machines

MDPI

Article

Force Sensing Control for Physical Human–Robot Interaction: A Transformer-Based Action Chunking Approach

Zhenyu Pan and Weiming Wang *

School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; pan13855965084@sjtu.edu.cn
* Correspondence: wangweiming@sjtu.edu.cn

Abstract

In human–robot collaboration (HRC) scenarios with direct physical contact, accurately estimating human intentions and adjusting robot behaviors based on multimodal information is the core factors that restrict the efficiency and precision of current HRC tasks. To enhance the performance of human–robot collaboration under physical contact conditions, we propose a joint network model named ACT_force_cooperative (AFC). This model leverages force sensing information as a representation of human intent to achieve human intent prediction during physical interaction, while simultaneously capturing visual information and robot state data, thereby enabling more efficient execution of human–robot collaborative tasks with multimodal information processing. Existing HRC methods often ignore humans' collaborative experience in the environment and fail to recognize the uniqueness of interactive force in expressing human intentions. Focusing on the special role of interactive force among various types of data in physical interaction environments, the proposed model predicts humans' future behavioral intentions and adopts a Transformer model to realize the fusion and representation of multimodal information, thus accomplishing more accurate and effective HRC tasks. Experimental results demonstrate that, through the processing of force sensing information and fusion of multimodal data, the proposed model reduces the motion error by 44.9% and increases the effective collaborative time ratio by 20.2% compared with the baseline Action Chunk Transformer (ACT) model. This not only improves the motion accuracy of the robot in collaborative tasks but also enhances the collaborative experience of human operators.

Keywords: human-robot collaboration; interaction force; action chunk transformer

Academic Editor: Brian Byunghyun Kang

Received: 27 January 2026
Revised: 14 February 2026
Accepted: 19 February 2026
Published: 23 February 2026

1. Introduction

HRC plays a pivotal role in manufacturing, healthcare, and service industries [1–3], requiring robots to achieve seamless collaboration by accurately estimating human intentions and dynamically adapting to evolving tasks and environments, thereby reducing the cognitive and physical burdens on human operators. Meanwhile, human collaborative experience typically expresses human intentions through haptic feedback during collaboration, such as changing the collaboration direction and adjusting the collaboration speed. As a key form of HRC, physical human–robot interaction (pHRI) involves direct contact and force transmission between humans and robots, and is widely applied in contact-rich tasks such as precision assembly and collaborative transportation [4–6]. These tasks often demand millimeter-level operational precision, delicate contact force coordination, and

closed-loop visual feedback—for instance, actions like threading cable ties, slotting batteries, and opening condiment cup lids [7]. Even minor deviations can lead to task failure, imposing stringent requirements on the robot's control performance.

Existing fine manipulation and pHRI control methods still face two core challenges: first, traditional high-precision operating systems rely on expensive robotic hardware and high-end sensors [7], resulting in high costs and complex deployment, while the precision limitations of low-cost hardware further exacerbate the difficulties in perception and planning; second, most existing control methods depend on single or dual-modal data such as vision or position [8–10], neglecting the critical role of interaction force as a direct carrier of human intentions. Interaction force contains rich intent information including motion direction and force adjustment, serving as an indispensable communication channel in the physical coupling process between humans and robots [11–13]. Traditional intent estimation methods primarily rely on short-term motion data (e.g., position and velocity) for prediction [14–16], making it difficult to respond to sudden changes in human intentions and prone to accuracy degradation in long-term collaboration [17–19]. Although some studies have attempted to integrate force sensing into intent estimation [20–22], most rely on complex dynamic modeling or computationally expensive diffusion policies [23–25], struggling to balance real-time performance and control precision.

In the field of imitation learning, the ACT model has demonstrated outstanding performance in fine-grained manipulation tasks by predicting action sequences in groups and incorporating temporal ensembling strategies to effectively mitigate error accumulation in imitation learning [7]. Leveraging the sequence modeling capabilities of the Transformer architecture, the ACT model can learn precise closed-loop control strategies from a small number of human demonstrations. Its core advantage lies in reducing the effective time horizon of tasks through action chunking while optimizing action smoothness via temporal ensembling [7]. However, the inputs of existing ACT models only include visual and joint state information [26–28], failing to utilize interaction force data. In contact-rich pHRI tasks, it is difficult for these models to accurately judge contact events (e.g., force feedback during object pressing and placement) and adjust collaboration compliance, limiting their application scope in pHRI scenarios [29–31].

To address the aforementioned issues, this paper proposes a human–robot interaction control method based on interaction force and the ACT model, whose specific structure is illustrated in Figure 1. By incorporating interaction force information into the ACT model framework, this method constructs a multimodal input mechanism that fully leverages the advantages of interaction force in intent expression and contact state perception, while utilizing the sequence prediction capability of the ACT model to enhance control precision and robustness.

The main contributions of this paper are as follows:

1. An extended ACT model framework based on interaction force encoding is proposed. Force signals are integrated as direct representations of human intentions into the action chunking Transformer architecture. A dedicated force signal encoder is designed to extract contact state and intent features, which compensates for the perceptual limitations of traditional ACT models that only rely on vision and joint states, enabling accurate capture of human intentions in contact-rich tasks.
2. A correlation mapping mechanism between interaction force and action sequences is established. Leveraging the advantages of action chunking and temporal ensembling in the ACT model, the intent information represented by force is converted into coherent robotic control actions. This effectively mitigates the error accumulation problem under pure visual drive and improves the accuracy and compliance of action execution in human–robot collaboration.

3. Without relying on complex dynamic modeling or additional intent estimation modules, end-to-end learning is adopted to realize the direct mapping of interaction force-intention-action, simplifying the system design of human–robot interaction control. Meanwhile, it maintains the adaptability of the ACT model to low-cost hardware, providing a more efficient solution for the engineering deployment of high-precision collaborative tasks.

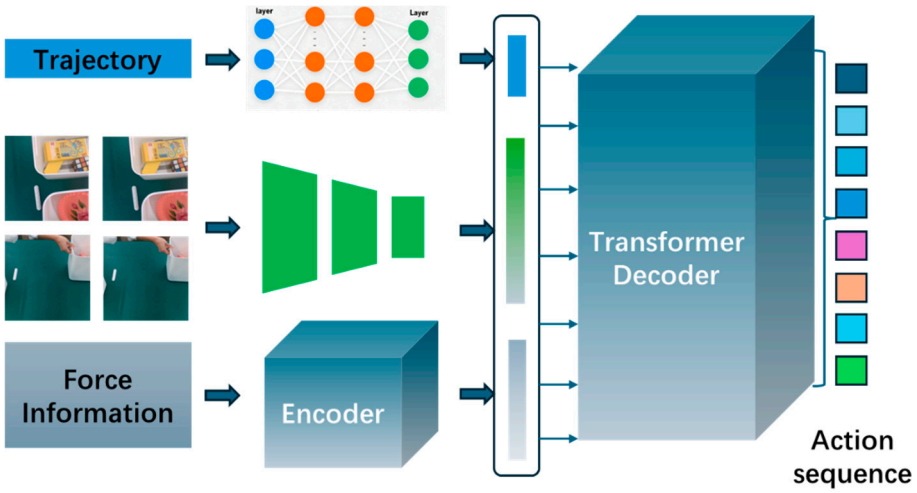

Figure 1. Overview of human–robot interaction control method integrating interaction force and ACT model. The framework consists of three core functional units: (1) interaction force information encoder for extracting human intent features from force/torque signals; (2) Transformer-based multi-modal feature fusion module; and (3) Transformer decoder for generating robot action sequences.

2. Related Work

2.1. Human Intention Perception

Human intention perception serves as the core foundation for seamless pHRI, enabling robots to proactively adjust behaviors and achieve safe and efficient collaboration. Current research on intention perception primarily relies on single-modal or multi-modal data fusion strategies, with significant differences in performance across complex contact-rich scenarios.

Early intention estimation methods mainly relied on short-term motion data such as human joint positions and velocities. For example, Gao et al. [16] proposed a hybrid recurrent neural network architecture to recognize human intentions by learning temporal features from motion sequences, but this method struggled to respond to sudden changes in human intentions and was prone to error accumulation in long-term collaboration. Huang et al. [14] developed an intent detection algorithm for intelligent walkers based on shared navigation control, which achieved basic intention recognition but lacked adaptability to dynamic environments. With the development of multi-modal perception technology, integrating multiple sensing modalities has become a mainstream trend to improve estimation accuracy. Tsunekawa et al. [32] proposed a Transformer-based dynamic attention analysis framework, revealing the patterns of attention transitions across body parts through kinematic optimization and attention visualization. Liu et al. [33] proposed the IDAGC framework, which achieves accurate human intent estimation via a conditional variational autoencoder (CVAE), integrates visual, linguistic, force, and robot state data through modality-specific encoders and a Transformer decoder, supports seamless switching between physical human–robot interaction (pHRI) and remote cooperation as well as multi-task policy learning, and provides a general paradigm for adaptive human–robot collaboration. Wong et al. [34] proposed a continuous multi-modal intention recognition

method combining vision and tactile sensing, which classified intentional and unintentional contacts by analyzing human posture, gaze direction, and touch location, achieving an F1-score of 86%. Similarly, Franceschi et al. [10] designed an assistive control system for pHRI that learned human motion intentions through visual and joint state fusion, but it failed to consider the direct reflection of interaction force on intent changes.

In recent years, Transformer-based models have shown significant advantages in capturing long-term dependencies in sequence data, promoting the development of intention perception towards long-horizon prediction. Liu et al. [18] proposed a dual Transformer-based Robot Trajectron (DTRT) framework, which integrated human-guided motion and force data through two Transformer-based conditional variational autoencoders (CVAEs), enabling rapid capture of human intention changes and accurate trajectory prediction. This method incorporated human dynamics into long-term prediction, addressing the limitation of short-term data-based methods in handling intent mutations. Additionally, gaze-based intention estimation has attracted attention as a non-contact sensing method. Fuchs et al. [35] developed Gaussian Hidden Markov Models (GHMMs) for pick-and-place tasks, leveraging eye movement scan paths to achieve early intention estimation, which generalized well across different users and spatial configurations but relied heavily on gaze tracking equipment and was less applicable in dynamic contact scenarios.

Despite the progress made, existing intention perception methods still face two key challenges: first, single-modal perception (e.g., vision or motion) struggles to fully capture intent information in contact-rich tasks, where human intentions are often directly reflected through interaction forces; second, although some multi-modal methods integrate force data, they lack effective modeling of the temporal correlation between force signals and intent changes, leading to insufficient responsiveness to sudden intent adjustments during physical contact.

2.2. Force Processing in pHRI

Force processing technology is crucial for pHRI, as it directly affects the safety, compliance, and precision of robot operations. This research direction mainly focuses on force sensing, signal processing, and force-based control strategy design, aiming to address issues such as contact force regulation and safety assurance in physical interaction.

In terms of force sensing systems, high-precision and multi-dimensional force acquisition has become a key research focus. D'Antona et al. [36] developed a variable stiffness testing device (VSITD) for collaborative robotics, which integrated FSR matrices and piezoelectric load cells to achieve high-resolution contact pressure mapping and point-specific force measurement. The system used an FPGA-based hardware platform for high-speed parallel acquisition of force and pressure data, ensuring compliance with ISO/TS 15066 safety standards [37]. Similarly, Liu et al. [20] designed a force-motion capture system for contact-rich manipulation tasks, which acquired high-fidelity force and motion data to support force-centric imitation learning. These sensing systems provide reliable data support for subsequent force processing, but the integration complexity and cost need to be further optimized for low-cost hardware deployment.

In force-based control strategies, impedance control and admittance control are widely used to adjust robot compliance. Wang et al. [38] proposed a Mamba-2-enhanced Transformer framework, integrating the Phase-Guided Action Chunk (PGAC) to fuse force-motion data and enabling human-inspired compliant manipulation in multi-stiffness assembly scenarios. Xing et al. [1] proposed an impedance learning method for human-guided robots, enabling robots to adapt to unknown environments through interaction force feedback and improve collaboration adaptability. Yu et al. [3] developed an adaptive-constrained impedance control strategy for human–robot co-transportation, which ad-

justed control parameters based on real-time force signals to maintain stable collaboration. With the development of learning-based methods, integrating force information into data-driven control frameworks has become a new trend. Ma et al. [39] proposed a Compliance-Aware Tactile Control and Hybrid Deformation Regulation-based Action Transformer (CATCH-FORM-ACTer), which dynamically adjusted stiffness and damping parameters through real-time force feedback during viscoelastic object manipulation, achieving sub-millimeter deformation accuracy. Watanabe et al. [22] extended the ACT model by incorporating force-torque information, improving the performance of pick-and-reorient tasks, but this method focused more on trajectory optimization rather than establishing a direct mapping between force signals and human intentions. Fusco et al. [40] proposed a hybrid architecture consisting of an MLP Transducer and a Transformer, which estimates and predicts contact forces and human motions in physical human–robot interaction (pHRI) based on human kinematic data, validating the advantages of Transformers in the processing of temporal force data.

Current force processing methods have made significant progress in safety assurance and precision control, but there are still limitations in pHRI scenarios: on the one hand, traditional impedance control relies on complex dynamic modeling and parameter tuning, lacking adaptability to individual differences in human operation; on the other hand, existing learning-based methods either ignore the intent expression function of force signals or require complex feature engineering, making it difficult to achieve end-to-end intent-force-action mapping. Therefore, developing a force processing method that can directly characterize human intentions and adapt to low-cost hardware remains an important research gap.

Table 1 systematically summarizes and compares representative existing models in the field of pHRI related to human intention perception and force signal processing, focusing on four key technical dimensions. Specifically, these dimensions include: force data utilization, temporal sequence processing, multimodal fusion, and low-cost hardware compatibility.

Table 1. Systematic comparison of core human–robot interaction models.

Research Team	Force Data	Temporal Sequence Processing	Multimodal Fusion	Low-Cost Hardware Compatible
Hybrid RNN (Gao et al. [16])	✗	✔	✗	✔
Tsunekawa et al. [32]	✗	✔	✗	✔
IDAGC (Liu et al. [33])	✔	✔	✔	✗
Wong et al. [34]	✗	✗	✔	✗
Franceschi et al. [10]	✗	✗	✔	✔
DTRT (Liu et al. [18])	✔	✔	✔	✗
Gaze-Based GHMM (Montesano et al. [35])	✗	✗	✗	✗
ForceMimic (Liu et al. [20])	✔	✗	✔	✗
Wang et al. [37]	✔	✔	✔	✗
FTACT (Watanabe et al. [22])	✔	✔	✔	✔
Fusco et al. [39]	✔	✔	✗	✗

Existing core models for human–robot interaction still have significant limitations across key technical dimensions: some models fail to integrate force sensing data, making it impossible to capture human intentions expressed through haptics in contact-rich scenarios and resulting in insufficient intent recognition accuracy; methods lacking effective temporal sequence processing mechanisms struggle to respond to sudden human intent changes and are prone to error accumulation in long-term collaboration; multimodal fusion models either suffer from insufficient fusion depth or rely on complex modal encoding modules, leading to high deployment costs and poor adaptability; certain approaches are highly dependent on high-precision sensing equipment (e.g., gaze tracking devices) or high-fidelity data, resulting in poor generalization and adaptability to low-cost hardware; furthermore,

most models integrating force data fail to establish direct force-intent-action mapping, either relying on complex dynamic modeling or focusing solely on trajectory optimization, making it difficult to balance control accuracy and collaboration compliance and limiting their engineering applications in practical physical human–robot interaction scenarios.

3. Methodology

This section introduces the AFC framework, which integrates interaction force signals, visual data, and robot state information with human intention characterization to enable end-to-end adaptive policy learning. By autonomously extracting intent-related features from force feedback and fusing multimodal perceptual information, the framework dynamically optimizes action sequences and adjusts collaboration compliance, thereby achieving precise and smooth physical human–robot interaction.

3.1. Problem Statement and Framework Overview

In pHRI scenarios involving direct contact (e.g., precision assembly, collaborative transportation), the core challenge lies in enabling robots to accurately perceive human intentions in real time and generate compliant and precise control actions. Specifically, human intentions in contact-rich tasks are often implicitly characterized by interaction force signals (e.g., force magnitude variations indicating movement direction preferences, contact force peaks reflecting task stage transitions). However, existing Transformer-based action learning methods typically rely solely on visual and robot joint state data, failing to effectively exploit the intent expression value of interaction force, leading to inadequate responsiveness to sudden intent changes and accumulated errors in long-term collaboration.

To address the aforementioned problem, this paper proposes the AFC framework. The framework is designed to fully exploit the intent characterization advantage of interaction force and leverage the sequence modeling capability of Transformer for action chunking, enabling end-to-end adaptive policy learning for pHRI tasks. The AFC framework consists of four core modules: (1) Multimodal Data Acquisition Module, which collects real-time interaction force signals, visual data, and robot joint state data; (2) Force-Intention Encoding Module, which extracts intent-related features from raw force signals through dedicated encoding networks, realizing the transformation from force signals to human intention representations; (3) integrated Transformer learning module, which fuses the encoded force-intention features with visual and joint state features, and performs action chunk prediction based on the Transformer architecture; (4) Action Optimization and Execution Module, which optimizes the predicted action chunks through temporal ensembling strategies, decodes single-step actions, and outputs them to the low-level controller to drive the robot to complete collaborative tasks.

The key design philosophy of the framework is to take interaction force as the direct characterization of human intention, avoiding complex explicit intent estimation modules. By integrating force-intention features into the Transformer-based action chunking process, the framework achieves tight coupling between human intention perception and robot action generation, thereby improving the accuracy and compliance of physical human–robot collaboration while ensuring real-time performance and adaptability to low-cost hardware.

3.2. Force Sensing and Feature Encoding Module

The force sensing and feature encoding module is a core prerequisite for the AFC framework to achieve human intention perception. Specifically, it processes real-time interaction force data collected by sensors, implements feature encoding for historical force sequences, predicts the impending force signal, and outputs the fused force feature vector

to serve as one of the input features of the subsequent integrated Transformer learning module. The overall workflow of this module is shown in Figure 2.

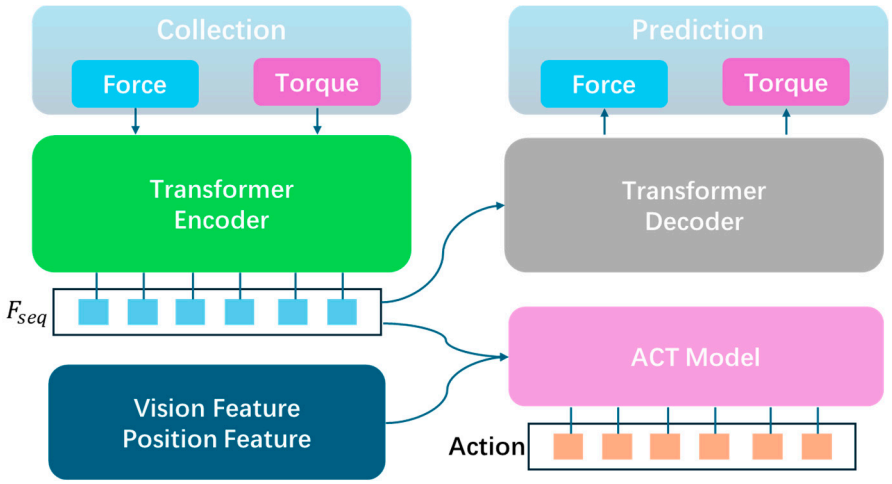

Figure 2. The overall workflow of force sensing and feature encoding module.

To accurately capture the human–robot interaction forces in contact-intensive tasks, the sensor sampling frequency is set to 1000 Hz, which can fully capture the dynamic changes in force signals during fine operations. The collected raw force signals are represented as a six-dimensional vector whose specific form is shown in Equation (1):

$$\mathcal{F}_t = \left[F_{x,t}, F_{y,t}, F_{z,t}, T_{x,t}, T_{y,t}, T_{z,t}\right]^T \in R^6 \tag{1}$$

Considering that raw force signals may contain high-frequency noise, preprocessing operations are required to ensure the quality of subsequent feature encoding. The preprocessing procedure mainly consists of two steps: (1) A 2nd-order Butterworth low-pass filter with a cutoff frequency of 10 Hz is employed to eliminate high-frequency noise in the raw force signals while preserving their effective dynamic characteristics. During model training, random noise with an amplitude of less than 3% of the maximum value is injected to enhance the model's anti-interference capability; (2) Normalization: the min-max normalization method is employed to normalize the denoised force signals to the range [0, 1], which eliminates the dimensional discrepancy between different components and avoids the impact of excessive signal amplitude on the training stability of the encoding network. The normalization formula is given in Equation (2):

$$\mathcal{F}_{t,\mathrm{nr}}^{i} = \frac{\mathcal{F}_t^i - \min\left(\mathcal{F}^i\right)}{\max\left(\mathcal{F}^i\right) - \min\left(\mathcal{F}^i\right)} \tag{2}$$

Force feature extraction is implemented using a lightweight Transformer encoder (MiniForceTransformer). This module takes 50 high-frequency sampled force points (six dimensions per point) at the current moment as input and extracts 256-dimensional global force features through a single-layer Transformer encoder. The specific processing flow is as follows: first, the input force data is projected into a 128-dimensional hidden space; then, feature extraction is performed via positional encoding and a single-layer Transformer encoder; finally, a 256-dimensional global force feature vector is obtained through the CLS token aggregation mechanism and an output projection layer. The force features are converted to 512 dimensions through a linear projection layer, unifying them with visual features and state features into the same dimensional space. Subsequently, all features are concatenated and fused in the form of a token sequence at the input layer of the Transformer encoder. This fusion mechanism enables the three modalities (visual, state,

and force) to learn cross-modal semantic correlations through the self-attention mechanism at the Transformer encoder layer.

The force prediction branch is implemented using an autoregressive predictor. This module takes the 256-dimensional force features of the current frame as input, and parallelly predicts the 256-dimensional force features of the next 50 frames through a Transformer decoder. Unlike the action prediction branch, the force prediction branch shares the force feature extractor with the action prediction branch but adopts an independent decoder for prediction, achieving a balance between feature sharing and task separation.

3.3. AFC Joint Training Module

The AFC Joint Training Module serves as the core computational unit of the framework. The overall architecture of this module is illustrated in Figure 3, which mainly consists of three key sub-modules: multi-modal feature input and embedding, Transformer-based feature fusion and position prediction, and an end-to-end joint training mechanism.

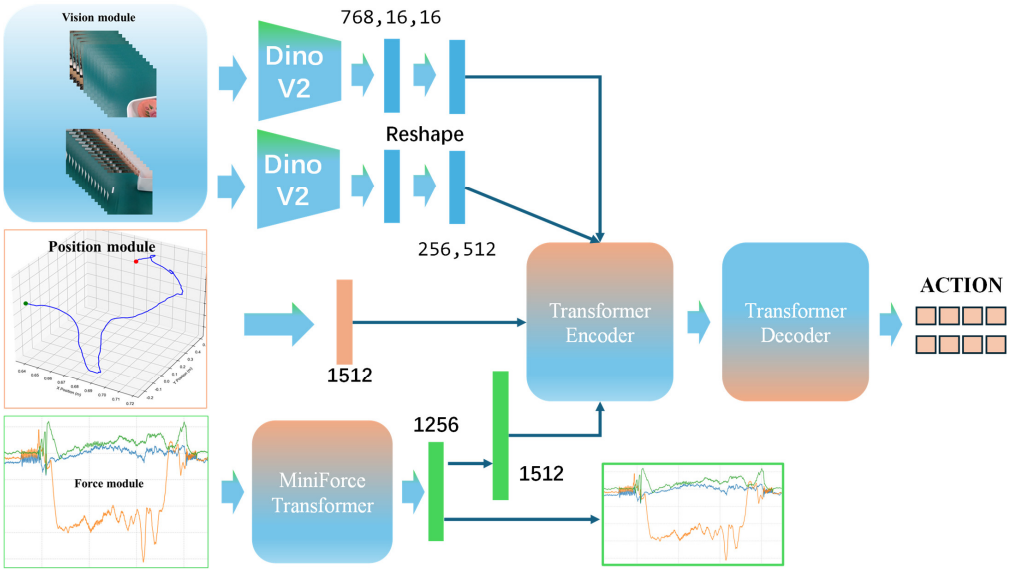

Figure 3. The overall architecture of AFC Joint Training Module. The module includes three modality-specific feature extraction submodules: (1) Force module: MiniForceTransformer for extracting 50-time-step 6D force features and projecting to 512-dimension; (2) Vision module: DinoV2 model for binocular image feature extraction and reshaping to 512-dimension; (3) Position module: robot end-effector state (9-dimension) encoding to 512-dimension.

The AFC method adopts a cooperative multimodal fusion mechanism to achieve deep integration of force information, visual information, and state information. The model inputs include binocular visual images, robot end-effector states (nine dimensions), and temporal force data. The force data are structured as a force sequence with 50 time steps, where each time step contains 6-dimensional force measurements, with a shape of (B, 50, 6). The DinoV2 [41] model is utilized for image feature extraction. In this paper, visual, force and state features are all projected to 512 dimensions: visual features are mapped from 768 dimensions via 1×1 convolution, while force and state features are mapped from 256 and nine dimensions respectively through linear projection, thus unifying all tokens in the same interoperable feature space. Latent variables, force and state features, as one-dimensional (1D) features, are each converted into a single token, and the two-dimensional (2D) features of binocular vision are reshaped into 512 tokens, which are finally concatenated to form a unified input sequence of 515 tokens. A separated positional encoding strategy is adopted: learnable positional encoding is assigned to 1D tokens including force and state tokens, and independent 2D positional encoding is configured for

visual 2D tokens. The encoding results are added to the corresponding tokens, enabling the model to accurately distinguish the semantic information of tokens across different modalities and spatial positions. The concatenated and encoded token sequence is fed into a 4-layer Transformer encoder, which accomplishes token weighting through the multi-head self-attention mechanism.

The total loss function of the jointly trained model is given by Equation (3):

$$L_{total} = L_{ACT} + \alpha \cdot L_{force_{weighted}} \tag{3}$$

$L_{force\_weighted}$ denotes the temporally dynamic weighted force prediction loss, whose specific form is shown in Equation (4):

$$L_{force\_weighted} = Smooth\_L1(f_{pred}, f_{get}) \odot w_{temporal} \tag{4}$$

The temporal weight vector $w_{temporal}$ is defined as follows: the weight for the first 20 frames (short-term prediction) is 1.4, and the weight for the subsequent 30 frames (long-term prediction) is 0.9. This design is based on the consideration that the accuracy of short-term prediction is more critical in practical applications, while accounting for the uncertainty of long-term prediction. The force loss weight $\alpha = 0.3$ is set to emphasize the effectiveness of collaborative learning.

4. Experiments

This paper verifies the effectiveness of the proposed method through a human–robot collaborative sorting and cooperative transportation task. The specific experimental scenario is illustrated in Figure 4. Two Rizon4s robots are employed as the platform for collaborative robots, two Intel RealSense SR300 cameras are adopted as the visual sensing devices with a sampling frequency of 15 Hz, and the sampling frequency of the force sensor is 1000 Hz.

Figure 4. Human–robot collaborative sorting and transportation scenario. Experimental platform for pHRI validation, consisting of two Rizon4s collaborative robots, two 15 Hz Intel RealSense SR300 cameras (vision), and 1000 Hz 6-axis force sensors.

To collect realistic interaction force data for human–robot collaboration, we employed a teleoperated arm to conduct operation experiments, as illustrated in Figure 5. Specifically, an operator manipulates the master manipulator to control another manipulator, which collaborates with a human operator to complete the transportation task. We conducted experiments with four participants (two males and one female), where all experimental tasks were identical, yet each participant exhibited distinct movement trajectories. What the robot needs to learn is the perception of haptic force variations, rather than mechanical adherence to a fixed path. Thus, individual differences in human force application patterns serve to enhance the diversity of the database. We collected 100 high-quality collaboration cases, and the main data captured include dual-channel camera information, time-aligned force sensing information, and robot end-effector position information, with the data from each participant split into training and test sets at an 8:2 ratio. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee for Science and Technology Research Involving Human Subjects of Shanghai Jiao Tong University (project identification code 20260054I) on 9 February 2026.

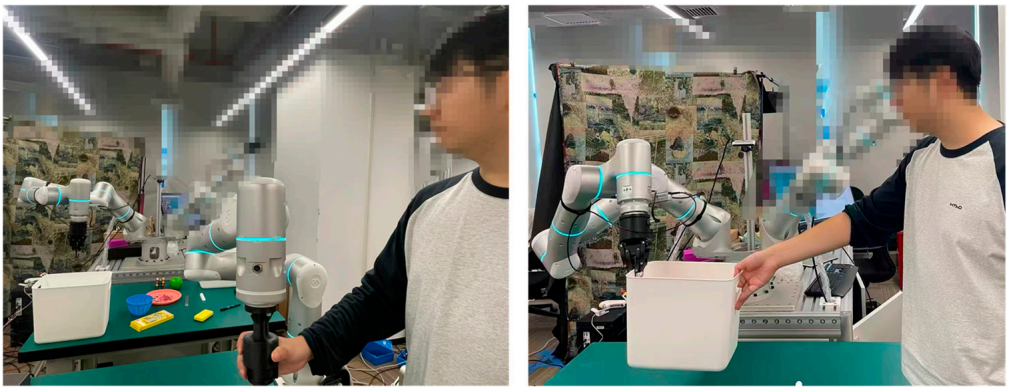

Figure 5. Teleoperation-based data collection scenario for.

Human–Robot Collaborative Tasks

We trained three models separately: the original ACT model, the ACT_FORCE model, and the AFC model. Specifically, the ACT_FORCE model extends the original ACT model by incorporating a dedicated force channel, while the AFC model implements joint training of force and action based on the algorithm proposed in this paper.

The training configurations are detailed as follows: the backbone network was fine-tuned with LoRA at a learning rate of 1×10^{-5}; the main ACT module adopted a learning rate of 1×10^{-4}. The AdamW optimizer was employed with a weight decay of 1×10^{-4}, a gradient clipping threshold of 10.0, and betas set to [0.9, 0.999]. The batch size was configured as 32, with a maximum of 50,000 training steps. Data augmentation techniques include color jittering (brightness [0.8, 1.2], contrast [0.8, 1.2], saturation [0.5, 1.5], hue $[-0.05, 0.05]$) and sharpness adjustment ([0.5, 1.5]). A maximum of three transformations are randomly applied to each sample. The model predicts a 50-step action sequence, executes the first 10 steps, then re-generates a new sequence based on the updated observation state; this cycle is repeated until a cumulative total of 150 steps are executed. The experiment was conducted on two NVIDIA RTX 4090 GPUs, using the Python + PyTorch (Python 3.10 with PyTorch 2.9.1) deep learning framework. The training time of the ACT model was 7.2 h, and that of the ACT_force model was 7.5 h. Owing to the additional branch network, the AFC model took 12.8 h to train. The final training results are presented in Figure 6.

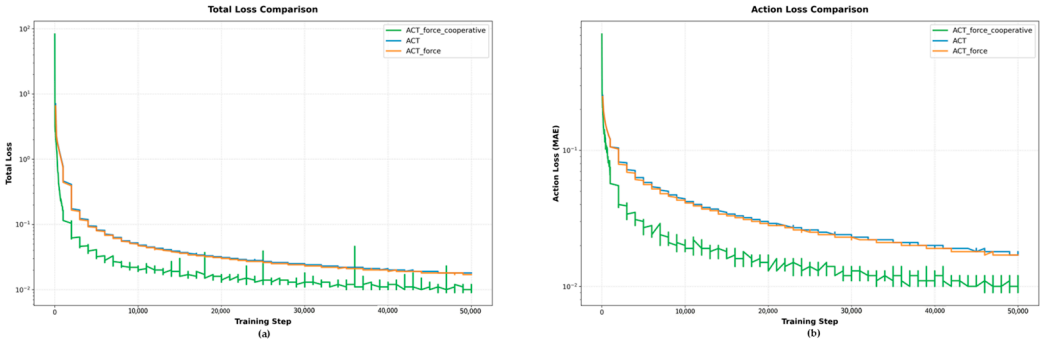

Figure 6. Training curve comparison for different methods. (**a**) Total loss variation with training steps; (**b**) action loss variation with training steps.

The AFC model exhibits a relatively high total loss at the initial training stage (approximately 83.0), primarily attributed to its more complex architecture that integrates multimodal fusion and force prediction branches, requiring the learning of more parameters. Despite the high initial loss, the model demonstrates rapid convergence capability, with the loss value dropping sharply in the early training phase (the first 1000 steps). At the later training stage (50,000 steps), the total loss reaches the lowest level, indicating that the cooperative fusion mechanism effectively enhances the overall learning capability of the model. The final losses of the two models incorporating force sensing information (ACT_force and AFC) are both lower than that of the baseline ACT model, which validates the effectiveness of force sensing in robot imitation learning. The AFC model achieves the lowest final loss, demonstrating that the cooperative multimodal fusion mechanism outperforms the independent branch design in terms of performance. The AFC model yields the lowest final total loss, with an average trajectory prediction error of 0.0103, which is significantly lower than those of the ACT (0.0187) and ACT_force (0.0181) models, representing an improvement of approximately 45–48%. The lowest action loss of the AFC model further verifies the advantages of the cooperative fusion mechanism in the action prediction task.

We conducted future end-effector position prediction for the robot using the three models, and the prediction results are presented in Figure 7. In this study, the computational overhead of the AFC model and the baseline ACT model during the inference phase was tested on the NVIDIA RTX 4090 GPU platform. Both models adopted a cyclic inference logic of predicting 50 steps and executing 10 steps; however, significant differences in computational overhead were observed due to architectural disparities between the models. The single-chunk inference latency of the ACT model was 0.784 ± 1.739 ms, that of the ACT_force model was 0.799 ± 1.794 ms, and the single-chunk inference latency of the AFC model was 0.872 ± 1.809 ms. The inference latencies of all three models can meet the real-time requirements for robotic action execution.

Average instantaneous error stands for the mean value of the robot end-effector position error at each time step, which reflects the short-term motion accuracy. The average instantaneous error of the AFC model is 0.0103, which is reduced by 44.9% and 43.1% respectively compared with the baseline ACT model (0.0187) and the ACT_force model (0.0181). Error stability analysis reveals that the error standard deviation of the AFC model is 0.0003, representing a 50.0% reduction compared with the baseline model, and its error fluctuation range is 0.0011, which is only half of that of the other two models (0.0022). Cumulative error analysis indicates that the cumulative errors of the three models all exhibit a linear growth trend, with slopes of 0.0103, 0.0181, and 0.0187 respectively. The AFC model achieves the lowest cumulative error growth rate, which reflects the long-term error accumulation speed, demonstrating its optimal error suppression capability in long-term prediction.

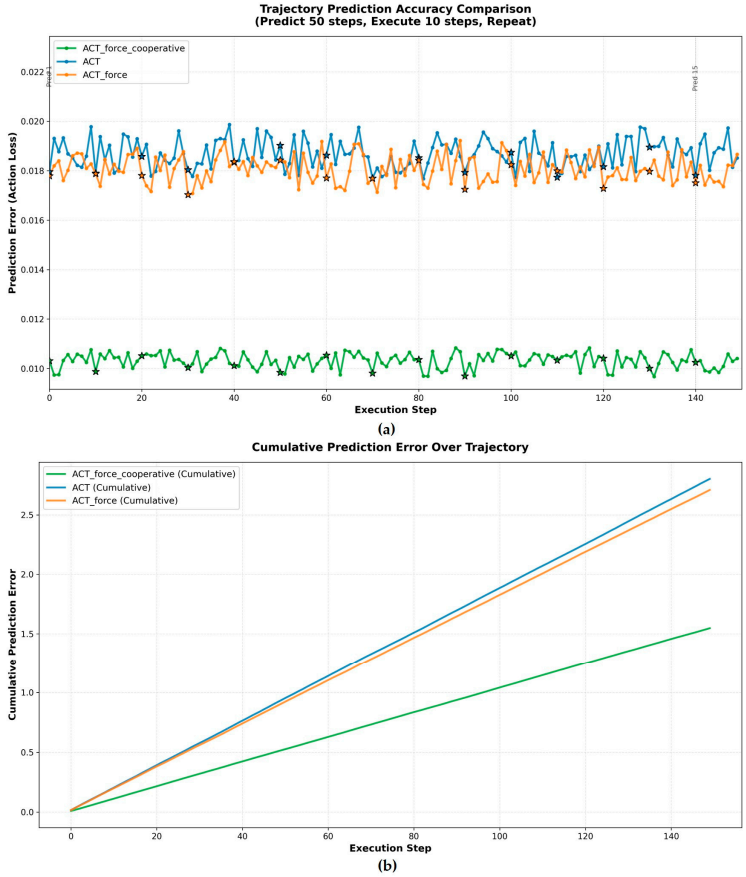

Figure 7. Trajectory prediction accuracy comparison for different methods (predict 50 steps, execute 10 steps, repeat). (**a**) Prediction error (action loss) variation with execution steps. The asterisks represent the prediction results every 10 steps; (**b**) cumulative prediction error over trajectory execution steps.

Comprehensive evaluation results show that the cooperative multimodal fusion mechanism can simultaneously improve the short-term accuracy and long-term stability of the model.

We further conducted 72 experiments on the manipulator for each of the three models separately. We proposed an effective collaborative time ratio to measure the compliance level of the robot during the collaboration process, which is defined as the proportion of time when the angle between the motion direction and the interaction force direction is less than 90 degrees relative to the total collaborative time. The results are presented in Figure 8.

The evaluation results demonstrate that the average effective collaborative time ratio of the AFC model reaches 75.43% (with a standard deviation of $\pm$4.87%), which is significantly higher than that of the ACT_force model (64.32%, $\pm$8.11%) and the baseline ACT model (62.76%, $\pm$13.30%).

Specifically, compared with the baseline model, the effective collaborative time ratio of the AFC model increases by 20.2%, its standard deviation decreases by 63.4%, and the coefficient of variation drops from 21.2% to 6.5%. Per-experiment analysis of the effective collaborative time ratio shows that the ratio of the AFC model mainly ranges from 65% to 87% across 72 experiments, exhibiting the narrowest fluctuation range; moreover, its minimum effective collaborative time ratio remains above 65%, which verifies the model's favorable performance stability and generalization capability for different participants under the same task. In contrast, the baseline ACT model has a fluctuation range of up to

60% (from 30% to 90%) in the effective collaborative time ratio, with a standard deviation of ±13.30% and a coefficient of variation of 21.2%, indicating unstable performance.

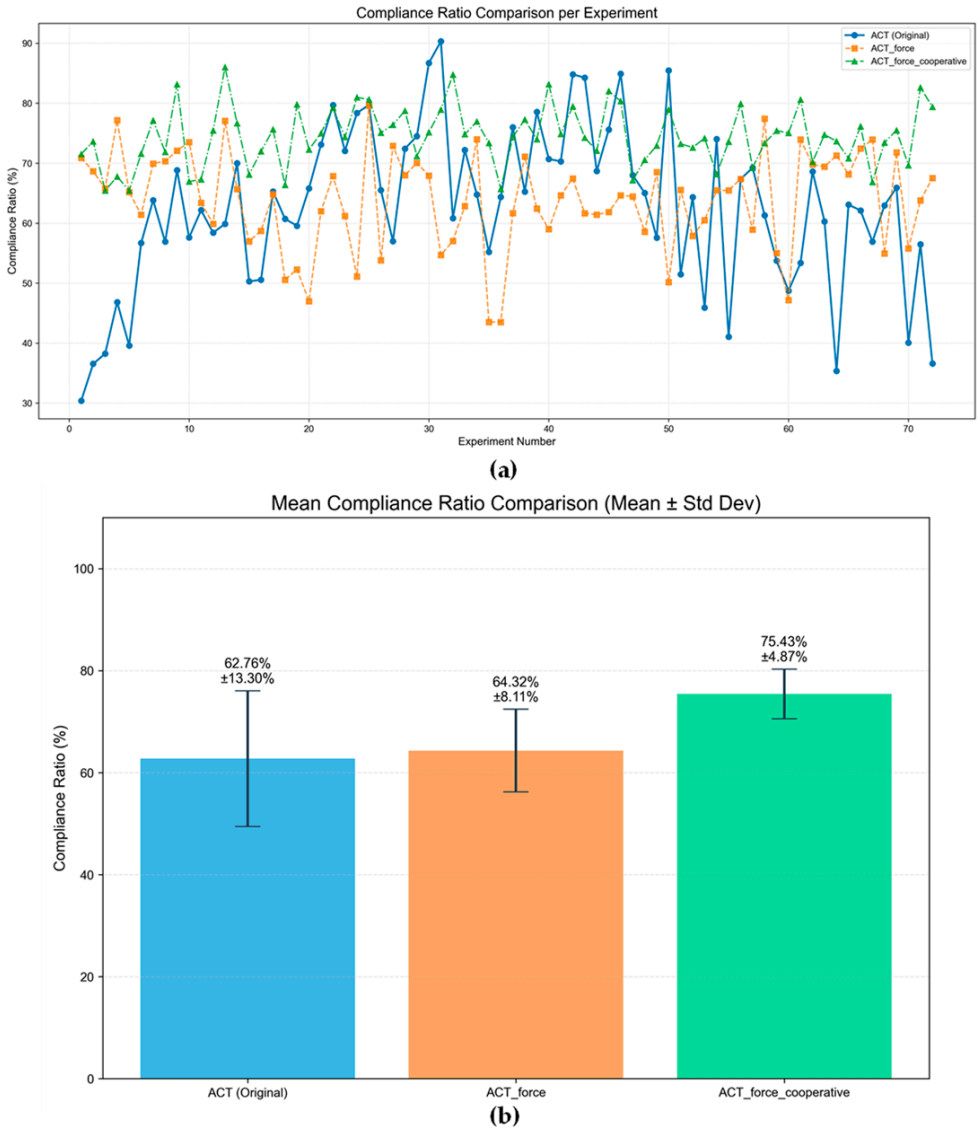

Figure 8. Compliance ratio comparison of different methods: per-experiment variation and mean values. (**a**) Compliance ratio variation across individual experiment numbers; (**b**) mean compliance ratio comparison with standard deviation.

These results are consistent with the findings from the training loss and prediction error analyses, further validating the effectiveness of the cooperative multimodal fusion mechanism in practical applications. Compared with force-aware imitation learning methods such as ForceMimic [20] and FTACT [22], the AFC model abandons the design concepts of complex dynamic modeling or exclusive focus on trajectory optimization. Instead, it integrates interactive force directly as the core representation of human intention into the action chunking Transformer architecture of the ACT model, realizing an end-to-end direct mapping of force-intention-action. It also retains good compatibility with low-cost hardware, effectively addressing the technical challenge of balancing real-time performance and control precision in traditional force-aware learning methods. Aiming at the limitations of shallow multimodal fusion and over-reliance on high-precision sensing equipment in multimodal Transformer models such as IDAGC [33] and DTRT [18], the AFC model is equipped with a dedicated force signal encoder to accurately extract intention features and

contact state features from interactive force. Combined with a modality-specific positional encoding strategy and a 4-layer Transformer encoder, it achieves deep collaborative fusion of interactive force, visual information and robot state information. In addition, a temporally dynamically weighted joint training mechanism is introduced to enhance the precision of short-time scale force prediction, thereby improving the collaboration stability in the process of physical human–robot interaction. As for independent branch architectures such as ACT_force that simply add a force channel, they extract features of each modality independently and lack cross-modal semantic interaction and feature correlation, which makes it impossible to capture the intrinsic coupling relationships between changes in interactive force, visual contact states and robot motion states, and thus difficult to realize the complementary utilization of multimodal information. In contrast, the AFC model projects all modal features into a unified 512-dimensional feature space and relies on the multi-head self-attention mechanism of the Transformer to complete the mutual weighting and semantic correlation learning of cross-modal features. Combined with the joint training strategy of force prediction and action prediction, the model can accurately perceive human interaction intentions based on complementary multimodal information. This avoids the problem of error accumulation caused by insufficient single-modal information from the perspective of model principles and greatly improves the perceptual robustness and action execution precision of the model in physical human–robot interaction tasks.

5. Conclusions

HRC plays a crucial role in smart manufacturing and autonomous robotics. However, traditional vision-state fusion methods often struggle to accurately perceive environmental states and predict long-term action sequences when confronting complex interactive tasks, limiting the reliability and accuracy of models in practical applications. To improve the performance of robot imitation learning in complex interactive tasks, it is necessary to fully leverage multimodal perceptual information, especially force sensing information, to achieve more accurate action prediction and more stable long-term execution.

This study proposes a cooperative multimodal fusion mechanism (AFC), which realizes the synergistic optimization of multimodal information by deeply fusing force features with visual and state features at the Transformer encoder layer. Experimental results demonstrate that the cooperative fusion mechanism can significantly improve model performance: the average instantaneous error is reduced by 44.9%, the error standard deviation is decreased by 50.0%, the cumulative error growth rate is lowered by 44.9%, and the effective collaborative time ratio is increased by 20.2%. These results validate the effectiveness of force sensing in robot imitation learning and the advantages of the cooperative fusion mechanism.

However, this study also has several limitations: (1) the method relies on high-quality force data acquisition, where the synchronization and annotation costs of multimodal data are high, and the calibration and maintenance of force sensors require professional expertise; (2) the computational complexity of the model during real-time inference is relatively high, as multimodal feature fusion and force prediction branches increase computational overhead, potentially affecting response speed in real-time applications; (3) the generalization capability to different robot platforms and task scenarios is limited, requiring data acquisition and model fine-tuning for specific platforms and tasks; (4) the quality and noise of force data significantly impact model performance, necessitating robust sensor fusion and noise suppression mechanisms in practical deployment; and (5) no dedicated subjective experience scale for participants was designed to quantitatively verify the correlation between the effective collaborative time ratio and human operational subjective experience. Due to the lack of quantified subjective experience data, it is difficult to conduct

a comprehensive and systematic evaluation of the model's actual performance at the level of human–robot collaboration experience.

Future research can be conducted from the following aspects. First, construct a "hybrid data augmentation framework" that integrates real-world data with simulation data, generating synthetic force data through physics simulation engines, and combining transfer learning and domain adaptation techniques to compensate for the high cost of real data acquisition while improving model adaptability to different sensor configurations. Additionally, explore a general multimodal imitation learning model based on "few-shot/zero-shot learning": by learning common features of different robot platforms and task scenarios during the pre-training stage, the model can quickly adapt to new tasks with only minimal task descriptions (such as sensor configuration and task type) as input, assisting the model in reasoning and decision-making in unknown scenarios and significantly improving generalization capability. Finally, further research can be conducted on adaptive force loss weight adjustment strategies, dynamically adjusting the weight ratio of multimodal losses according to task difficulty and training stage to achieve optimal performance balance.

Author Contributions: Conceptualization, W.W.; methodology, Z.P.; software, Z.P.; validation, Z.P.; formal analysis, Z.P.; investigation, Z.P.; resources, Z.P.; data curation, Z.P.; writing—original draft preparation, Z.P.; writing—review and editing, W.W.; visualization, Z.P.; supervision, W.W.; project administration, W.W.; funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets presented in this article are not immediately publicly available. Access to the data can be granted upon reasonable request, which should be directed to the author at pan13855965084@sjtu.edu.cn.

Conflicts of Interest: All the authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Xing, X.; Burdet, E.; Si, W.; Yang, C.; Li, Y. Impedance Learning for Human-Guided Robots in Contact with Unknown Environments. *IEEE Trans. Robot.* **2023**, *39*, 3705–3721. [CrossRef]
2. Liu, H.; Tong, Y.; Zhang, Z. Human Observation-Inspired Universal Image Acquisition Paradigm Integrating Multi-Objective Motion Planning and Control for Robotics. *IEEE/CAA J. Autom. Sinica* **2024**, *11*, 2463–2475. [CrossRef]
3. Yu, X.; Li, B.; He, W.; Feng, Y.; Cheng, L.; Silvestre, C. Adaptive-Constrained Impedance Control for Human–Robot Co-Transportation. *IEEE Trans. Cybern.* **2022**, *52*, 13237–13249. [CrossRef] [PubMed]
4. Pandya, R.; Wang, Z.; Nakahira, Y.; Liu, C. Towards Proactive Safe Human-Robot Collaborations via Data-Efficient Conditional Behavior Prediction. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13 May 2024; pp. 12956–12963.
5. Meng, L.; Yang, L.; Zheng, E. Hierarchical Human Motion Intention Prediction for Increasing Efficacy of Human-Robot Collaboration. *IEEE Robot. Autom. Lett.* **2024**, *9*, 7637–7644. [CrossRef]
6. Liu, Y.; Leib, R.; Franklin, D.W. Follow the Force: Haptic Communication Enhances Coordination in Physical Human-Robot Interaction When Humans Are Followers. *IEEE Robot. Autom. Lett.* **2023**, *8*, 6459–6466. [CrossRef]
7. Zhao, T.Z.; Kumar, V.; Levine, S.; Finn, C. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. *arXiv* **2023**. [CrossRef]
8. Cremer, S.; Das, S.K.; Wijayasinghe, I.B.; Popa, D.O.; Lewis, F.L. Model-Free Online Neuroadaptive Controller with Intent Estimation for Physical Human–Robot Interaction. *IEEE Trans. Robot.* **2020**, *36*, 240–253. [CrossRef]
9. Ma, M.; Cheng, L. A Human–Robot Collaboration Controller Utilizing Confidence for Disagreement Adjustment. *IEEE Trans. Robot.* **2024**, *40*, 2081–2097. [CrossRef]
10. Franceschi, P.; Bertini, F.; Braghin, F.; Roveda, L.; Pedrocchi, N.; Beschi, M. Learning Human Motion Intention for pHRI Assistive Control. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1 October 2023; pp. 7870–7877.

11. Shao, Y.S.; Li, T.; Keyvanian, S.; Chaudhari, P.; Kumar, V.; Figueroa, N. Constraint-Aware Intent Estimation for Dynamic Human-Robot Object Co-Manipulation. *arXiv* **2024**. [CrossRef]
12. Liu, H.; Tong, Y.; Zhang, Z. Human-Inspired Adaptive Optimal Control Framework for Robot-Environment Interaction. *IEEE Trans. Syst. Man Cybern. Syst.* **2025**, *55*, 6085–6098. [CrossRef]
13. Wang, C.; Zhao, J. Role Dynamic Assignment of Human–Robot Collaboration Based on Target Prediction and Fuzzy Inference. *IEEE Trans. Ind. Inf.* **2024**, *20*, 471–481. [CrossRef]
14. Huang, C.; Wasson, G.S.; Alwan, M.; Sheth, P.; Ledoux, A. Shared Navigational Control and User Intent Detection in an Intelligent Walker. In Proceedings of the AAAI Fall Symposium: Caring Machines, Arlington, VA, USA, 4–6November 2005; pp. 59–66.
15. Radmand, A.; Scheme, E.; Englehart, K. A Characterization of the Effect of Limb Position on EMG Features to Guide the Development of Effective Prosthetic Control Schemes. In Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; pp. 662–667.
16. Gao, X.; Yan, L.; Wang, G.; Gerada, C. Hybrid Recurrent Neural Network Architecture-Based Intention Recognition for Human–Robot Collaboration. *IEEE Trans. Cybern.* **2023**, *53*, 1578–1586. [CrossRef] [PubMed]
17. Song, P.; Li, P.; Aertbeliën, E.; Detry, R. Robot Trajectron: Trajectory Prediction-Based Shared Control for Robot Manipulation. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13 May 2024; pp. 5585–5591.
18. Liu, H.; Tong, Y.; Zhang, Z. DTRT: Enhancing Human Intent Estimation and Role Allocation for Physical Human-Robot Collaboration. *arXiv* **2025**. [CrossRef]
19. Ke, L.; Wang, J.; Bhattacharjee, T.; Boots, B.; Srinivasa, S. Grasping with Chopsticks: Combating Covariate Shift in Model-Free Imitation Learning for Fine Manipulation. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May 2021; pp. 6185–6191.
20. Liu, W.; Wang, J.; Wang, Y.; Wang, W.; Lu, C. ForceMimic: Force-Centric Imitation Learning with Force-Motion Capture System for Contact-Rich Manipulation. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA), Atlanta, GA, USA, 19 May 2025; pp. 1105–1112.
21. Xue, H.; Ren, J.; Chen, W.; Zhang, G.; Fang, Y.; Gu, G.; Xu, H.; Lu, C. Reactive Diffusion Policy: Slow-Fast Visual-Tactile Policy Learning for Contact-Rich Manipulation. *arXiv* **2025**. [CrossRef]
22. Watanabe, R.; Alvarez, M.; Ferreiro, P.; Savkin, P.; Sano, G. FTACT: Force Torque aware Action Chunking Transformer for Pick-and-Reorient Bottle Task. *arXiv* **2025**. [CrossRef]
23. Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; Song, S. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. *Int. J. Robot. Res.* **2025**, *44*, 1684–1704. [CrossRef]
24. Bharadhwaj, H.; Vakil, J.; Sharma, M.; Gupta, A.; Tulsiani, S.; Kumar, V. RoboAgent: Generalization and Efficiency in Robot Manipulation via Semantic Augmentations and Action Chunking. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13 May 2024; pp. 4788–4795.
25. Kim, M.J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Vuong, Q.; et al. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv* **2024**. [CrossRef]
26. Shafiullah, N.M.; Cui, Z.J.; Altanzaya, A.; Pinto, L. Behavior Transformers: Cloning k Modes with One Stone. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 28 November–9 December 2022; pp. 22955–22968.
27. Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Hsu, J.; et al. RT-1: Robotics Transformer for Real-World Control at Scale. *arXiv* **2023**. [CrossRef]
28. Pari, J.; Shafiullah, N.M.; Arunachalam, S.P.; Pinto, L. The Surprising Effectiveness of Representation Learning for Visual Imitation. *arXiv* **2021**. [CrossRef]
29. Zhang, T.; McCarthy, Z.; Jow, O.; Lee, D.; Chen, X.; Goldberg, K.; Abbeel, P. Deep Imitation Learning for Complex Manipulation Tasks from Virtual Reality Teleoperation. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 5628–5635.
30. Rahmatizadeh, R.; Abolghasemi, P.; Boloni, L.; Levine, S. Vision-Based Multi-Task Manipulation for Inexpensive Robots Using End-to-End Learning from Demonstration. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 3758–3765.
31. Ebert, F.; Yang, Y.L.; Schmeckpeper, K.; Bucher, B.; Georgakis, G.; Daniilidis, K.; Finn, C.; Levine, S. Bridge Data: Boosting Generalization of Robotic Skills with Cross-Domain Datasets. *arXiv* **2021**. [CrossRef]
32. Tsunekawa, Y.; Sekiyama, K. Dynamic Attention Analysis of Body Parts in Transformer-Based Human–Robot Imitation Learning with the Embodiment Gap. *Machines* **2025**, *13*, 1133. [CrossRef]
33. Liu, H.; Tong, Y.; Liu, G.; Ju, Z.; Zhang, Z. IDAGC: Adaptive Generalized Human-Robot Collaboration via Human Intent Estimation and Multimodal Policy Learning. In Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hangzhou, China, 19 October 2025; pp. 4480–4487.

34. Wong, C.Y.; Vergez, L.; Suleiman, W. Vision- and Tactile-Based Continuous Multimodal Intention and Attention Recognition for Safer Physical Human–Robot Interaction. *IEEE Trans. Automat. Sci. Eng.* **2024**, *21*, 3205–3215. [CrossRef]
35. Fuchs, S.; Belardinelli, A. Gaze-Based Intention Estimation for Shared Autonomy in Pick-and-Place Tasks. *Front. Neurorobot.* **2021**, *15*, 647930. [CrossRef]
36. D'Antona, A.; Farsoni, S.; Rizzi, J.; Bonfè, M. A Variable Stiffness System for Impact Analysis in Collaborative Robotics Applications with FPGA-Based Force and Pressure Data Acquisition. *Sensors* **2025**, *25*, 3913. [CrossRef]
37. *ISO/TS 15066:2016*. Robots and Robotic Devices—Collaborative Robots. International Organization for Standardization: Geneva, Switzerland, 2016.
38. Wang, R.; Cheng, Y.; Tay, F.E.H.; Ang, M.H., Jr. Human-Inspired Compliance Manipulation for Multi-Stiffness Assembly via a Mamba-2-Enhanced Transformer Framework. *IEEE Robot. Autom. Lett.* **2025**, *10*, 12349–12356. [CrossRef]
39. Kang, H.; Ma, H.; Li, W. CATCH-FORM-ACTer: Compliance-Aware Tactile Control and Hybrid Deformation Regulation-Based Action Transformer for Viscoelastic Object Manipulation. *IEEE Access* **2025**, *13*, 131998–132005. [CrossRef]
40. Fusco, A.; Modugno, V.; Kanoulas, D.; Rizzo, A.; Cognetti, M. Transformer-Based Prediction of Human Motions and Contact Forces for Physical Human-Robot Interaction. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13 May 2024; pp. 3161–3167.
41. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv* **2024**. [CrossRef]