper introduces the MIMO framework, which enhances action coherence in long-horizon tasks through efficient temporal modeling (ECM-Net) and improves mobile fine-manipulation accuracy via multi-receptive-field visual-motor alignment (MRF-Net). This integrated approach offers a robust solution for high-precision whole-body imitation learning.

III. Methodology

A. Integrated exoskeleton-VR teleoperation system

Teleoperation has become a widely adopted method for collecting robot demonstration data to train learning policies. For structured, table-top fixed tasks, existing teleoperation solutions can intuitively and reliably complete demonstration collection. However, when facing whole-body control tasks involving wheeled mobile robots, bipedal humanoid robots, and others with high degrees of freedom, dynamically unstable states, and requirements for interaction in unstructured environments, traditional teleoperation methods face significant challenges. Operators often require additional hardware devices and complex action combinations to achieve whole-body coordination, which not only increases the operational

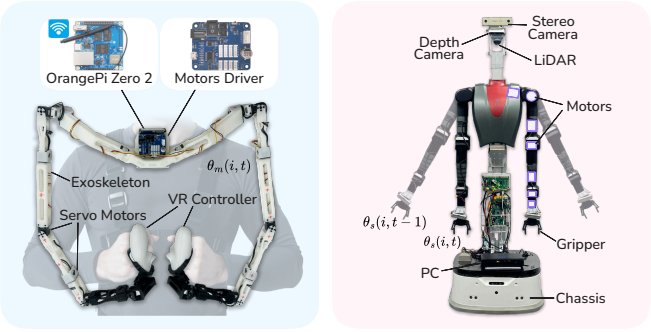

Fig. 2: *left:* The integrated exoskeleton-VR teleoperation system. *right:* The wheeled mobile robot Robint. The system enables joint-level mapping between the operator's arm motions and the robot's arms, providing real-time position tracking and simple force feedback through master–slave position monitoring.

burden but also reduces control precision and task execution accuracy. To address these limitations, inspired by [10], this paper proposes a novel whole-body teleoperation system integrating exoskeleton and VR technologies.

(1) System Overview: The system uses an exoskeleton to capture the operator's arm joint states in real-time and map them to the robot's arms, achieving joint-level precise control while providing simple force feedback to enhance operational perception, as shown in Fig. 2. VR controllers are used to coordinate the robot's overall movement, and a VR headset provides a real-time first-person view from the robot, offering immersive environmental feedback to the operator. This system allows a single operator to achieve coordinated whole-body control of a humanoid robot without additional devices or large body movements, improving intuitiveness, coordination, and precision while reducing demonstration data collection costs and improving data quality.

(2) Hardware: The hardware architecture is designed for integrated whole-body motion capture and interactive control, comprising four main components: the exoskeleton manipulation module, VR control unit, central control and data transmission system, and power management unit.

The **Exoskeleton Manipulation Module** employs a dual-arm design with each arm featuring 7 DoF, kinematically matching the human arm structure. It integrates servo motors with motors drivers and high-precision encoders, achieving an optimal balance among cost, size, weight, and accuracy. This module captures joint angles in real-time and provides basic force feedback. Its modular design facilitates adaptation to various robotic joint configurations and supports both wearable and fixed operational modes.

The **VR Control Unit** utilizes a Meta Quest 3 headset and controllers, acquiring real-time joystick and button states. Left and right controller inputs are mapped to the robot's base linear and angular velocities, respectively; the Trigger button controls gripper aperture; and the Grip button acts as a safety-enabled switch, ensuring command transmission only during active operation, thereby enhancing system reliability. The VR headset provides a first-person visual feedback from the robot's perspective, significantly improving operational intuitiveness and accuracy. Control data is transmitted via either wired ADB or wireless WiFi.

The **Central Control** and **Data Transmission System** uses an OrangePi Zero 2 as the core controller, which gathers sensor data from the exoskeleton via serial communication and transmits it to the robot via WiFi. The **Power System** employs an integrated lithium battery, enabling cordless operation and ensuring unrestricted mobility and comfort.

(3) Software: The system software is based on ROS, using a distributed node architecture for functional decoupling and inter-module communication. The Grip button functions as a safety enable switch, with control instructions output only when pressed. The Trigger button's analog input is linearly mapped to gripper opening commands. Inputs from left/right joysticks are parsed in real-time to generate linear/angular velocity controls for base movement. To improve data quality, the system applies real-time low-pass filtering to raw sensor signals to reduce high-frequency noise, syncs multi-source data between the exoskeleton and VR devices via timestamp alignment, and ensures spatiotemporal consistency between joint motion and control commands.

To achieve safe and reliable human-robot interaction, the system introduce a simple force feedback method based on overload detection mechanism. When overload of the humanoid robot's end-effector is detected, the exoskeleton locks its current position to apply simple force feedback to the operator, ensuring operational safety, helping the operator perceive the contact state between the robot and the environment. The overload criterion is defined as:

$$\Delta\Theta(t) = \sum_{i=1}^{n} w(i) \cdot |\theta_m(i,t) - \theta_s(i,t-1)| \quad (1)$$

Where $\theta_m(i,t)$ denotes the angle of the i-th joint of the master arm (exoskeleton arm) at time t; $\theta_s(i,t-1)$ denotes that of the i-th joint of the slave arm (robot arm) at time step $t-1$; and $w(i)$ is the weight coefficient for the i-th joint, higher weights are allocated to more critical root joints. The exoskeleton lock is triggered when $\Delta\Theta(t) > \tau_{\text{total}}$, where τ_{total} is the preset safety threshold.

The system employs a modular and distributed design to achieve collaborative control and real-time interaction across multiple hardware units. It offers high scalability and reliability, thereby providing crucial support for data collection and safety control in whole-body teleoperation.

B. Whole-Body Multimodal Imitation Learning Framework

With the rapid development of humanoid robots, their whole-body control algorithms provide new solutions for complex tasks requiring simultaneous movement and operation in the real world. Traditional fixed and table-top manipulation tasks can no longer adapt to complex dynamic environments and varying interaction needs.

Action trajectories in whole-body mobile manipulation tasks exhibit clear phased evolution characteristics: behavioral patterns differ significantly across stages, and state transitions strongly depend on historical information. Such

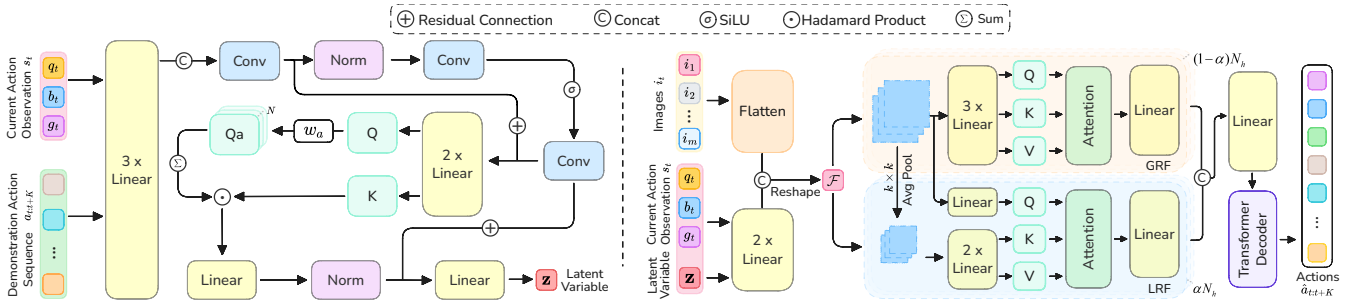

Fig. 3: *left:* ECM-Net Architecture: Captures long-range temporal dependencies in action sequences with linear complexity to mitigate long-horizon error accumulation. *right:* MRF-Net Architecture: Uses dual-path attention (Global Receptive Field, GRF, for macro motion planning; Local Receptive Field, LRF, for fine manipulation) to achieve precise cross-modal alignment between multi-scale visual features and action semantics, improving action generation accuracy.

complex temporal dynamics require the model to selectively memorize key states and forget redundant information to maintain action coherence and physical feasibility, thereby mitigating error accumulation in long-sequence prediction. Meanwhile, visual observation information, which forms the basis of action decision-making, has a significant spatial receptive field differentiation in its impact on action generation: large receptive field features dominate macro motion planning and overall behavioral fluency, while small receptive field features directly affect the fine operation accuracy of the end-effector. Therefore, to achieve the generation of high-quality fine manipulation actions, a new architecture that can collaboratively model the phased evolution of actions and multi-scale visual features is urgently needed.

This paper proposes a novel end-to-end robot whole-body imitation learning network MIMO, as shown in Fig. 3. The network takes RGB images $i_t = \{i_1, i_2, ..., i_m\}$ from multiple robot cameras (m is the number of cameras), the current joint position state q_t, gripper state g_t, and base velocity b_t as input, outputs the target joint position action sequence $\hat{a}_{t:t+K}$ for the next K timesteps, driving the robot to perform fine mobile manipulation.

(1) Overview: The MIMO network is based on a Conditional Variational Autoencoder (CVAE) framework [35]. During training, the encoder receives the demonstration action sequence $a_{t:t+K}$ and the current observation $\mathbf{s}_t = \{q_t, g_t, b_t\}$, outputting the distribution parameters of the latent variable $\mathbf{z} \sim \mathcal{N}(\mu_z, \sigma_z^2)$. The decoder then reconstructs the predicted action sequence $\hat{a}_{t:t+K}$ based on the current state $\mathbf{s}_t$, i_t and the latent variable $\mathbf{z}$ sampled from this distribution. During inference, the encoder is removed, the latent variable $\mathbf{z}$ is set to the prior mean (zero vector), and the decoding process becomes deterministic forward propagation. The loss function is the standard CVAE objective, including the L1 loss of the action sequence and the KL divergence between the latent variable $\mathbf{z}$ and the standard normal prior. Algorithms 1 and 2 outline MIMO's training and inference processes.

(2) Encoder: In generative model-based robot imitation learning, the core function of the encoder is to extract robust latent policy representations from long-horizon action demonstrations. Traditional methods like RNNs and LSTMs struggle to effectively capture long-range temporal dependencies due to vanishing or exploding gradient problems. While the self-attention mechanism of the standard Transformer can model long-range relations, its quadratic computational complexity creates significant efficiency bottlenecks when processing long sequences, limiting its application in real-time robot systems.

Algorithm 1 MIMO Training

Input: Demonstration data $\mathcal{D} = \{a_{t:t+K}\}$, current observation $\mathbf{s_t}$, images i_t and prediction time step K
Initialize: ECM-Net $\pi_\phi(\mathbf{z}|\mathbf{s}_t, a_{t:t+K})$ and MRF-Net $p_\theta(\hat{a}_{t:t+K}|\mathbf{s}_t, i_t, \mathbf{z})$

for iteration $n = 1, 2, \ldots$ **do**
  Sample $\mathbf{s}_t$, $a_{t:t+K}$ from $\mathcal{D}$
  Calculate $\mathbf{z}$ from $\pi_\phi(\mathbf{z}|\mathbf{s}_t, a_{t:t+K})$
  Predict $\hat{a}_{t:t+K}$ from $p_\theta(\hat{a}_{t:t+K}|\mathbf{s}_t, i_t, \mathbf{z})$
  Update ϕ, θ with ADAM
end for

Algorithm 2 MIMO Inference

Input: trained θ, the duration of the task T
Initialize: Buffers $\mathcal{B}[0:T]$, where $\mathcal{B}[T]$ stores actions predicted for timestep t, attenuation factor m

for iteration $n = 1, 2, \ldots, T$ **do**
  Predict $\hat{a}_{t:t+K}$ from $p_\theta(\hat{a}_{t:t+K}|\mathbf{s}_t, i_t, \mathbf{z})$ where $\mathbf{z} = 0$
  Store in buffer $\mathcal{B}[0:T] \leftarrow \hat{a}_{t:t+K}$
  Apply $\hat{a}_t = \sum_i w_i \cdot \mathcal{B}[i] / \sum_i w_i$, where $w_i = \exp(-m \cdot i)$
end for

To address these issues, we propose the Efficient Context Modeling Network (ECM-Net), which achieves efficient modeling of global context with linear computational complexity, significantly reducing computation and memory overhead. This mechanism simplifies the Query-Key-Value interaction structure, retains Query-Key interaction combined with linear transformation, and effectively captures long-range dependencies in the sequence while avoiding expensive

matrix operations. The specific process is as follows:

First, the input action sequence $a_{t:t+K}$ and current state q_t are projected into an embedding space via independent linear layers. After processing by convolutional and normalization layers with residual connections, they are linearly projected to generate Query (Q) and Key (K) matrices. To capture the global importance distribution of queries, a learnable parameter vector w_a is introduced: get the initial weights $Q_a = Qw_a$, and a global query vector is generated through weighted summation. This global query vector is element-wise multiplied with K; the result, after linear projection, normalization, is added back to the original Q to retain local features. Finally, the output dimension is adjusted via a linear layer to obtain the latent space representation.

This design explicitly enhances the encoder's ability to model dynamic transitions between action phases, effectively memorizing key states and filtering redundant historical information, thereby significantly mitigating error accumulation in long-sequence prediction and improving the coherence of action generation and overall task success rate.

(3) Decoder: The decoder combines the latent variable and current robot state to generate high-quality long-horizon action sequences. For whole-body mobile manipulation tasks with phased dynamic action characteristics, it must inherit the encoder's ability to capture long-range temporal dependencies and accurately map multi-scale visual receptive fields to action semantics. Neglecting the synergy between differentiated visual receptive fields and action temporal dynamics impairs action coherence, physical feasibility, and success rate.

We propose the Multi-Receptive Field Fusion Network (MRF-Net), which takes current observation $\mathbf{s}_t$ and latent variable $\mathbf{z}$ as input to output predicted action sequence $\hat{a}_{t:t+K}$. Its core innovation is a dual-path attention mechanism: the Local Receptive Field (LRF) pathway uses local window attention to focus on fine visual structures for precise end-effector control; the Global Receptive Field (GRF) pathway extracts macro visual features via pooling and global attention to ensure overall action coherence and coordination. The two pathways are dynamically weighted and fused to explicitly link visual scales with action semantics, supporting long-horizon action generation. Implementation details are provided below, with key steps as follows:

Visual Feature Extraction and Cross-Modal Fusion. Each RGB frame is processed by a ResNet [36] (with independent weights) to extract spatial features, which are flattened and augmented with 2D sinusoidal positional encoding. Joint state q_t and $\mathbf{z}$ are projected into feature spaces via linear layers. All visual, state, and latent features are concatenated along the sequence dimension to form global input sequence $\mathbf{X}_{MRF} = [\mathbf{X}_1^{\text{vis}}, \cdots, \mathbf{X}_m^{\text{vis}}, \mathbf{X}_q, \mathbf{X}_g, \mathbf{X}_b, \mathbf{X}_z]$, achieving unified cross-modal representation (integrating vision, state, and action styles) and avoiding information loss from modal isolation in traditional methods.

Multi-Receptive Field Attention Encoding. The concatenated sequence is reshaped into 2D feature map $\mathcal{F}$. Total attention heads N_h are split into GRF and LRF pathways by ratio α. GRF captures environmental layout and global motion temporal correlations via pooling and global attention; LRF performs attention within local windows (non-overlapping $k \times k$ regions from $\mathcal{F}$ via average pooling) to perceive fine end-effector-target interactions. They support macro planning and fine operation, respectively. Their outputs are concatenated, fused, and linearly projected to get $\mathbf{O}_{MRF}$. Impacts of α and k are discussed in the experiment's sensitivity analysis of key hyperparameters.

Action Sequence Decoding and Optimization. A standard Transformer decoder uses $\mathbf{O}_{MRF}$ as Key/Value and fixed positional embeddings as Query to predict $\hat{a}_{t:t+K}$. Decoder layers integrate multi-scale visual context and action semantics via cross-attention. The output is projected into the action space via an MLP to generate action chunks. Weighted averaging of overlapping chunks enhances action smoothness.

IV. Experiments

A. Data Collection

We conducted real-world experiments to validate the performance of the MIMO. The real robot used is Robint, a dual-arm wheeled robot, as shown in Fig2. Robint has a long arm span and is motor-driven, suitable for large-range and large-payload tasks. We used the exoskeleton-VR teleoperation system proposed earlier in this paper to collect robot actions and images from three cameras (head, left wrist, right wrist), recording 50 demonstrations for each task with a unified data sampling frequency of 50 Hz.

All tasks require fine-grained coordination between both arms and the base. During collection, human operators could use different strategies to complete tasks. Each action within an operation was essentially random, no strict consistency in actions was enforced.

B. Tasks

Box Stowing: A white plastic box and a black storage bin are placed on a table. Both arms grasp them respectively *(Subtask1 Grasp)* and place the box into the bin *(Subtask2 Place)*. The robot moves to a target position, and the right arm grasps another plastic box and places it into the bin *(Subtask3 Transfer)*. Due to the randomness of object and initial robot positions, and the small clearance between the gripper and the box, slight errors in bimanual coordination, post-movement operation, and multi-subtask transition can accumulate and cause collisions and task failure.

Fig. 4: Setup of box stowing task. The positions of the two white boxes are randomly placed within a 15cm radius, and the robot's initial position is also randomly initialized within a 30cm radius. The total base movement distance is about 3.5m, and the task duration is 40s. This task is conducted in a noisy and complex environment.

TABLE I: Task success rate comparison. Results for our full method (MIMO), baseline methods, and ablated versions (ECM-Net Only, MRF-Net Only) are reported. The last column reports the final success rate, which requires all subtasks to be completed successfully. **Bold** denotes the highest score in each column.

Method	Box Stowing			Box Transport			Object Sorting		
	Grasp	Place	Transfer	Grasp	Move	Place	Identify	Move	Place
BeT	40	28	0	38	0	0	32	0	0
DP	70	60	56	76	64	60	86	76	72
ACT	76	64	56	**82**	70	66	84	80	76
ECM-Net Only	74	66	58	80	68	64	82	78	74
MRF-Net Only	72	62	52	78	66	62	84	80	76
MIMO (Ours)	**78**	**72**	**64**	**82**	**76**	**72**	**88**	**86**	**80**

Box Transport: A turnover container is placed on the table, and both arms cooperate to grasp it *(Subtask1 Grasp)*. The robot moves to a target position *(Subtask2 Move)* and places the container on the table *(Subtask3 Place)*. The robot's movement path is long and passes through a narrow door. Any path planning deviation, manipulator posture control error, or positioning error may cause collision with the door or placement position offset, leading to task failure.

Fig. 5: Setup of box Transport task. The box is randomly placed within a 15cm radius, and the robot's initial position is randomly initialized within a 30cm radius. The total base movement distance is about 7m, and the task duration is 30s.

Object Sorting: White and black boxes are mixed and placed on a table. Both arms grasp boxes of different colors respectively *(Subtask1 Identify)*, move sequentially to the corresponding colored target containers *(Subtask2 Move)*, and place the black box in the black box and the white box in the white box *(Subtask3 Place)*. Task failure occurs if the color is misjudged during grasping, the box drops during movement, or it is placed into the wrong colored container.

Fig. 6: Setup of object sorting task. The boxes of both colors are randomly mixed within a 20cm radius. The robot's initial position is randomly initialized within a 30cm radius. The total base movement distance is about 1.5m, and the task duration is 30s.

C. Experiment Settings

We compared the MIMO with other imitation learning algorithms. BeT [12] discretizes continuous actions via k-means clustering to predict multimodal actions, but its visual perception and control networks are not jointly optimized; Diffusion Policy (DP) [32] models the probability distribution of actions through a diffusion process, gradually refining action predictions through network training; ACT [5] is an action generation method based on a Transformer backbone, whose encoder and decoder use identical encoder layers.

Subtask success rates are recorded. Each run utilized 1 seed with 50 evaluations. All experiments are conducted on a desktop computer with an i9-14900KF CPU and an RTX 5090 GPU. Hyperparameter settings are shown in Table II. Other parameters followed the original papers.

TABLE II: Hyperparameter settings.

Model	Learning Rate	Batch Size	Epochs	Other Parameters
BeT	3e-4	128	100	Momentum = 0.9
DP	1e-4	32	4000	Ema power = 0.75
ACT	1e-5	16	4000	Dropout = 0.1
MIMO	1e-5	16	4000	$\alpha = 0.4$, $k = 6$

Quantitative comparative experiments demonstrate that MIMO significantly outperforms all baseline methods across multiple real-world fine-manipulation tasks, effectively validating the efficacy of its core architectural designs, as summarized in Table I. Ablation studies confirm the individual contributions of each component: ECM-Net Only shows notable gains in long-horizon subtasks like Transfer (58%), while MRF-Net Only excels in visually-sensitive tasks like Identify (84%). However, neither variant matches the full MIMO framework, underscoring the necessity of their integration.

In Box Stowing, MIMO achieved a success rate of 64% in the challenging Transfer subtask, exceeding ACT by 8 percentage points, a improvement attributable to ECM-Net's linear-complexity contextual modeling which reduces error accumulation. In Box Transport, MIMO substantially outperformed ACT in Move (76% vs. 70%) and Place (72% vs. 66%), highlighting the GRF pathway's role in integrating large-receptive-field visual cues for coherent navigation. In Object Sorting, MIMO's superior performance in Identify (88%) and Place (80%) is largely due to the LRF pathway's focus on localized visual features for precise discrimination.

Furthermore, MIMO exhibited stronger generalization under varying initial conditions, facilitated by both its algorithmic design and the high-quality demonstration data collected via the proposed exoskeleton-VR teleoperation

system. In summary, the results confirm that MIMO's structure—featuring efficient long-sequence modeling and multi-receptive-field visual-motor fusion—significantly enhances performance in mobile manipulation tasks requiring both long-horizon coordination and fine motor control.

D. Sensitivity Analysis of Key Hyperparameters

To investigate the sensitivity of key hyperparameters in the MIMO, we conducted systematic experiments on two core parameters in the MRF-Net decoder: the Receptive Field Fusion Ratio α and the Local Window Size k . The success rates of different parameter combinations on the Box Stowing task are summarized in Fig. 7.

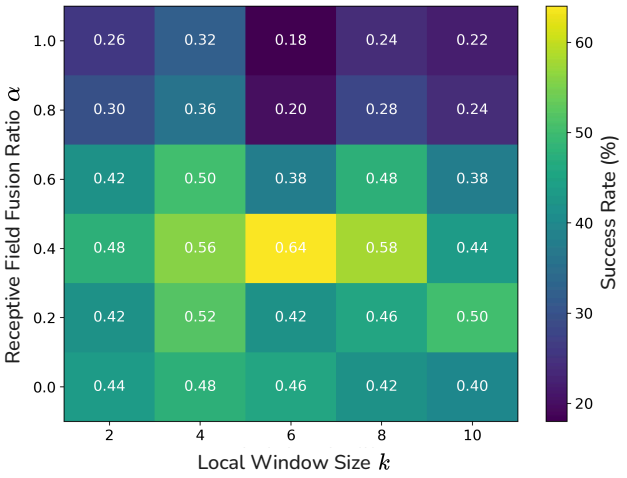

Fig. 7: Heatmap of the final success rate for the Box Stowing task under different values of receptive field fusion ratio α and local window size k, with each parameter combination tested 50 times. Warmer colors indicate higher success rates. Parameter combinations significantly affect task success rate, with the highest (0.64) at $\alpha = 0.4$ and $k = 6$.

Receptive Field Fusion Ratio (α). It regulates the proportion of attention heads assigned to the GRF pathway, thereby balancing the model's focus between global navigation cues and local hand-eye coordination features. Experimental results indicate that allocating all attention heads to the GRF pathway ($\alpha = 1$) leads to a significant performance decline in fine manipulation subtasks (e.g., Grasp and Place), due to insufficient capture of fine-grained visual details. Conversely, assigning all heads to the LRF pathway ($\alpha = 0$) results in unstable performance in long-horizon navigation and coordination tasks, owing to the lack of holistic environmental perception. Optimal performance is achieved at $\alpha = 0.4$, where the GRF pathway ensures coherent whole-body motion planning while the LRF pathway provides the necessary detail perception for precise end-effector control.

Local Window Size (k). It defines the spatial range for local feature aggregation in the LRF pathway. Excessively small windows introduce extraneous context, dilute focus on critical local features, and raise fine manipulation failure rates. Overly large windows, by contrast, restrict the receptive field and impair model's capture of essential spatial relationships between the end-effector and interaction objects. Results show $k = 6$ achieves an optimal balance: enabling accurate local detail perception while preserving relevant context, thus maximizing performance in tasks.

The results underscore the importance of the dual-path design. The optimal combination ($\alpha = 0.4$, $k = 6$) enables effective synergy between global navigation and precise manipulation, which is critical to the model's high performance.

E. Efficiency Analysis

We evaluate the computational efficiency and parameter economy of MIMO against the top-performing baseline, ACT, across two representative tasks. The results, summarized in Table III, demonstrate the superior efficiency of the proposed approach.

TABLE III: Efficiency & Parameter Comparison (NVIDIA RTX 5090). Best results are in **bold**.

Metric Category	Detail	Task	ACT	MIMO (Ours)
Inference Latency (ms)	Min.	Box Transport	0.48	**0.47** (-2.1%)
		Object Sorting	0.47	**0.45** (-4.3%)
	Avg.	Box Transport	1.94	**1.89** (-2.6%)
		Object Sorting	1.82	**1.79** (-1.6%)
	Max.	Box Transport	2.68	**2.57** (-4.1%)
		Object Sorting	2.77	**2.64** (-4.7%)
Model Params (M)	-	-	106.26	**103.77** (-2.3%)

Computational Efficiency. MIMO consistently achieves lower inference latency across all evaluated metrics. For the Box Transport task, the average inference time is reduced by 2.1% (0.47 ms vs. 0.48 ms), while the maximum latency sees a more substantial decrease of 4.1% (2.57 ms vs. 2.68 ms). Similar improvements are observed in the Object Sorting task, with the average and maximum latency reduced by 1.6% and 4.7%, respectively. These improvements can be attributed to the linear-complexity context modeling in ECM-Net, which processes long action sequences more efficiently than the quadratic self-attention mechanism used in ACT.

Parameter Economy. MIMO also exhibits higher parameter efficiency, utilizing only 103.77M parameters—a reduction of 2.3% compared to ACT's 106.26M. This reduction is achieved through the asymmetric encoder–decoder design: the lightweight ECM-Net captures essential temporal dependencies without redundant parameters, while the dual-path structure in MRF-Net enables effective multi-scale feature fusion with minimal parameter overhead.

In summary, MIMO not only enhances task performance but also improves computational and memory efficiency, making it particularly suitable for deployment on resource-constrained robotic platforms.

V. CONCLUSIONS

This study addresses two key challenges in whole-body mobile manipulation: inadequate teleoperation for high-quality demonstrations and imitation learning limitations in long-sequence modeling and visual-action alignment.

First, we developed an exoskeleton-VR teleoperation system for single-operator whole-body control, featuring simple force feedback and data synchronization to ensure high-quality demonstration data. Second, we proposed the MIMO framework: its ECM-Net uses linear-complexity modeling to reduce long-sequence errors, while MRF-Net's dual-path attention resolves visual-action misalignment.

Experiments on a wheeled mobile robot showed MIMO outperformed baselines in various real-world tasks, with lower latency and fewer parameters.

Future work will upgrade teleoperation with high-precision force sensors, add language for cross-task generalization, and validate MIMO in dynamic environments. This study provides a complete solution to advance robot deployment in real-world mobile manipulation tasks.

REFERENCES

[1] M. Mittal, D. Hoeller, F. Farshidian, M. Hutter, and A. Garg, "Articulated object interaction in unknown scenes with whole-body mobile manipulation," in *2022 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2022, pp. 1647–1654.

[2] J. Hu, P. Stone, and R. Martín-Martín, "Causal policy gradient for whole-body mobile manipulation," *arXiv preprint arXiv:2305.04866*, 2023.

[3] A. Purushottam, C. Xu, Y. Jung, and J. Ramos, "Dynamic mobile manipulation via whole-body bilateral teleoperation of a wheeled humanoid," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1214–1221, 2023.

[4] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[5] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation using low-cost whole-body teleoperation," in *8th Annual Conference on Robot Learning*, 2024.

[6] M. Murooka, T. Hoshi, K. Fukumitsu, S. Masuda, M. Hamze, T. Sasaki, M. Morisawa, and E. Yoshida, "Tact: Humanoid whole-body contact manipulation through deep imitation learning with tactile modality," *IEEE Robotics and Automation Letters*, 2025.

[7] P. Sundaresan, R. Malhotra, P. Miao, J. Yang, J. Wu, H. Hu, R. Antonova, F. Engelmann, D. Sadigh, and J. Bohg, "Homer: Learning in-the-wild mobile manipulation via hybrid imitation and whole-body control," *arXiv preprint arXiv:2506.01185*, 2025.

[8] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu, "Airexo: Low-cost exoskeletons for learning whole-arm manipulation in the wild," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 15 031–15 038.

[9] P. Wu, Y. Shentu, Q. Liao, D. Jin, M. Guo, K. Sreenath, X. Lin, and P. Abbeel, "Robocopilot: Human-in-the-loop interactive imitation learning for robot manipulation," *arXiv preprint arXiv:2503.07771*, 2025.

[10] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang, "Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit," *arXiv preprint arXiv:2502.13013*, 2025.

[11] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[12] N. M. Shafiullah, Z. Cui, A. A. Altanzaya, and L. Pinto, "Behavior transformers: Cloning k modes with one stone," *Advances in neural information processing systems*, vol. 35, pp. 22 955–22 968, 2022.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 17 425–17 436.

[15] Z. Pan, J. Cai, and B. Zhuang, "Fast vision transformers with hilo attention," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 541–14 554, 2022.

[16] M. Hejrati and J. Mattila, "Robust immersive bilateral teleoperation of dissimilar systems with enhanced transparency and sense of embodiment," *arXiv e-prints*, pp. arXiv–2505, 2025.

[17] Y. Jiang, R. Zhang, J. Wong, C. Wang, Y. Ze, H. Yin, C. Gokmen, S. Song, J. Wu, and L. Fei-Fei, "Behavior robot suite: Streamlining real-world whole-body manipulation for everyday household activities," *arXiv preprint arXiv:2503.05652*, 2025.

[18] S. Pandian Arunachalam, S. Silwal, B. Evans, and L. Pinto, "Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation," *arXiv e-prints*, pp. arXiv–2203, 2022.

[19] T. Tao, M. K. Srirama, J. J. Liu, K. Shaw, and D. Pathak, "Dexwild: Dexterous human interactions for in-the-wild robot policies," *arXiv preprint arXiv:2505.07813*, 2025.

[20] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto, "Open teach: A versatile teleoperation system for robotic manipulation," *arXiv preprint arXiv:2403.07870*, 2024.

[21] S. Dass, W. Ai, Y. Jiang, S. Singh, J. Hu, R. Zhang, P. Stone, B. Abbatematteo, and R. Martín-Martín, "Telemoma: A modular and versatile teleoperation system for mobile manipulation," *arXiv preprint arXiv:2403.07869*, 2024.

[22] S. Yang, "Ace: A cross-platform visual-exoskeleton system for low-cost dexterous teleoperation," Master's thesis, University of California, San Diego, 2025.

[23] M. Seo, S. Han, K. Sim, S. H. Bang, C. Gonzalez, L. Sentis, and Y. Zhu, "Deep imitation learning for humanoid loco-manipulation through human teleoperation," in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*. IEEE, 2023, pp. 1–8.

[24] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 12 156–12 163.

[25] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.

[26] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 991–1002.

[27] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[28] P. De Haan, D. Jayaraman, and S. Levine, "Causal confusion in imitation learning," *Advances in neural information processing systems*, vol. 32, 2019.

[29] J. Pari, N. M. Shafiullah, S. P. Arunachalam, and L. Pinto, "The surprising effectiveness of representation learning for visual imitation," *arXiv preprint arXiv:2112.01511*, 2021.

[30] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on robot learning*. PMLR, 2022, pp. 894–906.

[31] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[32] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[33] Y. Duan, M. Andrychowicz, B. Stadie, O. Jonathan Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba, "One-shot imitation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[34] K. Bousmalis, G. Vezzani, D. Rao, C. Devin, A. X. Lee, M. Bauzá, T. Davchev, Y. Zhou, A. Gupta, A. Raju *et al.*, "Robocat: A self-improving generalist agent for robotic manipulation," *arXiv preprint arXiv:2306.11706*, 2023.

[35] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.