BotHub：複数の物理リモコンに対する意味理解と動的制御 UI 生成を行うフィジカル AI システム

三浦 晴大 * 渡邊 恵太 †

概要. ソフトウェア領域では AI による柔軟な対話が実現しているが，物理世界の家電操作は静的なインターフェースに縛られている．既存のスマートリモコンは信号を模倣するのみで起こる動作の意味を理解できず，手動設定が必須である．本研究は，AI が物理インターフェースの意味を自律的に理解し，ソフトウェアの対話的柔軟性を物理世界へ拡張することを目的とする．そこで，LLM がカメラ映像からリモコンの「機能的意味」と「物理座標」を即座に理解し，動的 UI を生成，さらに XY プロッタによる物理エージェントで操作を実行するシステムを提案する．本デモでは，複数デバイス（LED・ファン）による「雰囲気」の協調制御，および複雑なリモコン UI のタスクベースな最適化，という 2 つのユースケースを通じて，本提案を実証する．

1 はじめに

複数の家電や家屋設備を利用して暮らすことが一般化しており，ユーザーは複数のリモコンやスイッチの操作を要求される．これらを統合的に操作するため，学習リモコンや SwitchBot[2] など，家電を個人や家庭の要求に応じてカスタマイズする手法が存在する．例えば，アラームが鳴ったら部屋の電気やカーテンが制御されたり，GPS と組み合わせて家に近づいたら空調が on になるといったことが可能になる．また，スマートスピーカーなどのスマートデバイスは，障がい者支援や介護支援にも使われている．[3]

しかし，その統合や設計の難易度が非常に高いという課題がある．セットアップは複雑であり，一度設定した動作は状況や文脈を考慮できず，柔軟性に欠ける．さらに，新しい家電を追加する際の再設定もユーザーの大きな負担となる．多岐の機能を使いこなせなかったり，正しい使い方をしていなかったりする．このように，既存の手法は，多数の機能を持つデバイスを「統合する」という最も複雑な設計作業をユーザー自身に要求してしまっている．

また，AI はマルチモーダルな入力を得意としていて，物理世界と仮想世界の双方に適用可能である [1]．しかし，その柔軟な対話能力はソフトウェア領域での活用に留まっている．

そこで本研究では，LLM と XY プロッタを組み合わせ，所有する製品のリモコンを置くだけで，機能の「意味」を自律的に理解し，制御を実現するシステム「BotHub」を提案する．

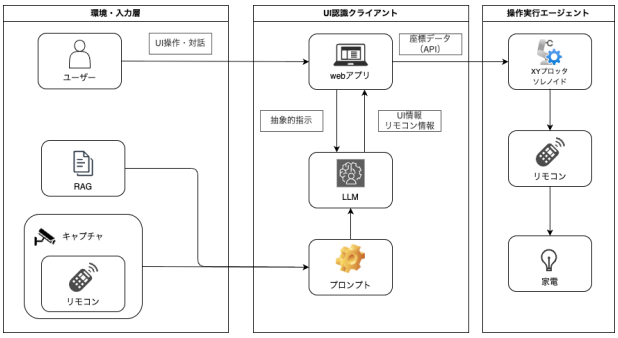

図 1. システムアーキテクチャ

本システムは，AI が各デバイスの機能を意味を理解することによって，統合された UI をユーザーへ自動的に提示する．さらに，ユーザーの抽象的な要望に対し，AI が複数の家電を柔軟に連携させ，統合的に制御することを可能にする．

本稿では，この概念を具現化したプロトタイプシステムを実装した．その有効性を，(1) 複数デバイス（LED とサーキュレーター）による「雰囲気」の協調制御，および (2) 複雑なリモコン UI の最適化，という 2 つのユースケースを通じて実証する．

2 提案手法

2.1 システムアーキテクチャ

本システムは，図 1 に示すように「環境・入力層」，「UI 認識クライアント」，「操作実行エージェント」の 3 コンポーネントで構成される．ハードウェアおよび web アプリは図 2 のようになっている．UI 認識クライアントは，入力（画像・テキスト）の取得，UI 描画，キャリブレーション，ユーザー操作の受

* 明治大学大学院先端数理科学研究科
† 明治大学総合数理学部

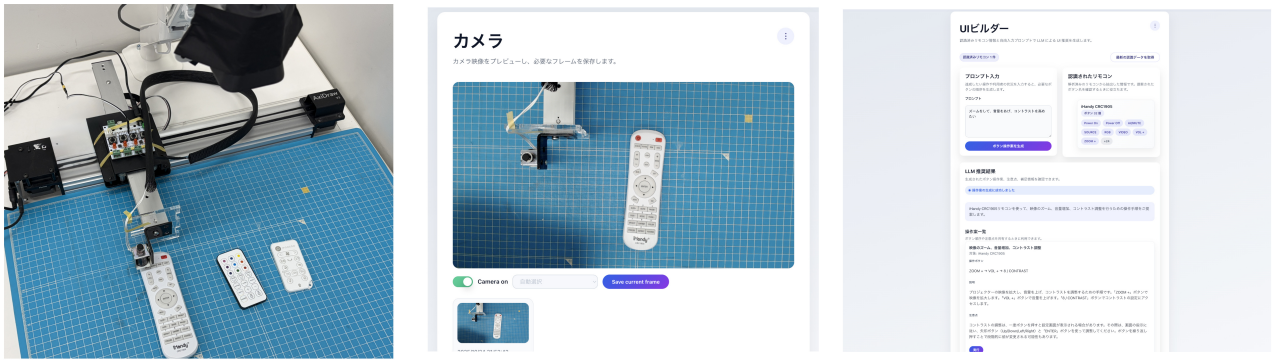

図 2. 左：ハードウェア機構, 中央：web アプリ（カメラ），右：web アプリ：UI 表示部分

付・送信を担う．AI サーバーは，LLM と通信し「意味・座標解析」と「動的 UI 生成」の推論処理を実行する．操作実行エージェントは，XY プロッタとソレノイドで構成され，サーバーからの座標指示に基づき物理ボタンを押下する．本システムの動作は，「キャリブレーション」,「解析と UI 生成」,「操作実行」の 3 プロセスで構成される．

2.2 キャリブレーション

カメラのピクセル座標とプロッタの物理座標を紐付けるため，OpenCV を用いた射影変換でキャリブレーションを行う．クライアント上でクリックした 4 点と，物理座標 4 点との対応から変換行列を計算・保存する．これにより，LLM の認識誤差は残存するが，幾何学的な誤差は補正される．

2.3 解析と UI 生成

2.3.1 UI の認識

クライアントは，カメラ画像とコンテキスト（取扱説明書など）を AI サーバーに送信する．LLM は，`gemini-2.5-pro` を用い，RAG のアプローチで両情報を総合的に判断し，各ボタンの機能とピクセル座標を JSON 形式で出力する．また，用いたプロンプトは下記である．プロンプトに加えて，Json スキーマを指定している．

```
"画像内のすべてのリモコンを検出してください．",
"ボタンの機能と，ボタンの中心のピクセル座標を正確に検出してください．",
```

ソースコード 1. LLM への指示プロンプト

2.3.2 UI の生成

AI サーバーは，得られた機能一覧とユーザーからの抽象的な指示を LLM で推論し，操作すべき機能を特定する．クライアントは，その推論結果に基づき，UI を動的に再構築・描画する．

2.4 操作実行

ユーザーがクライアント上の UI を操作，または AI が操作を決定すると，算出された物理座標に対し，操作実行エージェント（XY プロッタ）に送信され，ソレノイドが物理ボタンを押下する．

3 アプリケーション

本システムを用いて，2 つのアプリケーションシナリオを構築した．

第 1 は，抽象的な指示による複数デバイスの協調動作である．ワークスペースに多色 LED 電球のリモコンとサーキュレーターのリモコンを並べて配置し，ユーザーが望む「雰囲気」を LLM との対話によって創出する．

第 2 は，複雑な UI のタスクベース・リマッピングである．ボタンおよび機能が複雑なプロジェクターのリモコンと，その取扱説明書をシステムに入力し，LLM によってユーザーのタスク（例：「プレゼン」）に最適な UI を動的に再構築する．

4 考察と今後の展望

本デモでは，抽象的意図によるデバイス協調とタスクに基づく UI リマッピングの 2 点を実証した．これらは「インターフェイスの意味理解」によって実現され，既存手法がユーザーに要求していた複雑な統合設計を AI が自動化できる優位性を示した．

一方で，位置の認識誤差（キャリブレーションと LLM 精度への依存），解釈の依存性（LLM のハルシネーションリスク），静的な認識（デバイスの現状態の不認識）といった技術的課題が残る．

今後は，これらの課題解決を進めると共に，環境のセンシングによってデバイスの状態をコンテキストに加える．最終的な目標は,「はじめに」で述べたユーザーによる統合設計の負担をゼロにし，AI による柔軟な物理世界制御を実現することである．

参考文献

[1] Z. Durante, Q. Huang, N. Wake, R. Gong, J. S. Park, B. Sarkar, R. Taori, Y. Noda, D. Terzopoulos, Y. Choi, K. Ikeuchi, H. Vo, L. Fei-Fei, and J. Gao. Agent AI: Surveying the Horizons of Multimodal Interaction. 2024.

[2] SwitchBot. SwitchBot Bot(online). https://www.switchbot.jp/products/switchbot-bot, 10 2025.

[3] M. Tsurumi, M. Miyagi, and Y. Tamura. Survey on Use of Smart Speakers and Smart Home Devices by Visually Impaired People in Japan. *ICCHP Open Access Compendium–Future Perspectives of AT, eAccessibility and eInclusion–*, pp. 55–62, 9 2020.