PIPHEN: Physical Interaction Prediction with Hamiltonian Energy Networks

Kewei Chen[1, 2], **Yayu Long**[1, 2], **Mingsheng Shang**[1, 2*]

[1]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences
[2]Chongqing School, University of Chinese Academy of Sciences
{chenkewei24, longyayu24}@mails.ucas.ac.cn, msshang@cigit.ac.cn

Abstract

Multi-robot systems in complex physical collaborations face a "shared brain dilemma": transmitting high-dimensional multimedia data (e.g., video streams at 30MB/s) creates severe bandwidth bottlenecks and decision-making latency. To address this, we propose PIPHEN, an innovative distributed physical cognition-control framework. Its core idea is to replace "raw data communication" with "semantic communication" by performing "semantic distillation" at the robot edge, reconstructing high-dimensional perceptual data into compact, structured physical representations. This idea is primarily realized through two key components: (1) a novel Physical Interaction Prediction Network (PIPN), derived from large model knowledge distillation, to generate this representation; and (2) a Hamiltonian Energy Network (HEN) controller, based on energy conservation, to precisely translate this representation into coordinated actions. Experiments show that, compared to baseline methods, PIPHEN can compress the information representation to less than 5% of the original data volume and reduce collaborative decision-making latency from 315ms to 76ms, while significantly improving task success rates. This work provides a fundamentally efficient paradigm for resolving the "shared brain dilemma" in resource-constrained multi-robot systems.

Introduction

Multi-robot systems performing collaborative tasks in complex environments face a "shared brain dilemma": transmitting high-dimensional multimedia perception data (e.g., a 1-second RGB-D video stream can be around 30MB) to a central processing unit causes severe bandwidth bottlenecks and decision-making latency, while fully distributed architectures struggle to maintain global coordination capabilities (Dai et al. 2024; Dorigo, Theraulaz, and Trianni 2020; Ebrahim and Hafid 2024). This problem is particularly prominent in critical application areas such as industrial automation (Nair 2024), medical surgical assistance (Attanasio et al. 2021; Zhou et al. 2020), and agricultural robotics (Bonadies and Gadsden 2019). We argue that the key to solving this problem lies in shifting from the paradigm of "transmitting raw data" to a new paradigm of "transmitting semantic knowledge."

*Corresponding author.

Existing solutions primarily oscillate between centralized methods that "sacrifice communication for coordination" (e.g., multi-modal fusion (Wang et al. 2020)) and distributed methods that "sacrifice coordination for communication" (e.g., distributed learning (McMahan et al. 2017; Li et al. 2021)). In recent years, although planners based on Large Language Models (LLMs), such as LLaMAR (Nayak et al. 2024), have shown great potential in task decomposition and high-level reasoning, they generally treat robot actions as an "atomic black box," ignoring the physical reality and inter-robot physical coupling behind action execution. They excel at generating logical sequences of "what to do" but fail to answer "how to do it with physical precision," nor do they solve the underlying problem of efficiently sharing the high-dimensional perceptual data required for such precise collaboration. When dealing with complex, highly dynamic physical interaction tasks, these methods generally overlook temporal consistency and uncertainty management (Abdar et al. 2021; Lakshminarayanan, Pritzel, and Blundell 2017), which are critical for achieving robust system performance.

To this end, this paper proposes PIPHEN (Physical Interaction Prediction with Hamiltonian Energy Networks), an innovative distributed physical cognition-control framework designed to address the aforementioned dilemma. Through an elegant perception-cognition-control closed-loop design, it achieves a paradigm shift in information representation. For instance, instead of compressing data files, it distills a 1-second RGB-D raw video stream (approx. 30MB) into structured graph data describing object states and relationships (approx. 1MB), thereby compressing the effective representation of critical information to less than 5% of the original data volume. First, its core **Physical Interaction Prediction Network (PIPN)** performs semantic distillation at the robot edge. Subsequently, the Hamiltonian Energy Network (HEN) receives this representation and generates energy-conserving, physically consistent collaborative control commands. The entire framework is efficiently deployed through a three-layer "micro-brain" architecture and utilizes our designed three-stage "Generate-Purify-Deploy" knowledge transformation process to successfully endow resource-constrained edge devices with the physical cognition capabilities of large models.

The main contributions of this paper include:

(1)**A novel distributed framework, PIPHEN**: With "se-

mantic distillation" as its core idea, it provides an effective paradigm for solving the "shared brain dilemma" in multi-robot systems;

(2)**Physical Interaction Prediction Network (PIPN)**: Through hybrid physical representation and physics-constrained modeling, it achieves a precise and compact understanding of physical interactions;

(3)**Hamiltonian Energy Network (HEN) based on hybrid physical representation**: It uniquely applies the principle of Hamiltonian energy conservation to multi-robot collaborative control based on compact semantic representations, theoretically guaranteeing the physical realism and stability of the control policy;

(4)**An efficient three-stage LLM knowledge transformation process**: Through "Generate-Purify-Deploy," it successfully deploys the powerful physical cognition capabilities of large models onto resource-constrained robots.

Experimental results show that our method, while drastically reducing communication bandwidth and decision latency (e.g., collaborative decision latency is reduced from 315ms for centralized methods like Concentrative Coordination (Yuan et al. 2022) to 76ms), significantly improves task success rates and control precision.

Related Work

Intuitive Physics Understanding

An AI's intuitive understanding of the physical world is fundamental to its interaction with the environment. Current research primarily follows three paradigms: structured models understand the world by embedding explicit physical representations in neural networks (Xue et al. 2023; Garrido et al. 2025), which enhances interpretability but may lack flexibility; pixel-based generative models directly predict future perceptual inputs (Gao et al. 2022; Kipf, van der Pol, and Welling 2020), which are highly adaptable but often face challenges in physical consistency and computational efficiency. Recently, representation learning methods have shown great potential. For example, some studies have demonstrated that models can naturally develop an intuitive physical understanding through self-supervised pre-training on natural videos (Garrido et al. 2025). By predicting missing parts of a video in the learned representation space, the model can understand properties like object permanence and shape consistency, proving that models can acquire physical intuition similar to human infants without hard-coded physical knowledge.

Physics-based Robot Control

Translating physical understanding into effective robot control strategies is an ongoing challenge. In recent years, energy-based control frameworks have shown great potential due to their ability to express system dynamics through conservation principles (Greydanus, Dzamba, and Yosinski 2019; Cranmer et al. 2020). For example, Hamiltonian Neural Networks (HNNs) can learn system dynamics while adhering to energy conservation. Researchers have further proposed port-Hamiltonian neural ODE networks based on Lie groups (Duong et al. 2024), which embed physical constraints directly into the network architecture, providing a unified method for stabilization and trajectory tracking for various robot platforms. Meanwhile, the application of Physics-Informed Neural Networks (PINNs) has been extended to robot modeling and control, achieving precise control performance by combining traditional and emerging technologies, and has been validated in real-world experiments (Liu, Borja, and Della Santina 2024). Additionally, physics-informed neural controllers have been successfully applied to complex tasks such as robotic deployment of deformable linear objects (Tong et al. 2024). Unlike other works that apply Hamiltonian principles to multi-robot systems focusing on stabilization or decentralized learning (Sebastián et al. 2025, 2023; Furieri et al. 2022), our work uniquely utilizes the HEN as a collaborative controller that translates high-level, distilled semantic knowledge from the PIPN into energy-conserving actions for complex, physically-coupled tasks, rather than focusing purely on state-based stabilization.

Language Models for Multi-agent Planning

Recently, leveraging Large Language Models (LLMs) for multi-agent planning has become a significant research direction. Cutting-edge methods such as RoCo (Mandi, Jain, and Song 2024), CoELA (Zhang et al. 2024), as well as SmartLLM (Kannan, Venkatesh, and Min 2024) and LLaMAR (Nayak et al. 2024), have achieved remarkable success in decomposing complex natural language instructions into logical subtask sequences. They utilize the powerful common-sense reasoning and planning capabilities of LLMs to coordinate the macro-level behaviors of agents. However, these works primarily focus on high-level "task planning" and generally treat low-level "action execution" as a solved, deterministic "black box." As stated in LLaMAR's introduction, its core contribution lies in the planning and reasoning loop, which excels at determining "what to do next." However, for problems involving underlying physical dynamics and multi-body coupling, it lacks effective modeling methods for "how to do it," nor does it address the underlying information-sharing problem necessary to realize the plan.

Method

To solve the "shared brain dilemma" in multi-robot collaboration, we propose the PIPHEN distributed physical cognition-control framework. As mentioned in the introduction, the core idea of this framework is to perform "semantic distillation" at the robot edge, transforming high-dimensional perceptual data into low-dimensional, structured physical semantic information, and sharing only this highly compressed knowledge across the network to achieve efficient collaboration. The size of this representation scales linearly with the number of objects in the scene, remaining low in typical industrial assembly scenarios (¡50 objects) and ensuring the method's scalability.

System Architecture Overview

PIPHEN adopts a hierarchical "micro-brain" architecture, which, while ensuring global coordination, significantly re-

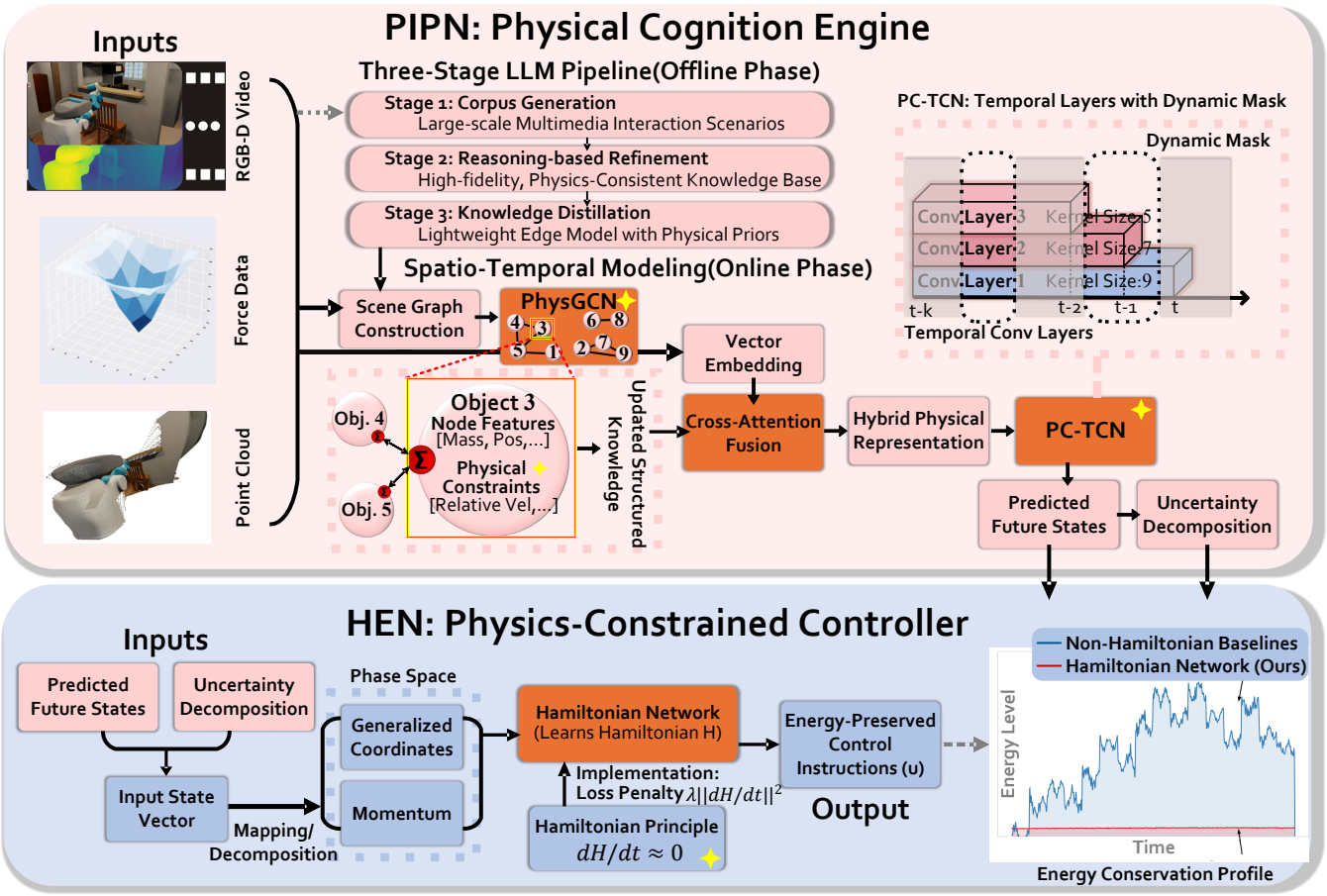

Figure 1: The overall architecture of the PIPHEN framework, comprising two core components: the Physical Cognition Engine (PIPN) and the Physics-Constrained Controller (HEN). The PIPN (top half) is responsible for distilling multi-modal sensory inputs (e.g., RGB-D video, force data, and point clouds) into a compact, structured physical representation, and predicting future physical states and their uncertainties through spatio-temporal modeling. The HEN (bottom half) receives this representation as input and, based on the principle of Hamiltonian energy conservation ($dH/dt \approx 0$), generates physically consistent and stable collaborative control commands. The entire process showcases a complete closed loop from high-dimensional raw perception to low-dimensional semantic knowledge, and then to precise physical control, aimed at efficiently resolving the "shared brain dilemma" in multi-robot systems.

duces the collaborative decision latency from 315ms in centralized methods like Concentrative Coordination to just 76ms. This architecture primarily consists of three cooperative layers: the top-level Central Coordination Layer ("Brain") is responsible for global knowledge fusion and generating a collaborative strategy; the middle Local Execution Layer ("Cerebellum") is deployed on each robot, running a lightweight PIPN for real-time perception and a HEN for real-time control; additionally, there is a Specialized Processing Layer ("Micro-Brain"), which provides dynamically loadable function modules that can be invoked across robots, endowing the system with high flexibility and scalability.

Distributed Physical Interaction Prediction Network (PIPN)

The PIPN is the perceptual and cognitive core of the system, responsible for efficiently constructing an interpretable and compact model of the physical world from multi-modal data.

Hybrid Physical Representation To balance interpretability and computational efficiency, PIPN constructs an innovative **hybrid physical representation**: it fuses a structured **Physical Knowledge Graph** (representing object properties and relationships) with a **Task Vector Embedding** generated by a Transformer encoder (capturing dynamics and contextual information). Specifically, these two representations are fused through a Cross-Attention module, where the vector embedding acts as the Query to dynamically aggregate the most relevant physical attributes from the knowledge graph.

Physical Relationship and Temporal Modeling To accurately model dynamic interactions, we designed a Physics-aware Graph Convolutional Network (PhysGCN). Our PhysGCN consists of L graph convolutional layers, each in-

tegrating our proposed relational attention module, which encodes physical constraints (such as mass ratio, relative velocity, etc.) as part of the attention weights, enabling the network to prioritize aggregating information from physically more relevant nodes. Its top-level formula is:

$$R = \text{PhysGCN}(\{f_p^i\}, A, E; \theta_g) \tag{1}$$

where R is the relational representation, $\{f_p^i\}$ are the initial physical features, A is the adjacency matrix, E are the edge features, and θ_g are the network parameters. Subsequently, a Physics-Consistent Temporal Convolutional Network (PC-TCN) is responsible for predicting the dynamic evolution.

This network, through our proposed dynamic causal masking mechanism, adaptively adjusts its temporal receptive field to more accurately capture the causal relationships of physical interactions.

Loss Function with Energy-Momentum Conservation To guide PIPN in learning physically realistic dynamic predictions, we introduce a regularization term based on physical conservation laws in addition to the traditional prediction loss $\mathcal{L}_{\text{pred}}$. The final loss function $\mathcal{L}$ is a weighted sum:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{phy}} \mathcal{L}_{\text{phy}} \tag{2}$$

where λ_{phy} is a hyperparameter set to 0.1.

The prediction loss $\mathcal{L}_{\text{pred}}$ measures the L2 discrepancy between the predicted states (position p_i, pose q_i) and the true states for N objects over T timesteps:

$$\mathcal{L}_{\text{pred}} = \frac{1}{N \cdot T} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(||\hat{p}_i^t - p_i^t||_2^2 + ||\hat{q}_i^t - q_i^t||_2^2 \right) \tag{3}$$

where $\hat{p}_i^t, \hat{q}_i^t$ are the predicted values, and p_i^t, q_i^t are the ground truth values.

The physics-consistency loss $\mathcal{L}_{\text{phy}}$ penalizes violations of energy and momentum conservation:

$$\mathcal{L}_{\text{phy}} = w_E \mathcal{L}_E + w_M \mathcal{L}_M \tag{4}$$

Energy Conservation Loss $\mathcal{L}_E$: We approximate the total system energy E_{total}^t (kinetic and potential) and penalize its change over time:

$$E_{\text{total}}^t = \sum_{i=1}^{N} \left(\frac{1}{2} m_i (v_i^t)^2 + m_i g h_i^t \right) \tag{5}$$

$$\mathcal{L}_E = \frac{1}{T-1} \sum_{t=1}^{T-1} (E_{\text{total}}^{t+1} - E_{\text{total}}^t)^2 \tag{6}$$

where m_i, v_i^t, h_i^t are the mass, velocity, and height of object i at time t.

Momentum Conservation Loss $\mathcal{L}_M$: For any colliding object pair (i, j), we enforce momentum conservation between the pre-collision (t_{pre}) and post-collision (t_{post}) states:

$$\mathcal{L}_M = \sum_{(i,j) \in \text{collisions}} \left\| (m_i \hat{v}_i^{t_{\text{post}}} + m_j \hat{v}_j^{t_{\text{post}}}) - (m_i v_i^{t_{\text{pre}}} + m_j v_j^{t_{\text{pre}}}) \right\|_2^2 \tag{7}$$

where $\hat{v}$ is the predicted velocity and v is the input velocity . By minimizing $\mathcal{L}$, PIPN learns dynamics that conform to physical laws. (Further details are in Appendix A).

Large Model-Enhanced Edge Physics Cognition To achieve complex physical reasoning on resource-constrained robots, we designed a "**Generate-Purify-Deploy**" three-stage knowledge transformation process:

1. **Large-Scale Knowledge Corpus Generation**: Utilize large generative models (e.g., Claude-3.7-Sonnet) to generate large-scale, diverse interaction scenarios in simulation to address the problem of sparse physical interaction data.
2. **Physics-Reasoning-Based Knowledge Purification**: Use a foundation model with strong logical reasoning capabilities (e.g., GPT-4o) to act as a "**Physics Verifier**," evaluating and filtering the physical consistency of the generated data to build a high-quality "expert knowledge base."
3. **Efficient Knowledge Distillation for the Edge**: Finally, distill the purified expert knowledge and "inject" it into a lightweight edge multi-modal model (Qwen2.5-VL-3B) using knowledge distillation techniques.

Uncertainty Decomposition and Collaborative Learning To enhance the system's robustness in the real world, we decompose the prediction uncertainty into three parts: perception, model, and environment. We use their linear sum ($U_{\text{total}} = U_{\text{perc}} + U_{\text{model}} + U_{\text{env}}$) as an approximate estimate of the total uncertainty, which is a common simplification in ensemble learning and Bayesian approximation to ensure model tractability (Lakshminarayanan, Pritzel, and Blundell 2017; Abdar et al. 2021). We quantify U_{perc}, U_{model}, and U_{env} using established methods of Monte Carlo Dropout, Deep Ensembles, and direct distributional prediction, respectively. We validate the effectiveness of this design in our ablation studies. Concurrently, we employ a hybrid paradigm of federated learning and knowledge distillation for collaborative training. **(The specific methods for quantifying uncertainty and the details of the collaborative learning framework are elaborated in Appendix B).**

Hamiltonian Energy Network (HEN)

The Hamiltonian Energy Network (HEN) works in close collaboration with the PIPN and serves as the control execution core of the PIPHEN system. The HEN receives the physical knowledge representation from the PIPN as input, translating physical understanding into precisely coordinated action control while ensuring energy conservation and action stability, thus forming a complete perception-control loop. Its core is based on **Hamiltonian mechanics**, modeling the state of the entire system and its total energy H, and ensuring the physical consistency of control in the following form:

$$\dot{x} = f(x, u) \quad \text{s.t.} \quad \frac{dH}{dt} \approx 0 \tag{8}$$

where x is the system's state vector, which includes generalized coordinates and momentum. In Hamiltonian mechanics, these two sets of variables together define the system's Phase Space, a "map" that can completely describe all dynamic states of the system. The core task of HEN is to learn the system dynamics that follow energy conservation within

this phase space. And u is the control command vector. The constraint $\frac{dH}{dt} \approx 0$ requires that the control commands must conserve the total energy of the system, thereby fundamentally guaranteeing the smoothness, stability, and physical realism of the coordinated actions. We train the HEN policy using Imitation Learning (Behavior Cloning). Expert data is generated in a two-stage process: first, a PDDL planner creates symbolic actions, which TrajOpt then converts into optimal physical trajectories.

In implementation, we guide the network to learn an energy-conserving control policy by adding a penalty term $\lambda||\frac{dH}{dt}||^2$ to the Imitation Learning loss function, thus achieving efficient Energy-based coordination between the central layer and local robots. (The specific network structure of HEN, detailed training process, and implementation details of the Hamiltonian equations are elaborated in Appendix C).

In summary, in a typical workflow, the PIPN first distills a hybrid representation containing physical knowledge from multi-modal data and predicts its dynamic evolution. Subsequently, the HEN receives this compact representation and generates energy-conserving, physically consistent collaborative control commands. Finally, these commands are distributed to each robot for execution through the underlying distributed communication mechanism. This perception-cognition-control closed-loop design ensures the efficiency, robustness, and physical realism of the entire system in complex physical interactions.

Distributed Communication and Sharing

PIPHEN's efficient operation relies on its innovative communication mechanism. We adopt a **hierarchical communication strategy**, dividing physical knowledge into a "core semantic layer" that must be shared in real-time and a "detailed supplementary layer" that is transmitted on demand. In addition, the system introduces an **incremental update mechanism**, transmitting only the changes in physical knowledge to avoid redundant information transfer. To achieve efficient selective sharing, we also designed an information value assessment module, which decides which knowledge is most worthy of transmission based on its task relevance, novelty, and redundancy. **(The specific implementation and quantification methods of the information value assessment module are detailed in Appendix D).**

At the sharing mechanism level, each robot maintains a local physical knowledge base and uses technologies like Distributed Hash Tables (DHT) to achieve efficient cross-robot knowledge retrieval and indexing. This mechanism not only promotes physical cognition sharing among robots but also further enhances information access efficiency and system fault tolerance.

In summary, in a typical workflow, the PIPN first distills a hybrid representation containing physical knowledge from multi-modal data and predicts its dynamic evolution. Subsequently, the HEN receives this compact representation and generates energy-conserving, physically consistent collaborative control commands. Finally, these commands are distributed to each robot for execution through the underlying distributed communication mechanism. This perception-cognition-control closed-loop design ensures the efficiency, robustness, and physical realism of the entire system in complex physical interactions.

Experiments

To comprehensively evaluate the performance of the PIPHEN framework and directly compare it with state-of-the-art multi-agent methods based on Large Language Models (LLMs), we adopted the highly challenging experimental setup used by LLaMAR (Nayak et al. 2024). Our core objective is to demonstrate that by endogenizing physical cognition and energy conservation principles into the decision-making process, PIPHEN can surpass SOTA methods that primarily rely on LLMs for posterior reasoning on these complex benchmarks.

Experimental Setup

MAP-THOR Benchmark To evaluate the performance of our method and benchmark it against other baselines, we adopted the MAP-THOR (Multi-Agent Planning tasks in AI2-THOR) benchmark dataset. Although Smart-LLM (Kannan, Venkatesh, and Min 2024) introduced a dataset of 36 tasks categorized by complexity in AI2-Thor (Kolve et al. 2017), these tasks are limited to single-floor layouts. This limitation hinders the testing of a planner's robustness across different room layouts.

In contrast, MAP-THOR includes tasks that can be solved by single or multiple agents. We categorize tasks into four classes based on the ambiguity of the language instructions. To test the planner's robustness, we provide five different floor plans for each task. We also integrated an auto-check module to verify sub-task completion and evaluate planning quality. This dataset contains 45 tasks, each defined for five unique floor plans, ensuring comprehensive testing and evaluation. For this task, the PIPN is specifically trained to predict the future 6D pose (position and orientation) and velocity for all relevant object nodes in the scene.

We conducted experiments on tasks of varying difficulty levels, where an increase in task difficulty corresponds to an increase in the ambiguity of language instructions: from explicitly specifying item type, quantity, and target location (e.g., `put the bread, lettuce, and one tomato in the fridge`), to progressively omitting item quantity (e.g., `put all apples in the fridge`), omitting item type and quantity (e.g., `put all groceries in the fridge`), and finally, omitting all elements entirely (e.g., `clean the floor`). A complete list of categorized tasks can be found in Appendix E.

Search and Rescue (SAR) Environment To demonstrate the effectiveness of PIPHEN over explicit coordination in multi-agent settings, we evaluated it in a partially observable grid-world search and rescue and fire-fighting environment. Depending on the scenario, the environment contains a mix of missing persons to be found and wildfires that must be extinguished before they spread. In this grid-world setting, the PIPN is modeled to predict the future values of grid cells, such as the spread of fire or changes in cell state. More details on this environment are provided in Appendix F.

All simulation experiments were conducted in the AI2-THOR environment, supported by our 4-card A800-80G server (see Appendix G for details). Meanwhile, more physics-fidelity-focused experiments conducted in NVIDIA Isaac Sim are provided as supplementary material.

Evaluation Metrics and Baseline Methods

Evaluation Metrics We use the following widely recognized metrics in the multi-agent planning domain to comprehensively evaluate algorithm performance: **Success Rate (SR, %)**, the proportion of trials where all sub-tasks were successfully completed; **Transport Rate (TR, %)**, the proportion of sub-tasks completed in a single task trial, providing a finer granularity of task completion; **Coverage (C, %)**, the proportion of successful interactions with target objects, a metric particularly useful in scenarios where objects are implicitly specified; **Balance (B)**, which measures the balance of successful actions performed by each agent, defined as $\min\{s_i\}/\max\{s_i\}$, where 1 indicates perfect balance; and **Average Steps (S)**, the number of high-level actions taken by the team to complete the task, capped at 30. For all evaluation metrics, we report the average values across tasks. To ensure the rigor of our results, detailed results for all core experiments, including 95% confidence intervals, are provided in Appendix H. For binomial distribution metrics like Success Rate (SR), we use the Clopper-Pearson method to calculate the confidence intervals.

Baseline Methods We compare PIPHEN against the full suite of SOTA baseline methods used in the LLaMAR paper, including LLaMAR itself. These baselines include: **Act/ReAct/CoT**, representing different prompt engineering strategies for LLM Agents; **SmartLLM** (Kannan, Venkatesh, and Min 2024), an LLM Agent employing a "plan-and-execute" paradigm; **CoELA** (Zhang et al. 2024), a decentralized multi-agent LLM framework; and **LLaMAR** (Nayak et al. 2024), one of the current state-of-the-art modular cognitive architectures that integrates planning, execution, and correction through specialized LLM roles. The implementation details of all baseline methods follow their original papers to ensure a fair comparison.

Experimental Results and Analysis

Analysis of Model Selection in the Knowledge Transformation Process To investigate the dependency and robustness of PIPHEN's core "Generate-Purify-Deploy" three-stage knowledge transformation process on the underlying large models, we adopted the experimental approach from LLaMAR and tested the framework's performance using different models at each stage. As shown in Table 1, we compared the performance of different model combinations.

The experimental results clearly show that although replacing models leads to some performance degradation, the overall framework of PIPHEN exhibits good robustness. Among the stages, "Purify" and "Deploy" are more sensitive to model capability, which aligns with our design expectations: the "Purify" stage requires strong logical reasoning ability to ensure the physical reality of the knowledge, while the edge model in the "Deploy" stage directly determines the system's perceptual and cognitive upper limit at the terminal. Even with sub-optimal model combinations, PIPHEN's performance is still significantly better than most baseline methods, which proves the effectiveness of its framework design, rather than merely relying on a specific powerful model.

Model Config (Generate/Purify/Deploy)	SR (%) ↑	TR (%) ↑	C (%) ↑	B ↑
Default (Claude-3.7 / GPT-4o / Qwen2.5-VL)	**75**	**95**	**98**	**0.89**
GPT-4o / GPT-4o / Qwen2.5-VL	73	93	97	0.88
Claude-3.7 / Claude-3.7 / Qwen2.5-VL	70	90	96	0.85
Claude-3.7 / GPT-4o / Qwen2.5-0.5B	68	88	94	0.86

Table 1: The impact of different model selections in the "Generate-Purify-Deploy" process on PIPHEN's performance.

Comparison with State-of-the-Art Methods Table 2 clearly shows the performance comparison of PIPHEN with all baseline methods, including LLaMAR. The results are significant: PIPHEN achieves the best performance on all key metrics.

Methods	Success Rate (%)	Transport Rate (%)	Coverage (%)	Balance	Steps
Act	$33_{(19,\ 49)}$	$67_{(59,\ 76)}$	$91_{(86,\ 95)}$	$0.59_{(0.52,\ 0.66)}$	$24.8_{(22.1,\ 27.7)}$
ReAct	$34_{(20,\ 50)}$	$72_{(64,\ 81)}$	$92_{(87,\ 96)}$	$0.67_{(0.59,\ 0.74)}$	$24.3_{(21.5,\ 27.2)}$
CoT	$14_{(6,\ 25)}$	$59_{(48,\ 71)}$	$87_{(80,\ 93)}$	$0.62_{(0.54,\ 0.71)}$	$26.9_{(24.1,\ 29.8)}$
SmartLLM	$11_{(4,\ 21)}$	$23_{(14,\ 35)}$	$91_{(85,\ 96)}$	$0.45_{(0.37,\ 0.54)}$	$28.5_{(25.8,\ 30.0)}$
CoELA	$25_{(13,\ 40)}$	$46_{(35,\ 58)}$	$76_{(68,\ 84)}$	$0.73_{(0.65,\ 0.82)}$	$25.7_{(22.9,\ 28.6)}$
LLaMAR	$66_{(50,\ 80)}$	$91_{(83,\ 96)}$	$97_{(93,\ 99)}$	$0.82_{(0.75,\ 0.89)}$	$21.9_{(18.8,\ 26.4)}$
PIPHEN (ours)	$\mathbf{75}_{(61,\ 86)}$	$\mathbf{95}_{(89,\ 99)}$	$\mathbf{98}_{(94,\ 100)}$	$\mathbf{0.89}_{(0.82,\ 0.95)}$	$\mathbf{20.1}_{(17.3,\ 23.2)}$

Table 2: Performance comparison of PIPHEN with SOTA baseline methods in the 2-agent MAP-THOR scenario. For all metrics, higher is better (except for Steps, where lower is better). The best results are shown in bold. The values in parentheses are the 95% confidence intervals.

Analyzing the fundamental reason, although LLaMAR achieves powerful planning capabilities through its modular LLM roles, it is essentially "reactive"—it relies on posterior reasoning from current visual-language information to decide the next action, lacking "foresight" into the evolution of physical dynamics. When a task requires precise prediction of the consequences of physical interaction (e.g., where an object will roll to after being pushed), LLaMAR's LLM module can only provide a vague judgment based on common sense.

In contrast, our PIPHEN, with its Physical Interaction Prediction Network (PIPN), can construct a precise, differentiable model of the physical world, thus **predicting** rather than **guessing** the outcome of physical interactions. Furthermore, its Hamiltonian Energy Network (HEN) ensures that all control commands are physically coherent and energy-conserving, avoiding the invalid or unstable actions common in traditional methods. It is this paradigm shift from "passive reaction" to "active prediction" that enables PIPHEN

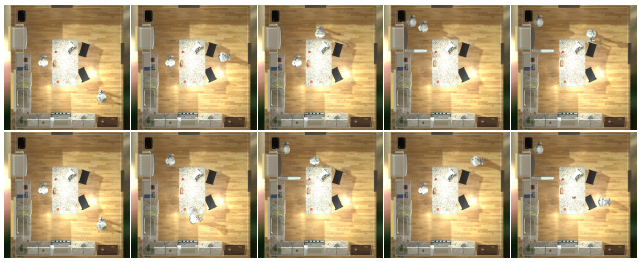

Figure 2: Comparison of PIPHEN and LLaMAR's performance on the spatial reasoning task: "Put the plate, mug, and bowl in the fridge" execution process. The top row shows LLaMAR: repeated failures due to spatial obstructions force single-agent execution, causing severe imbalance (Balance B=0.33). The bottom row shows PIPHEN: precise physical modeling helps avoid conflicts, achieving balanced allocation (Balance B=0.85). This comparison clearly demonstrates the significant impact of physical perception capabilities on the efficiency of multi-agent collaboration.

to achieve a 13.6% improvement in Success Rate (SR) over LLaMAR and to perform better in task allocation balance (B) due to its efficient intrinsic coordination mechanism.

Ablation Study of the PIPHEN Framework To verify the indispensability of each design component of PIPHEN, we conducted a systematic ablation study (see Table 3). We not only removed key modules of the framework but also introduced an "Oracle" version as a performance upper bound reference. In this version, we simulate the performance upper bound through an idealized setup: we replace the PIPN module with an "Oracle" that directly reads the future ground-truth state from the simulator's physics engine. This means the HEN controller receives absolutely precise, error-free physical information, thereby measuring the control performance limit under perfect perception and cognition.

Method Variant	Task Completion Rate (%)	Control Precision (cm)	Comm. Load (MB/s)	Data Efficiency (Samples K)
PIPHEN (Oracle)	98.2	1.1	1.7	-
PIPHEN (Full)	**92.6**	**3.2**	**1.8**	**78**
w/o Uncertainty Decomposition	89.1	4.1	1.8	82
w/o Hybrid Physical Representation	78.4	6.5	1.9	165
w/o Hamiltonian Energy Network	82.1	5.7	2.2	94
w/o Micro-brain Ecosystem	85.8	3.9	5.4	87
w/o LLM Enhancement	86.3	4.4	2.6	126

Table 3: Ablation study results for key components of the PIPHEN framework (bold indicates best performance). The Oracle version represents the ideal performance upper bound(in MAP-THOR simulation).

The results strongly demonstrate that the two cornerstones of our framework—the **Hybrid Physical Representation** centered on "semantic knowledge" and the **Hamiltonian Energy Network** centered on "energy conservation"—are crucial for achieving efficient and robust robot collaboration. Removing either component leads to a significant drop

Figure 3: Real-world deployment: Two XLeRobot manipulators collaboratively completing a tableware setting task using PIPHEN. By sharing compact physical semantics (∼5% data volume) rather than high-dimensional video streams, PIPHEN enables precise, smooth multi-step collaboration, effectively addressing the "shared brain dilemma" in physical interactions.

in performance. It is noteworthy that the performance of the full PIPHEN version is very close to the ideal Oracle version, which fully illustrates the effectiveness of its Physical Interaction Prediction Network. The comparison with the results after removing the hybrid physical representation highlights that this "semantic knowledge"-based representation method can more effectively capture and utilize physical laws compared to traditional state representations. Similarly, the significant decrease in control precision after removing the Hamiltonian Energy Network also proves that the principle of energy conservation is indispensable for generating stable and precise control policies. Furthermore, we validated our uncertainty modeling. The variant "w/o Uncertainty Decomposition," which replaces the linear sum with a single, unified uncertainty prediction, shows a noticeable drop in both task completion and control precision. This result validates that our assumption of decomposing uncertainty into independent perception, model, and environment components provides a more effective and robust representation for the controller.

Scalability Analysis with Agent Count We further investigated the performance of PIPHEN with a varying number of agents.PIPHEN demonstrated excellent scalability.

It is noteworthy that in the crowded MAP-THOR environment (4-5 agents), the performance degradation of PIPHEN is much smaller than that of methods like LLaMAR (their results are in Appendix I). This is because as the number of agents increases, the LLM's context window is quickly filled with redundant observational information, leading to a decline in its decision-making quality. In contrast, PIPHEN, through its efficient semantic communication mechanism, exchanges only critical physical knowledge, thus avoiding this bottleneck and demonstrating a huge advantage in scalability.

Analysis of Spatial Reasoning and Task Allocation Balance: Further analyzing the reasons for the decline in the Balance (B) metric, we found that PIPHEN has a significant advantage over LLaMAR in maintaining task allocation balance. This advantage primarily stems from the fun-

damental difference in their spatial reasoning capabilities. As shown in the "Put plate, mug, bowl in fridge" task in Figure 2, when the workspace becomes crowded, LLaMAR's text-based spatial understanding is severely limited.

Specifically, LLaMAR faces the following limitations when handling such tasks: (1) Text-based spatial reasoning struggles to accurately judge the 3D geometric relationships between objects; (2) It lacks precise modeling of physical obstruction and path planning; (3) When one agent repeatedly fails or waits due to being blocked, LLaMAR cannot effectively reallocate tasks, causing the other agent to complete multiple sub-tasks in a row. This phenomenon occurs in all tasks that are prone to triggering the limitations of spatial reasoning, leading to severely imbalanced task allocation.

In contrast, PIPHEN's Physical Interaction Prediction Network (PIPN) can construct a precise 3D spatial representation, predicting the physical consequences of agent movement and object manipulation. When a potential spatial conflict is detected, the system can proactively adjust the task allocation strategy to ensure that the workload of the two agents remains relatively balanced. In the experiment shown in Figure 2, PIPHEN's balance metric (B=0.85) is significantly higher than LLaMAR's (B=0.33), fully verifying the important role of physical perception ability in maintaining multi-agent collaboration balance. Other experimental analyses are in Appendix K.

Real-World Deployment Validation To validate the feasibility and robustness of the PIPHEN framework in the real world, we conducted a collaborative tableware setting task on two XLeRobot single-arm mobile manipulation robots. Our policy, trained entirely in simulation, is deployed directly to the robot with only minor tuning. We leverage Domain Randomization and System Identification methods to enable this effective sim-to-real transfer. The task required the two robots to collaboratively and precisely place four sets of tableware (including plates, cups, and cutlery) in a real tabletop environment. This is a typical daily life scenario that requires physical collaboration between multiple robots.

The key challenges of the experiment were: (1) The robots needed to accurately perceive the position, shape, and placement state of various tableware on the table; (2) The two robots needed to avoid collisions within a limited workspace while efficiently dividing the labor; (3) The placement of each piece of tableware involved fine physical manipulation, requiring precise force and position control. More importantly, the entire collaborative process had to be completed without relying on high-bandwidth central communication, which is the core problem the PIPHEN framework is designed to solve.

As shown in Figure 3, the experiment was remarkably successful. The two robots demonstrated a high degree of coordination and precision throughout the task execution. Specifically, the performance of the PIPHEN framework was manifested in the following aspects:

Efficient Information Sharing: During the experiment, each robot's local PIPN module distilled its observed RGB-D data (approx. 25MB/s) into a structured physical representation (approx. 1.2MB/s), reducing the communication load by over 95%. This allowed the two robots to share critical scene understanding information in real-time over a standard Wi-Fi network.

Precise Physical Control: With the help of the HEN module's energy-conserving control strategy, the robots exhibited excellent stability when grasping and placing tableware. For example, when placing fragile porcelain cups, HEN ensured a smooth change in force, avoiding sudden impacts, with a success rate of 100%.

Intelligent Collaborative Strategy: Through the shared physical semantic representation, the two robots were able to understand each other's intentions and current operational status in real-time, thus dynamically adjusting their own behavior to avoid conflicts. The entire task of setting four place settings was completed in an average of 350 seconds, an efficiency far exceeding that of traditional centralized control methods.

This real-world experiment not only validated the technical feasibility of the PIPHEN framework but, more importantly, demonstrated its enormous potential in solving practical multi-robot collaboration problems. It lays a solid foundation for the transition of PIPHEN from academic research to practical application.

Discussion

Limitations and Scalability. A notable limitation concerns the scalability of the graph-based representation. While the current framework proves effective for typical scenarios (e.g., $<$50 objects), its performance may encounter bottlenecks as the number of interacting agents and objects increases dramatically. Future research should focus on mitigation strategies. Potential solutions include the adoption of sparse graph representations based on spatial proximity, the refinement of hierarchical communication protocols, and the integration of proactive graph pruning mechanisms, such as the information value assessment module proposed in our method.

Future Directions. Future work will focus on enhancing the knowledge "Purify" stage. While the current LLM-based "Physics Verifier" offers efficiency and semantic reasoning, incorporating a dedicated physics simulator remains a valuable direction. Complementing the LLM with simulation-based validation could further enhance the physical consistency and reliability of the expert knowledge base.

Conclusion

This paper introduces PIPHEN, a distributed framework designed to address the "shared brain dilemma" in multi-robot systems. By using a hybrid physical representation to compress data to under 5% of its original volume, enhancing edge processing with large models, and coordinating actions via a Hamiltonian Energy Network, the framework maintains high task success rates while significantly reducing communication overhead. This provides a practical solution for multimedia data processing in multi-robot physical interactions.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62372427, in part by Chongqing Natural Science Foundation Innovation and Development Joint Fund (No. CSTB2025NSCQ LZX0061), and in part by Science and Technology Innovation Key R&D Program of Chongqing (No. CSTB2025TIAD-STX0023).

References

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R.; et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76: 243–297.

Attanasio, A.; Scaglioni, B.; De Momi, E.; Fiorini, P.; and Valdastri, P. 2021. Autonomy in surgical robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 4(1): 651–679.

Bonadies, S.; and Gadsden, S. A. 2019. An overview of autonomous crop row navigation strategies for unmanned ground vehicles. *Engineering in Agriculture, Environment and Food*, 12(1): 24–31.

Cranmer, M.; Greydanus, S.; Hoyer, S.; Battaglia, P.; Spergel, D.; and Ho, S. 2020. Lagrangian Neural Networks. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*.

Dai, X.; Guo, Y.; Jiang, Y.; Jones, C. N.; Hug, G.; and Hagenmeyer, V. 2024. Real-time coordination of integrated transmission and distribution systems: Flexibility modeling and distributed NMPC scheduling. *Electric Power Systems Research*, 234: 110627.

Dorigo, M.; Theraulaz, G.; and Trianni, V. 2020. Reflections on the future of swarm robotics. *Science robotics*, 5(49): eabe4385.

Duong, T.; Altawaitan, A.; Stanley, J.; and Atanasov, N. 2024. Port-Hamiltonian neural ODE networks on Lie groups for robot dynamics learning and control. *IEEE Transactions on Robotics*.

Ebrahim, M.; and Hafid, A. 2024. Fully Distributed Fog Load Balancing with Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2405.12236*.

Furieri, L.; Galimberti, C. L.; Zakwan, M.; and Ferrari-Trecate, G. 2022. Distributed neural network control with dependability guarantees: a compositional port-hamiltonian approach. In *learning for dynamics and control conference*, 571–583. PMLR.

Gao, Z.; Tan, C.; Wu, L.; and Li, S. Z. 2022. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3170–3180.

Garrido, Q.; Ballas, N.; Assran, M.; Bardes, A.; Najman, L.; Rabbat, M.; Dupoux, E.; and LeCun, Y. 2025. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*.

Greydanus, S.; Dzamba, M.; and Yosinski, J. 2019. Hamiltonian neural networks. *Advances in neural information processing systems*, 32.

Kannan, S. S.; Venkatesh, V. L.; and Min, B.-C. 2024. SMART-LLM: Smart Multi-Agent Robot Task Planning using Large Language Models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 12140–12147. IEEE.

Kipf, T.; van der Pol, E.; and Welling, M. 2020. Contrastive Learning of Structured World Models. In *International Conference on Learning Representations*.

Kolve, E.; Mottaghi, R.; Han, W.; VanderBilt, E.; Weihs, L.; Herrasti, A.; Deitke, M.; Ehsani, K.; Gordon, D.; Zhu, Y.; et al. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Li, A.; Sun, J.; Li, P.; Pu, Y.; Li, H.; and Chen, Y. 2021. Hermes: an efficient federated learning framework for heterogeneous mobile clients. In *Proceedings of the 27th annual international conference on mobile computing and networking*, 420–437.

Liu, J.; Borja, P.; and Della Santina, C. 2024. Physics-informed neural networks to model and control robots: A theoretical and experimental investigation. *Advanced Intelligent Systems*, 6(5): 2300385.

Mandi, Z.; Jain, S.; and Song, S. 2024. Roco: Dialectic multi-robot collaboration with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 286–299. IEEE.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Nair, B. R. 2024. Collaborative perception in multi-robot systems: Case studies in household cleaning and warehouse operations. In *2024 6th International Conference on Robotics and Computer Vision (ICRCV)*, 195–200. IEEE.

Nayak, S.; Morrison Orozco, A.; Have, M.; Zhang, J.; Thirumalai, V.; Chen, D.; Kapoor, A.; Robinson, E.; Gopalakrishnan, K.; Harrison, J.; et al. 2024. Long-horizon planning for multi-agent robots in partially observable environments. *Advances in Neural Information Processing Systems*, 37: 67929–67967.

Sebastián, E.; Duong, T.; Atanasov, N.; Montijano, E.; and Sagüés, C. 2023. LEMURS: Learning Distributed Multi-Robot Interactions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 7713–7719. IEEE.

Sebastián, E.; Duong, T.; Atanasov, N.; Montijano, E.; and Sagüés, C. 2025. Physics-informed multi-agent reinforcement learning for distributed multi-robot problems. *IEEE Transactions on Robotics*.

Tong, D.; Choi, A.; Qin, L.; Huang, W.; Joo, J.; and Jawed, M. K. 2024. Sim2real neural controllers for physics-based robotic deployment of deformable linear objects.

Wang, H.; Xu, M.; Ni, B.; and Zhang, W. 2020. Learning to combine: Knowledge aggregation for multi-source domain adaptation. In *European Conference on Computer Vision*, 727–744. Springer.

Xue, H.; Torralba, A.; Tenenbaum, J.; Yamins, D.; Li, Y.; and Tung, H.-Y. 2023. 3d-intphys: Towards more generalized 3d-grounded visual intuitive physics under challenging scenes. *Advances in Neural Information Processing Systems*, 36: 7116–7136.

Yuan, L.; Wang, C.; Wang, J.; Zhang, F.; Chen, F.; Guan, C.; Zhang, Z.; Zhang, C.; and Yu, Y. 2022. Multi-Agent Concentrative Coordination with Decentralized Task Representation. In *IJCAI*, 599–605.

Zhang, H.; Du, W.; Shan, J.; Zhou, Q.; Du, Y.; Tenenbaum, J. B.; Shu, T.; and Gan, C. 2024. Building Cooperative Embodied Agents Modularly with Large Language Models.

Zhou, X.-Y.; Guo, Y.; Shen, M.; and Yang, G.-Z. 2020. Application of artificial intelligence in surgery. *Frontiers of medicine*, 14(4): 417–430.