異種マルチロボットを用いた潜在ニーズ推定に関する研究

A Study on Latent Need Estimation Using Heterogeneous Multi-Robot Systems

田代 大愛 [1*]　田中 文英 [2]
Taito Tashiro[1]　Fumihide Tanaka[2]

[1] 筑波大学 知能機能システム学位プログラム
[1] Master's Programs in Intelligent and Mechanical Interaction Systems, University of Tsukuba
[2] 筑波大学 システム情報系 知能機能工学域
[2] Institute of Systems and Information Engineering, University of Tsukuba

Abstract: Existing research on human needs estimation has primarily focused on predicting human behaviors, such as movement destinations and utterance timing. However, real-time multimodal information obtained from multiple robots, including visual information and dialogue history, has the potential to contribute to estimating latent needs that users do not explicitly express. Therefore, this study proposes a latent needs estimation method that integrates dialogue information between humans and conversational robots with nonverbal information of humans captured by cameras, using a vision-language model.

1　はじめに

医療・教育・介護など幅広い領域でサービスロボットの導入が進み，ロボットが人間の指示を理解して適切な支援行動を提供することが期待されている．一方で，人間は日常生活の中で支援を必要としていても，その要求を常に明示できるとは限らない．例えば，要求を適切に言語化できないことや自ら支援を要求することによる心理的抵抗感 [1]，そもそも必要性を自覚していない [2]，といった要因により，要求が明示的に表れない場合がある．本稿では，このように人間が明示的に表出しない要求を潜在ニーズと呼ぶ．潜在ニーズを適切に推定できれば，ロボットはユーザの状況に応じてプロアクティブな支援を実行できる．例えば，ユーザが自覚していない危険や不調の兆候を早期に提示し，事故や状態悪化を未然に防ぐことにつながる．さらに，このようなプロアクティブな支援は，人間の負担を低減しつつ状況に即した支援を提供することで，人とロボットの信頼関係やロボットの受容性に寄与する可能性がある [3]．

物体操作や発話タイミング予測などの意図推定 [4, 5] の場合は，人間の行動がその背後にある目的に基づいて選択されるため，視覚的な手がかりから意図を一定程度予測できるのに対し，潜在ニーズの推定では，その目的自体が暗黙的であり，人間は言語的・視覚的に手がかりを表出しないため，観測可能な視覚情報のみから推定することは困難である．したがって潜在ニーズの推定には，環境や行動の観測に加えて，人間の内部状態に関わる情報を引き出すための対話インターフェースを備えたマルチエージェントシステムが必要となる．近年，大規模言語モデル（Large Language Model: LLM）や視覚言語モデル（Vision-Language Model: VLM）の台頭により，視覚情報や自然言語を含む高レベルな指示から，複数のロボットの役割や能力を考慮したタスクの計画・割当・実行までを一貫して行うマルチロボットシステムの研究が急速に進展している [6, 7, 8, 9]．しかし，既存研究の多くはナビゲーションやマニピュレー

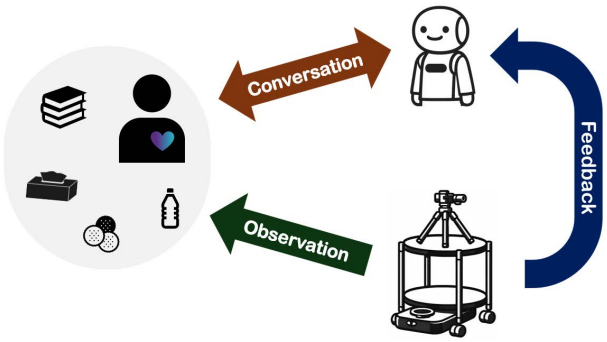

図 1: システムコンセプト

*連絡先：筑波大学 知能機能システム学位プログラム
〒 305-8573 茨城県つくば市天王台 1–1–1
E-mail: taito.tashiro@ftl.iit.tsukuba.ac.jp

ションといった物理的なタスクに焦点が当てられており，人とのインタラクションを前提としたものは限られている．

そこで本研究では，役割の異なる複数のロボットから得られるマルチモーダル情報に基づき，実環境下のユーザが明示的に表出しない潜在ニーズをリアルタイムに推定するマルチロボットシステムを提案する．具体的には，会話ロボットと移動ロボットの 2 種類のロボットを用いて，それぞれのロボットから得られる異なる視覚情報に基づいて，会話ロボットの質問内容に反映させることで，システムはユーザの潜在ニーズを効率的に探索することができる．

2 関連研究

2.1 意図推定

既存研究では，人間の行動は背後の目的に基づいて決定され，その意図が言語的・視覚的に観測可能な手がかりとして表れるという前提のもとで，次の行動を推定する枠組みが検討されてきた．環境情報に基づくアプローチとして，Patel ら [10] は，環境中の物体配置の時系列変化から物体の移動パターンをグラフニューラルネットワークで学習し，次の物体の移動先を予測することで先回り支援を行う手法を提案している．同様に，Mascaro ら [4] は，カメラ映像から人間と物体の相互作用を検出し，直前までの物体状態と人間の行動履歴に基づいて，次に人間が物体に対して行う動作を予測する研究を報告している．一方，人間の非言語行動に着目した研究も進められている．Russell ら [5] は，姿勢変化や顔向きなどの非言語行動から発話開始・終了のタイミングを予測する手法を示している．近年では，マルチモーダル情報を統合したアプローチも提案されている．Ali ら [11] は，発話や身振り，表情といった人間に関するマルチモーダル情報と環境状態を統合し，LLM を用いて次の行動意図を推論する枠組みを提案している．さらに，Wan ら [12] は，人間の指示生成までの過程をモデル化し，環境情報と人間の行動履歴から，指示の背後にある目的を推定する手法を提案している．

2.2 既存研究の課題と本研究の位置づけ

これらの既存研究では，物体操作や発話タイミング予測といった意図推定を扱っている．これらの場合，人間の行動はその背後にある目的に基づいて選択されるため，視覚的な手がかりから意図を一定程度予測できる．一方，潜在ニーズの推定では，その目的自体が暗黙的であり，人間は言語的・視覚的に手がかりを表出しないため，観測可能な視覚情報のみから推定することは困難である．例えば，要求の言語化困難や心理的抵抗感 [1]，あるいは必要性の非自覚 [2] といった要因により，ユーザが支援を必要としていても，その要求が行動として表れない場合がある．したがって潜在ニーズの推定には，環境や行動の観測に加えて，人間の内部状態に関わる情報を引き出すための対話インターフェースを備えたシステムが必要となる．

そこで本研究では，会話ロボットと移動ロボットからなる異種マルチロボットシステムを用いて，両ロボットから得られるマルチモーダル情報に基づき，実環境下のユーザが明示的に表出しない潜在ニーズをリアルタイムに推定するシステムを提案する．具体的には，移動ロボットが環境情報やユーザの全身行動を観測し，会話ロボットが対話を通じてユーザの内在的情報を探索することで，観測可能な手がかりが限定された状況においても効率的な推定を実現する．

3 提案手法

3.1 潜在ニーズ推定システムの概要

本研究では，会話ロボットと移動ロボットからなる異種マルチロボットシステムを用いて，ユーザが明示的に表出しない潜在ニーズを対話を通じて推定し，推定結果に基づく支援行動へ繋げるシステムを提案する．提案システムのコンセプトを図 1に示す．

会話ロボットはユーザと音声対話を行い，頭部に取り付けられたカメラからユーザの表情や視線などの局所的な視覚情報を取得する．一方，移動ロボットは底部および上部に取り付けられたカメラを用いて，環境に存在する物体群やユーザの服装や姿勢，所作といった大域的な視覚情報を取得する．両ロボットから得られる会話情報と視覚情報を統合して潜在ニーズを推定し，推定された対象に基づいて移動ロボットが探索や運搬といった行動を生成することで，ユーザへ直接的な支援を提供する．

3.2 潜在ニーズ推定のための問題設定

潜在ニーズは，視覚から推定しやすい外顕的な意図と比較して異なる性質を持つ．具体的には，曖昧で状況依存的であること，時間的に不安定で変化し得ること，多因子的で複雑であることが挙げられる．さらに，ユーザ本人が明示的に述べないため，教師あり学習において正解ラベルを一意に付与することが困難である．

そこで本研究では，視覚情報に加えて，ユーザの内在的情報を引き出すインターフェースとして対話を位置づける．すなわち，ロボット側から質問や提案を能動

的に行い，ユーザの反応を逐次観測することで，不確実性の高い潜在ニーズを段階的に絞り込む．また，ニーズが発散することを避けるため，ユーザに一定のコンテクストを付与する事前課題を導入し，その遂行過程の観測を通じて生起しやすいニーズの範囲を現実的に制約する．事前課題としては，計算課題や組立課題などを想定している．

加えて，本研究では潜在ニーズを環境に存在する物体に対する欲求に限定することで，推定を分類問題として定式化する．環境から検出された物体の集合を $\mathbf{O}$ とし，潜在ニーズ y をユーザが所望している物体 $y \in \mathbf{O}$ として表す．提案システムの目的は，視覚情報と対話履歴から得られるマルチモーダル情報に基づき y の事後分布を逐次更新し，限られた対話ターン内で最終的に最も確信の高い対象 $\hat{y}$ を推定することである．

3.3 視覚情報に基づく会話方策の生成

本研究の推定フェーズは，図2に示す3ステップから構成される．各ステップで得られる情報を統合し，会話ロボットが次に発すべき質問を決定する会話方策を生成する．

ステップ1：環境中のオブジェクトラベルの生成
移動ロボットは搭載カメラにより環境を撮影し，環境中に存在する物体候補を抽出する．物体の検出とラベル生成にはVLMを用い，OpenAI社が提供するGPT-4o[13]を利用する．VLMにより得られた物体ラベル列を

$$\mathbf{O} = \{O_1, O_2, \ldots, O_N\} \tag{1}$$

と表す．ここで N は物体ラベルの総数である．この物体ラベル列 $\mathbf{O}$ は会話ロボットの質問生成時に参照される．

ステップ2：コンテクストの理解
ユーザには所定時間の事前課題を行ってもらい，移動ロボットがその様子を観測する．観測データはVLMにより要約・解釈され，ユーザの状態や状況を示すコンテクスト情報 C として会話ロボットへ提供される．C は自然言語で記述され，事前課題の遂行状況やユーザの状態を反映した情報を含む．これにより，会話ロボットは無関係な質問を避けつつ，ユーザの現状に整合した話題選択や質問の導入を行うことができる．

ステップ3：視覚情報に基づく質問生成
事前課題終了後，ユーザは会話ロボットと音声対話を行う．この際，ユーザは明示的に欲しいものを表出しないことを前提とする．会話ロボットの目的は，コンテクストや質問に対するユーザの反応から，最も確信の高い対象 $\hat{y}$ を推定し，提案することである．

ユーザの潜在ニーズに関する信念状態をニーズ確信度として確率分布

$$\mathbf{P}^{(t)} = \{P_1^{(t)}, P_2^{(t)}, \ldots, P_N^{(t)}\}, \qquad \sum_{i=1}^{N} P_i^{(t)} = 1 \tag{2}$$

で表す．ここで $P_i^{(t)}$ は対話ターン t（$t \geq 1$）直後において，ユーザの潜在ニーズが候補物体 O_i である尤もらしさを表す．また，ニーズ確信度の初期値として，コンテクスト情報 C と物体ラベル列 $\mathbf{O}$ をLLMに入力し，コンテクストの内容に基づいて各候補物体に対する初期スコアを算出する：

$$\mathbf{P}^{(0)} = f_{\mathrm{LLM}}(C, \mathbf{O}). \tag{3}$$

対話開始時には，コンテクスト情報 C とニーズ確信度の初期値 $\mathbf{P}^{(0)}$ をLLMに入力し，初期質問 Q_1 を生成する：

$$Q_1 = f_{\mathrm{LLM}}(C, \mathbf{P}^{(0)}). \tag{4}$$

対話中，会話ロボットと移動ロボットはユーザに関する視覚情報を連続的に取得する．会話ロボットのカメラからは表情や視線などの局所的情報を，移動ロボットはユーザの服装や姿勢，所作などの全身情報を取得する．対話ターン t（$t \geq 1$）において，会話ロボットが質問 Q_t を発した直後を開始点，ユーザの応答が終わる瞬間を終了点として，両ロボットのカメラから数フレーム分の画像列を同期取得する：

$$\begin{aligned} \mathbf{V}_t^{\mathrm{global}} &= \{I_{t,1}^{\mathrm{global}}, I_{t,2}^{\mathrm{global}}, \ldots, I_{t,K}^{\mathrm{global}}\}, \\ \mathbf{V}_t^{\mathrm{local}} &= \{I_{t,1}^{\mathrm{local}}, I_{t,2}^{\mathrm{local}}, \ldots, I_{t,K}^{\mathrm{local}}\}, \\ \mathbf{V}_t &= \big(\mathbf{V}_t^{\mathrm{global}}, \mathbf{V}_t^{\mathrm{local}}\big). \end{aligned} \tag{5}$$

ここで K はフレーム数，$\mathbf{V}_t^{\mathrm{global}}$ は移動ロボット視点の画像列，$\mathbf{V}_t^{\mathrm{local}}$ は会話ロボット視点の画像列，$\mathbf{V}_t$ は二視点の画像列を統合した変数である．

対話ターン t（$t \geq 1$）において，ユーザの応答を含む対話履歴を H_t とする．各ターン終了ごとに，コンテクスト情報 C，これまでのニーズ確信度 $\mathbf{P}^{(t-1)}$，対話履歴 H_t，二視点の画像列 $\mathbf{V}_t$ をVLMに入力し，更新されたニーズ確信度 $\mathbf{P}^{(t)}$，次質問 Q_{t+1}，および提案フラグ F_t を得る：

$$\Big(\mathbf{P}^{(t)}, Q_{t+1}, F_t\Big) = f_{\mathrm{VLM}}\big(C, \mathbf{P}^{(t-1)}, H_t, \mathbf{V}_t\big). \tag{6}$$

提案フラグ $F_t \in \{0, 1\}$ は，特定の物体に対する確信度が十分に高まり，会話ロボットが直接的な提案を行うべきかどうかを示すトリガーである．$F_t = 1$ のとき，会話ロボットは「〜はいかがでしょうか」といった直接的な提案を行う．

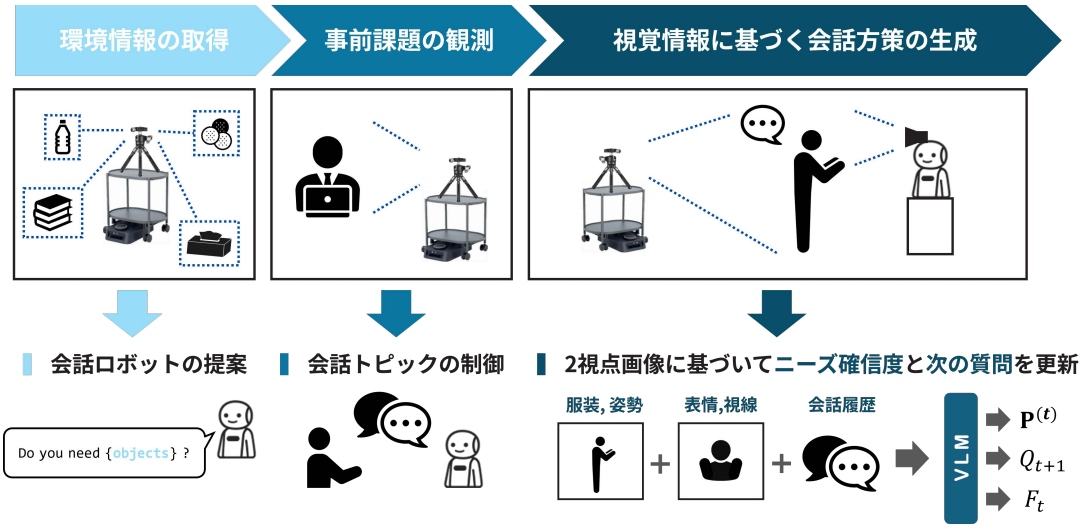

図 2: 潜在ニーズの推定ステップ

会話ロボットの質問戦略として，最初から直接的に「〜が欲しいですか」と尋ねるのではなく，特定の物体を意図した間接的な質問から対話を開始する．これは，ユーザ自身が自覚していない潜在ニーズに対処するためである．ユーザが必要性を自覚していない段階で直接的に物体を提案しても効果は限定的であるが，間接的な質問を段階的に繰り返すことで，ユーザ自身がその必要性に気づき，本来自覚していなかったニーズを認識できるようになることが期待できる．

具体的には，ある候補物体に対して，まずその物体が必要となる状況や用途について尋ね，ユーザの反応を観察する．式 (6) により各ターンでニーズ確信度が更新され，更新後の確信度分布において最大値を示す候補物体に関連する質問が次に選択される．ユーザの応答が肯定的であれば，その物体に関する質問をより具体的な内容へ深堀りを行い，否定的であれば別の候補物体へと探索を移行する．このように対話が進むにつれて確信度が特定の候補へ収束し，提案フラグ $F_t = 1$ となった時点で，最も確信度の高い対象 $\hat{y} = \arg\max_i P_i^{(T)}$ を直接的に提案する．同定された $\hat{y}$ は移動ロボットへ共有され，ユーザへの物体の運搬といった支援行動の生成に利用される．

以上のように，本提案手法は環境に存在する物体情報，事前課題の観測に基づくコンテクスト情報，異なるロボットからの二視点映像，会話ロボットとの対話履歴を統合することで，ユーザが明示的に開示しなくとも会話ロボットが能動的に候補物体を探索・提案できる潜在ニーズ推定システムを実現する．

4　ロボットシステムの実装

4.1　会話ロボットのための音声処理

会話ロボットは，ユーザとの音声対話を担当し，頭部カメラから取得した局所的視覚情報 $\mathbf{V}_t^{\text{local}}$ を移動ロボットから得られる大域的視覚情報 $\mathbf{V}_t^{\text{global}}$ と統合することで，式 (6) に基づく潜在ニーズの推定を実現する．

対話処理は，ユーザからの音声入力区間の検出，文字起こし，応答生成，音声合成の一連の処理により構成される．音声入力区間の検出には WebRTC Voice Activity Detection (VAD)[14] を用い，ユーザの発話開始と終了をリアルタイムに判定する．VAD により発話終了が検出されると，録音された音声データを対話システムへ送信する．

音声認識および音声合成には，低遅延な音声対話を実現するために Gemini Live API[15] を利用する．ただし，式 (6) で示したニーズ確信度の更新や次質問の生成には，第 3 節で述べた VLM を用いた推論処理が必要となるため，Gemini Live API による応答生成の前に，移動ロボットから得られる視覚情報と対話履歴を統合した推論を別途実行する．

頭部カメラからは，対話中のユーザの表情や視線方向といった局所的な視覚情報を取得する．式 (5) で定義した画像列 $\mathbf{V}_t^{\text{local}}$ は，会話ロボットが質問 Q_t を発した直後からユーザの応答が終了するまでの区間において，数フレーム分の画像を同期取得することで構成される．取得された画像列は移動ロボットから得られる $\mathbf{V}_t^{\text{global}}$ とともに VLM へ入力され，言語的応答と非言語的な視覚情報の整合性を考慮したニーズ確信度の更新に利用される．

4.2 移動ロボット

推定された潜在ニーズに対する物理的支援を実現するため，移動ロボットとして Preferred Robotics 社製の Kachaka Pro[16] を採用した．その外観を図 3(a) に示す．Kachaka Pro は自律移動機能を備えた家庭用サービスロボットであり，カチャカシェルフと呼ばれる専用棚の底部にドッキングすることで，図 3(b) のように棚ごと物品を運搬することができる．

移動ロボット制御には Kachaka API[17] を利用した．Kachaka API は目的地への移動指示，棚とのドッキング，現在位置の取得などの高レベル命令を提供する．本研究では，ROS2 ノードとして実装したコントローラが，式 (6) により同定された潜在ニーズ $\hat{y}$ に対応する支援行動の指示を Kachaka API の命令シーケンスへ変換して実行する．

(a) Kachaka Pro　　(b) カチャカシェルフ

図 3: Kachaka Pro とカチャカシェルフの外観

図 4は，移動ロボット底部カメラから取得した RGB 画像に対して MediaPipe Pose[18] を適用し，ユーザの全身姿勢を推定した例である．検出された関節キーポイントにより，頭部，肩，肘，手首，腰，膝，足首などの主要な身体部位の位置関係が抽出される．このような全身情報は $\mathbf{V}_t^{\text{global}}$ の一部として式 (6) における推論に利用され，大域的な身体状態を会話方策の生成に反映させることを可能にする．

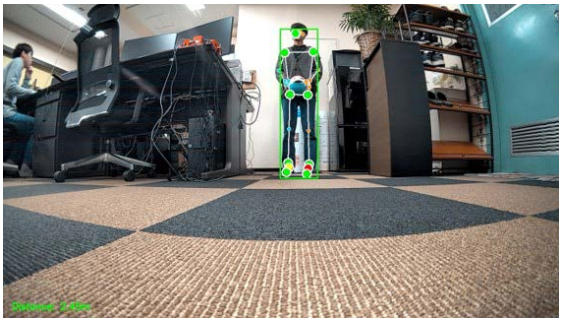

図 4: MediaPipe Pose による全身検出

5 おわりに

本研究では，会話ロボットと移動ロボットからなる異種マルチロボットシステムを用いて，ユーザが明示的に表出しない潜在ニーズをリアルタイムに推定する手法を提案した．提案手法は，各ロボットから得られる異なる視点の視覚情報と対話履歴を視覚言語モデルにより統合することで，環境中の物体に対する潜在ニーズを段階的に絞り込み，プロアクティブな支援行動を実現する．

今後の課題として，提案システムの有効性を適切に評価するための実験デザインの検討が挙げられる．潜在ニーズは人間自身が明示的に認識していない場合があるため，推定精度の評価指標や実験プロトコルの設計には慎重な検討が必要である．また，本研究では潜在ニーズを環境中の物体に対する欲求に限定したが，実際のニーズは物質的なものから精神的なものまで多様である．例えば，疲労や孤独感の緩和，励ましや共感といった情緒的支援への対応も重要であり，このような非物質的ニーズへの拡張が今後の展開として期待される．

謝辞

本研究は，JSPS 科研費 23H00484 の助成を受けて実施した．

参考文献

[1] Veronica Pavedahl, Åsa Muntlin, Ulrica Von Thiele Schwarz, Martina Summer Meranius, and Inger K. Holmström. Fundamental care in the emergency room: insights from patients with life-threatening conditions in the emergency room. *BMC Emergency Medicine*, Vol. 24, p. 217, 2024.

[2] L. H. Andrade, J. Alonso, Z. Mneimneh, J. E. Wells, A. Al-Hamzawi, G. Borges, et al. Barriers to mental health treatment: Results from the who world mental health (wmh) surveys. *Psychological Medicine*, Vol. 44, No. 6, pp. 1303–1317, 2013.

[3] Jimmy Baraglia, Maya Cakmak, Yukie Nagai, Rajesh P. N. Rao, and Minoru Asada. Initiative in robot assistance during collaborative task execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 67–74, 2016.

[4] Esteve Valls Mascaro, Daniel Sliwowski, and Dongheui Lee. HOI4ABOT: Human-object interaction anticipation for human intention reading collaborative roBOTs. In *7th Annual Conference on Robot Learning*, 2023.

[5] Sam O'Connor Russell and Naomi Harte. Visual cues enhance predictive turn-taking for two-party human interaction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 209–221, 2025.

[6] Zhao Mandi, Shreeya Jain, and Shuran Song. RoCo: Dialectic Multi-Robot Collaboration with Large Language Models. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 286–299, 2024.

[7] Bangguo Yu, Qihao Yuan, Kailai Li, Hamidreza Kasaei, and Ming Cao. Co-NavGPT: Multi-Robot Cooperative Visual Semantic Navigation Using Vision Language Models. *arXiv:2310.07937*, 2023.

[8] Pengying Wu, Yao Mu, Kangjie Zhou, Ji Ma, Junting Chen, and Chang Liu. CAMON: Cooperative agents for multi-object navigation with LLM-based conversations. In *RSS 2024 Workshop*, 2024.

[9] Junting Chen, Checheng Yu, Xunzhe Zhou, Tianqi Xu, Yao Mu, Mengkang Hu, Wenqi Shao, Yikai Wang, Guohao Li, and Lin Shao. EMOS: Embodiment-aware Heterogeneous Multi-robot Operating System with LLM Agents. In *International Conference on Learning Representations (ICLR)*, 2025.

[10] Maithili Patel and Sonia Chernova. Proactive robot assistance via spatio-temporal object modeling. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, Vol. 205, pp. 881–891, 2023.

[11] Hassan Ali, Philipp Allgeuer, and Stefan Wermter. Comparing apples to oranges: Llm-powered multimodal intention prediction in an object categorization task. In *International Conference on Social Robotics (ICSR)*, pp. 292–306, 2024.

[12] Yanming Wan, Yue Wu, Yiping Wang, Jiayuan Mao, and Natasha Jaques. Infer human's intentions before following natural language instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, pp. 25309–25317, 2025.

[13] OpenAI. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/, 2024. Accessed: 2025-02-01.

[14] Google. WebRTC: Real-Time Communication for the Web. https://webrtc.org/. Accessed: 2025-02-01.

[15] Google. Gemini Live API: Get started with Live API. https://ai.google.dev/gemini-api/docs/live. Accessed: 2025-02-01.

[16] Preferred Robotics. Kachaka Pro: 自律搬送ロボット. https://kachaka.life/. Accessed: 2025-02-01.

[17] Preferred Robotics. スマートファニチャープラットフォーム「カチャカ」API. https://github.com/pf-robotics/kachaka-api. Accessed: 2025-02-01.

[18] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A Framework for Building Perception Pipelines. 2019.