MistyPilot: Enabling Social-Robot Control through Multi-Agent LLM Skill Orchestration

Xiao Wang* · Lu Dong*

Ifeoma Nwogu · Srirangaraj Setlur · Venu Govindaraju

* Equal contribution

University at Buffalo

National AI Institute for Exceptional Education

Social Robots in Everyday Interaction

Social robots

- Education, assistance, and social interaction.

For a specific task, users still need to:

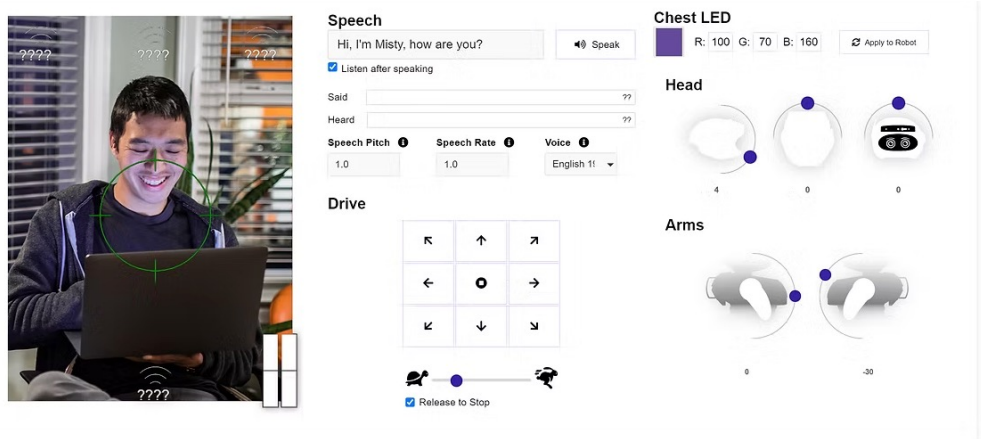
- Program robot actions.
- Define triggers and conditions.
- Design responses and interaction flows.

Example: Misty

Speech · Motion

Vision · Touch sensing

Misty Studio: action-level controls



How can we translate user requirements into executable robot behaviors?

From Skill Generation to Skill Orchestration

AutoMisty

PRIOR WORK · IROS 2025

Generate skills (control code) for open-ended tasks

- Generate robot skills from natural-language requests.
- Execute and refine the skills to complete the task.

How can available skills be selected and coordinated during interaction?

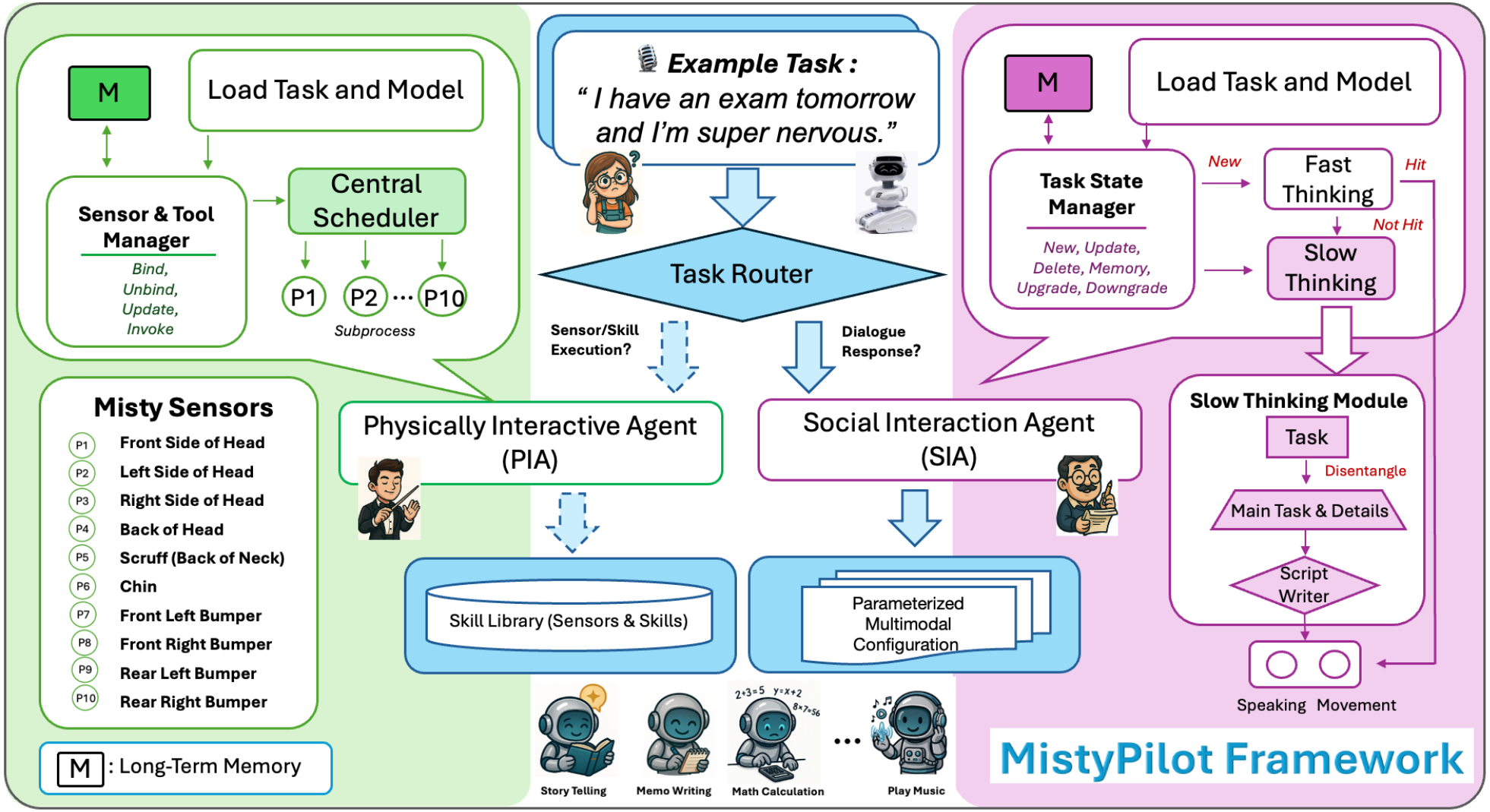
MistyPilot

THIS WORK · ECCV 2026 ACVR

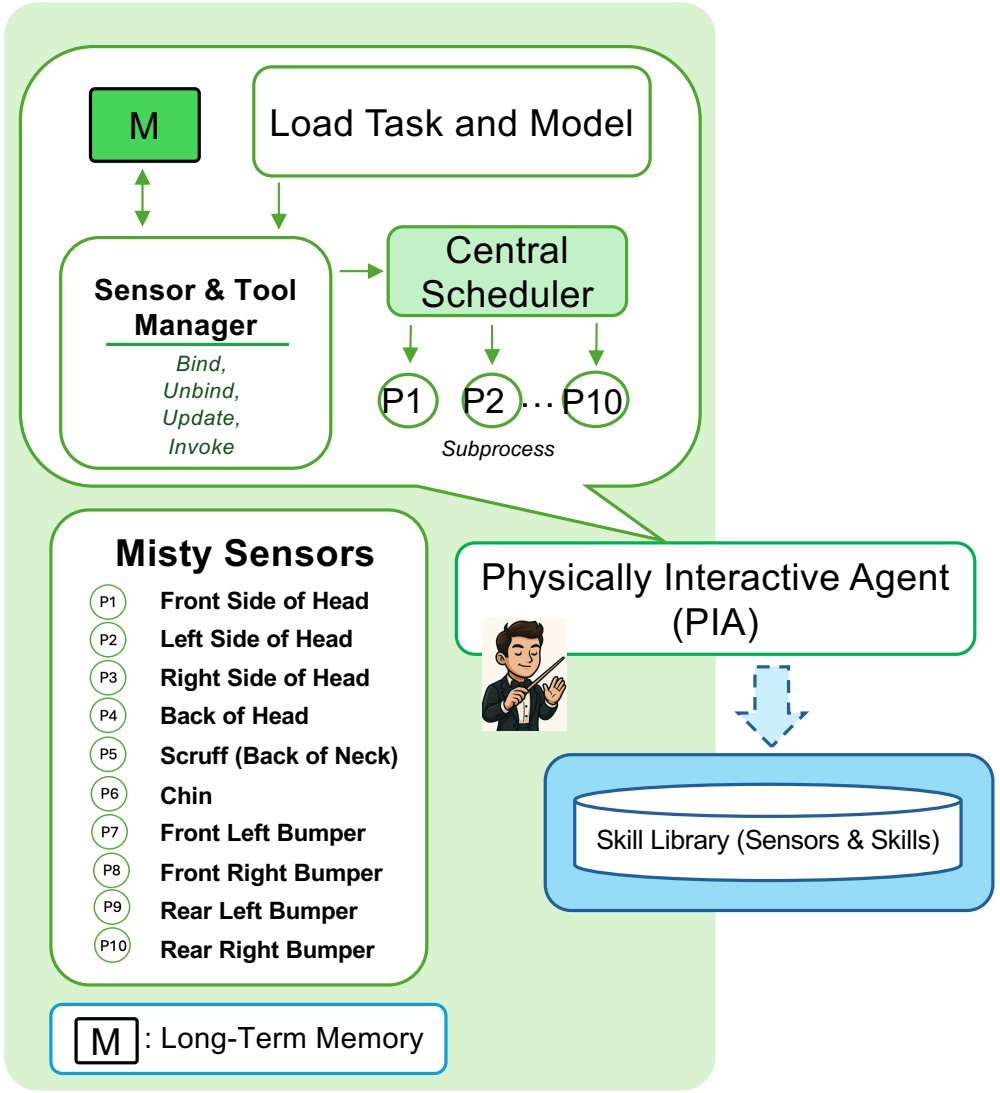
Orchestrate skills during interaction

- Select and configure skills based on user requests.
- Use sensor bindings and dialogue state to guide execution.

Method: MistyPilot Framework



Method: Physically Interactive Agent (PIA)



Natural-language control

Map high-level user instructions to robot skills.

Two execution modes

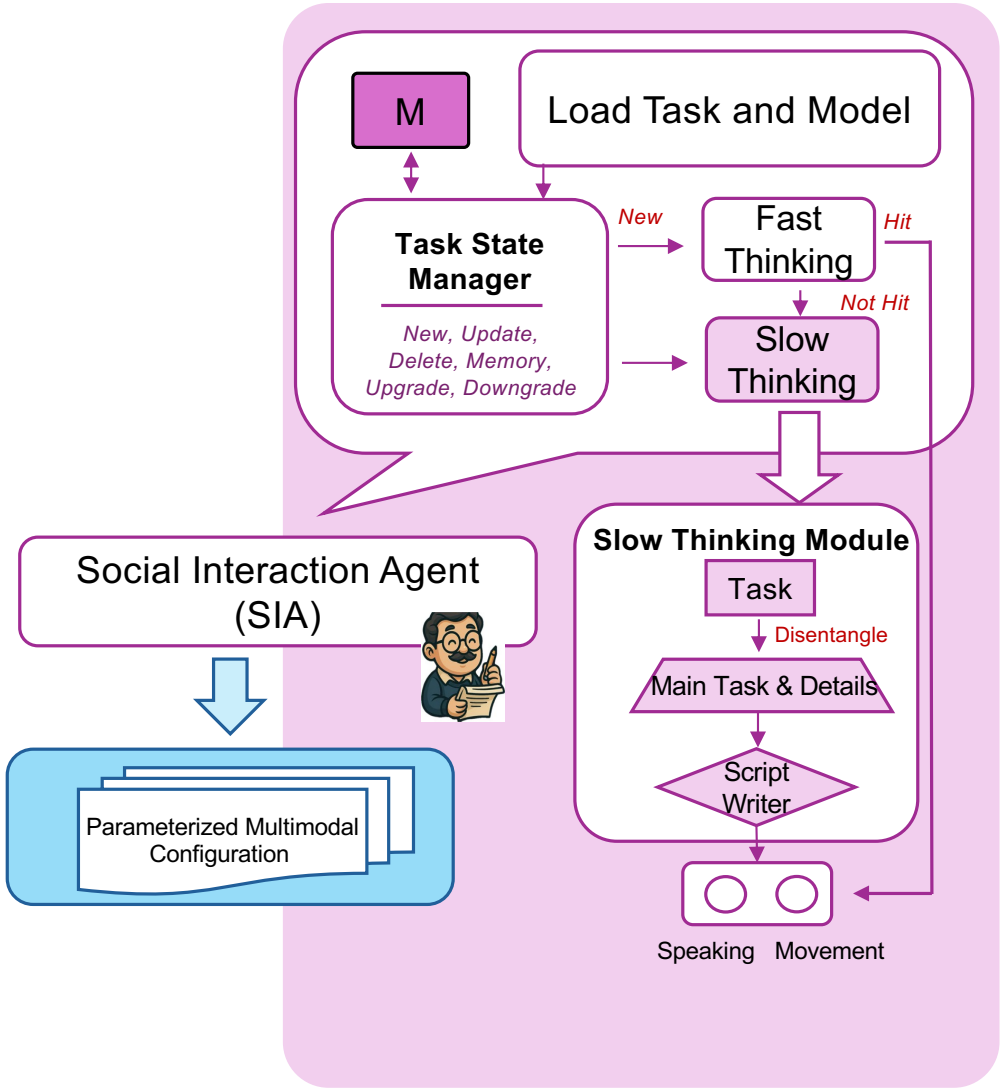
Persistent sensor bindings

One-time execution

An expandable skill library

Make newly generated skills available for invocation at runtime.

Method: Social Interaction Agent (SIA)



Emotionally expressive responses

Align emotional tone, intonation, and speaking rate with motion and facial expressions.

Fast path: reuse saved results

Reuse suitable results that users value and choose to save.

Slow path: generate new responses

Generate a new response when no suitable saved result is available.

Current task-state management

Maintain explicit task state with selective updates, rather than relying solely on dialogue history.

Technical Evaluation

Dataset Construction

Expert Seeds
Instructions from experts

→

LLM Expansion
Gemini 2.5 Pro + GPT-5
Rule-guided rewriting

→

Filtering & Sampling
ROUGE-L filtering
+ random sampling

→

Human Annotation
Human-assigned ground-truth labels

Dataset	Scale	Evaluation Focus
Route100	100 instructions	Routing accuracy between PIA and SIA
SensorBind40	40 instructions	Sensor binding and direct skill invocation accuracy
TaskParser256	40 dialogues; 256 turns	Task-state and system-control accuracy across turns
FastThinking230	46 instructions × 5 variants	Top-1 retrieval accuracy for paraphrased requests
SkillExtension100	30 / 50 / 70 / 100 skills	Skill selection from descriptions across library sizes

Technical Evaluation Results

Consistently high accuracy

With GPT-5-mini as its backbone, MistyPilot achieved over **95%** mean accuracy across all evaluated components.

Faster responses through result reuse

Core-task retrieval reuses user-saved results, reducing mean response time from **5.09 s** to **2.26 s**—a **55.5%** reduction.

Robust skill-library expansion

No degradation in skill-selection performance was observed as the library expanded from **30** to **100** skills.

Full experimental details, quantitative results, and discussion are provided in the paper.

User Study

Subjective evaluation results of the MistyPilot prototype.

Evaluation Dimension and Questionnaire Item	Mean	Std
Interaction Naturalness: Communicating with the robot through natural language to perform specific tasks was intuitive and straightforward.	4.75	0.45
Comprehension Accuracy: The robot accurately understood my intentions and provided responses that met my expectations.	4.42	0.67
Emotional Expressiveness: The robot's speech and reactions were emotionally expressive, giving it a highly anthropomorphic and human-like presence.	4.75	0.45
System Responsiveness: The overall response latency was within an acceptable range.	4.67	0.49
User Satisfaction: Overall, I am satisfied with the multimodal interaction experience provided by the MistyPilot system.	4.75	0.45

Note: All scores are on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). Mean is the average rating; Std is the standard deviation.

12 participants interacted with MistyPilot for **at least 30 minutes** each.

They then completed a **five-item questionnaire** on a **5-point Likert scale**.

All five mean ratings exceeded **4.4/5**.

MistyPilot Demonstration

PIA

Conclusion & Future Work

MistyPilot: natural-language skill orchestration

A multi-agent framework for coordinating physical and social interaction.

Initial evaluation completed

Component tests and a preliminary user study.

Next steps

Scale up experiments and user studies and continuously refine the framework.

Project page

wangxiaoshawn.github.io/MistyPilot.html

Project code and datasets are publicly available.

Q&A