行動認識に基づく実世界手続きタスク支援 AI：

センシングと適応的対話によるインタラクションフレームワーク PrISM

Riku Arakawa　Mayank Goel

Carnegie Mellon University

1　はじめに

大規模言語モデル（LLM）の急速な発展により，ソフトウェア上で計画・推論・行動を行う AI エージェントは目覚ましい進歩を遂げた。一方で，人間が物理世界で日々遂行する**手続きタスク**——複数の工程からなる料理，慣れない機器の操作，手術後の創傷ケアといった医療的セルフケアなど——を支援する AI は，依然として未解決の難問である。これらのタスクでは「重要な工程の抜け落ち」や「次に何をすべきかの失念」が頻繁に起こり，人間がそばで助言・介入するように [1]，状況に即した文脈認識（context awareness）[2] をもつデジタルアシスタントの価値は大きい。Amazon Alexa や Siri のような音声アシスタントは，音声入力という簡便な方法で質問できる利点をもつ一方，**利用者が今どのような状況にあるかを把握していない**ため，そもそも利用者は何を質問して良いか分からなかったり，ミスをしてしまっている時に能動的に支援できなかったりする。

こうした文脈に応じた能動的な支援を物理世界で実現することを本質的に難しくしているのが，**接地（grounding）**の問題である。ソフトウェアエージェントが扱う状態が API やテキストとして明示的に与えられるのに対し，物理タスクの支援 AI は，利用者の現在の状況（例えばどの工程にいるか）をセンサ信号から推定しなければならない。このセンシングには多様な選択肢がある。カメラは豊富な情報を与える一方で，オクルージョンや照明環境に依存し，設置の制約やプライバシーの懸念も伴う。マイクと慣性センサ（IMU）を備えたスマートウォッチや，LiDAR・mmWave レーダといった環境センサは，より偏在的でプライバシーにも配慮しやすい。いずれを選ぶかは利用者や環境次第であり，PrISM はこうしたマルチモーダルなセンシングに対応する。ただし，どのセンサであっても知覚は完璧ではなく，とくにカメラ以外のセンサに基づく人間行動認識（Human Activity Recognition; HAR）には，AI 上の二つの根本的な困難が伴う。

第一に，**データ効率の問題**である。インターネット規模のテキストで学習できる言語モデルと異なり，手続きタスクの行動データはタスクごと・利用者ごとに少量しか得られない。新しいタスクのために何十回もの実演データを収集することは非現実的であり，理想的には利用者が一度実演するだけで実用的な精度の知覚モデルを初期化できる必要がある。これは少数事例（one-shot/few-shot）学習の問題であり，物理世界への AI 展開におけるスケーラビリティの鍵を握る。

第二に，**不可避な不確実性下でのロバスト性の問題**である。微細で表現に乏しい動作（例えばカップを置く動作）や，工程間の曖昧な境界のため，非接触・非視覚のセンサによる知覚は原理的に不完全であり，フレーム単位の HAR の精度は複雑なタスクで 37–57% 程度にとどまる。誤った文脈理解

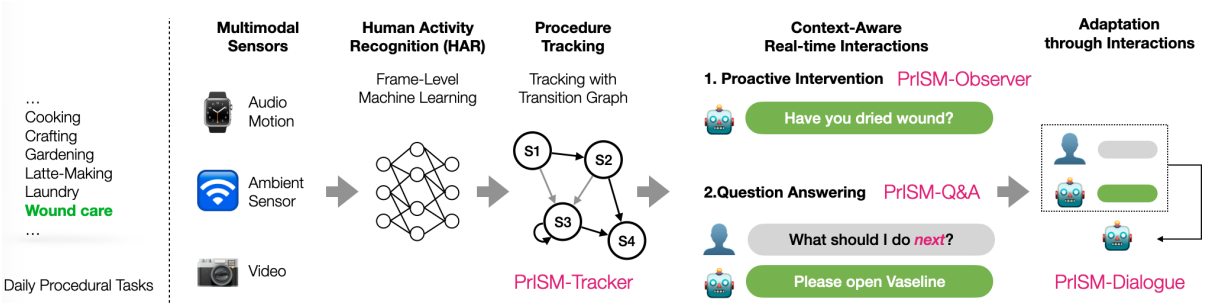


図 1: PrISM フレームワークの概念図。スマートウォッチ・環境センサ・カメラなどのマルチモーダルセンサを用いて，今利用者が何をしているのかを追跡する PrISM-Tracker [4]。その追跡したコンテキストに基づいて，ミスを防ぐためにタイミングよく能動的に介入支援を行う PrISM-Observer [5]。利用者の曖昧な質問に対しても，現在のコンテキストに基づいて適切な応答を提供する PrISM-Q&A [6]。そして，会話を通じて利用者と協調して文脈を適応させる PrISM-Dialogue [7]。

は，見当違いの応答や不適切なタイミングの通知といった「むしろ煩わしい」支援を生み，利用者の信頼を損なう。実際，本稿の Study #2 でも，受動条件では，制御された実験室環境ですら応答の 27% が誤認識に起因して不適切であったことが実証された。完璧な知覚を前提にできない以上，**支援 AI は不確実性のもとで推論し行動する能力をもたねばならない**。

これら二つの困難は，物理世界の AI がソフトウェアの AI と根本的に異なる理由を浮き彫りにする。言語モデルがインターネット規模のテキストから恩恵を受け，その入出力が明示的な記号として与えられるのに対し，物理タスクの支援 AI は，少量かつノイズの多いセンサ信号から世界の状態を推し量らねばならない。近年の AI とセンシングの進歩は，物理タスクを支援する知的アシスタントの可能性を大きく高めているが [3]，それが真に役立つためには，新規タスクへ低コストで展開できる（スケーラブル）だけでなく，配備後に**個別の環境・利用者・センシングの不確かさへ継続的に適応していく**ものでなければならない。本研究は，この要件を満たすための方法論を，アルゴリズムと支援設計の両面から提示する。

本稿が提案する PrISM（Procedural Interaction from Sensing Module）は，これらの課題に取り組むフレームワークである (図 1 参照)。PrISM は，マルチモーダルセンシングに基づく四つの貢献——ステップ追跡，文脈認識の質問応答，先回りの観察・介入，そして対話によるオンライン適応——を，**追跡 (tracking) → 支援 (interactions) → 適応 (adaptation)** という一貫した三層の流れとして統合する。

- **追跡（Tracking）**：マルチモーダルなセンサ信号から，利用者が手続きのどの工程にいるかを推定する。状態遷移グラフと拡張ビタビアルゴリズムにより，ノイズの多い HAR 出力を手続きの構造で補正し，一度の実演からの効率的な学習を可能にする（第 4 節）。
- **支援（Interactions）**：追跡された文脈を LLM に接続し，受動的な**質問応答**（利用者の曖昧な問いに文脈で答える）と，不確かな観測中で利用者のミスを防ぐための能動的な**介入**（重要工程を適切な時点で先回りして促す）という二様式の支援を実現する（第 5 節）。
- **適応（Adaptation）**：上記の支援を双方向に捉え直し，混合主導対話から得られる手がかりで追跡器の文脈理解をリアルタイムに更新する。これにより，不完全なセンシングを個々の状況へ適応させながら信頼性高く運用する（第 6 節）。

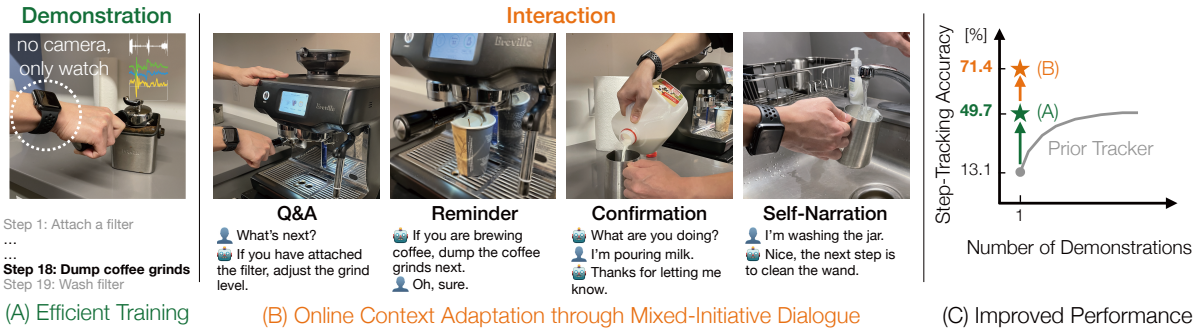


図 2: 本稿のコアである一度のデモンストレーションからタスクを学習し，混合主導対話によるオンライン文脈適応を実現するフレームワークの概念図。(A) マルチモーダルセンサとして，例えばスマートウォッチを用い，一度の実演 (例えばラテ作成のタスク) から手続きの「ステップ追跡器」を効率的に学習する。(B) 学習した追跡器は不完全だが，混合主導対話ー質問応答 (Q&A)・リマインド (Reminder)・確認 (Confirmation)・自己ナレーション (Self-Narration) を通じてオンラインで文脈理解を適応させる。(C) こうした拡張により，追加のデータ収集やモデルの再学習なしに追跡精度と支援の質が向上する（図中の数値は 19 ステップからなるラテ作成タスクにおける代表的な結果を示す。本文では各実験条件の精密値を報告する）。

PrISM のコア主張には，物理世界の手続き支援 AI の鍵が，知覚の精度を限界まで高めることではなく，人間と AI の自然な対話を，リアルタイムの適応信号として活用することにある，という点に集約される。すなわち，限られたデータから得た不完全な追跡を初期推定として受け入れたうえで，質問応答・利用者の自己ナレーション・システムからの確認やリマインドといった日常的な対話のやり取りから文脈を抽出し，知覚モデルの内部状態をオンラインで更新する。これは，モデルの再学習を要しない軽量なテスト時・人間協調型（human-in-the-loop）適応である。本フレームワークの全体像を図 2 に示す。

本稿は，「不完全なセンシングを，対話を通じた適応によって信頼性高く運用する」という一つの設計原理のもとに，物理世界へ AI を展開する際の方法論と実証的根拠を提示する。広範な物理タスクへの応用が想定されるが、著者らは特に本フレームワークを，**皮膚がんの手術後に患者が在宅で行う創傷セルフケアの支援や，認知症の患者の在宅での日常生活支援**として，臨床現場および患者の自宅で実際に展開している [8, 9]（第 9 節 で詳述）。創傷ケアは工程の抜けや誤りが感染や治癒遅延に直結する一方，患者は医療者の常時監督なしに手続きを正確に遂行せねばならず，まさに「不確実性のもとでと人間と AI が共存しながら安全な支援をいかに成立させるか」という本稿の問いが，文字どおり患者の予後を左右する。本稿では，この医療応用を貫く動機として位置づけたうえで，知能情報処理技術の実社会応用としての意義を論じる。

2　関連研究と本研究の位置づけ

本研究は，人間行動認識，文脈認識タスク支援，およびリアルタイムの人間協調（human-in-the-loop）システムという三つの研究領域の交点に位置する。

2.1　人間行動認識（HAR）とその限界

タスク支援には，利用者の現在の状況（どの段階にいるか）という動的文脈の把握が不可欠である。これを実現するセンサには，カメラ，ウェアラブル，環境センサなど多様な選択肢があり，PrISM は

そのいずれにも対応しうる。カメラは豊富な情報を与え，エゴセントリックな支援データセットも整備されつつある [1, 10]。一方でカメラはオクルージョンや照明環境に依存し，常に正確に認識できるわけではない。加えて，設置の制約やプライバシーへの懸念から，環境にカメラを置きたくない利用者も多い。特に，自宅での医療ケアの支援においては，カメラの設置が困難な場合が多く，従来のHAR 手法が適用しづらい。そうした場合には，スマートウォッチや，LiDAR・ドップラーレーダといった環境センサに基づく偏在的な行動認識（HAR）が用いられる [11]。マイク [12]，IMU，ドップラーレーダ [13]，2D LiDAR [14] など多様なセンサやそれらの組合せ [15] が研究されているが，カメラを避けるほど，また**実世界の複雑なタスクに適用するほど，知覚精度は下がらざるをえない**。手続きタスクには微弱な信号しか生まない工程が含まれ，センサで検出しづらいためである [16]。フレーム単位 HAR の後処理としてタスクの遷移情報を組み込むことで精度向上を図る試みもあるが，依然としておよそ半数の誤りが報告されており [17]，しかも多くは新規タスクごとに多数の実演を要して展開コストが高い。PrISM は，どのセンサ構成でも避けがたいこの不完全さを前提に，一度の実演からの学習と対話による適応で実用化を図る。

2.2　文脈認識タスク支援とのインタラクション

Human-Computer Interaction (HCI) 分野では，電子工作 [18]，調理 [19–22]，組立 [23, 24]，機器操作 [10, 25] など多様な物理タスク支援のためのインタラクションが長く研究されてきた。なかでも音声インタラクションは，スマートウォッチやスマートスピーカなど多様なプラットフォームで利用でき，LLM の発展も相まって広く普及している。質問応答（Q&A）は複雑なタスクに従事する利用者にとって主要な支援形態であり，Wizard-of-Oz 法を用いた研究は，手続きタスクでは「次に何をすべきか」といったステップ関連の問いが多いことを報告している [26, 27]。ここで効果的な対話には文脈の共有が不可欠だが，利用者はしばしば問いを明確に言語化できず，代名詞のような言語モデルには解釈の難しい参照に頼る。そのため，センシングによって得た文脈を対話に組み込むことが重要となる。また，先回りの介入も誤りを防ぐ重要な支援形態であり，ウェアラブルカメラによる服薬エラー検出 [28] や，スマートウォッチによる逐次誘導 [12] が提案されてきた。しかし，質問応答と先回り介入のような複数の支援様式を統合し，文脈理解に不確実性があるなかで利用者を効果的に支援する方法は，ほとんど探究されてこなかった。多くの研究がセンシングの誤りに直面しており，これは「**もしセンシングが完璧なら」という仮想的に理想的な支援を Wizard-of-Oz 法で探る先行研究 [27, 29] の限界を浮き彫りにする**。本研究は，支援は相互的（reciprocal）であるべき——**アシスタントは利用者からの動的なフィードバック（それが対話のなかの自発的なものであっても）を通じて文脈理解を更新すべき**——だと主張し，そのオンライン文脈適応機構を導入して有効性を示す。

2.3　リアルタイムの人間協調システム

人間協調（human-in-the-loop）システムは，自動処理の実行に人間が能動的に参加し，フィードバックや調整，必要に応じた介入によって性能や結果を改善するものである [30]。例えば，ロボットが操作者と対話して曖昧性を解消しながら物体を把持する研究がある [31]。その有効性は広く示されてきたが，多くは明示的で専用のフィードバック手続きに依存しており，それは利用者の負担となり，実世界では十分なデータが得られないことも指摘されている [32]。これに対し，システムの振る舞いへのリアルタイムの反応のような，自然だが構造化されていない信号の利用が探究されている

が，こうした暗黙的フィードバックは正確な捕捉が難しい。本研究は，**構造化されていない対話のやり取りを，タスク支援アシスタントの構造化されたステップ追跡状態とリアルタイムに結びつける計算論的アプローチ**を提案する。

3　PrISM フレームワークの全体像

図 1 に PrISM の全体構成を示す。マルチモーダルなセンサ信号（例えばスマートウォッチの音響・運動）をステップ追跡器が処理して利用者の現在の状況（どの工程にいるか）を推定し，その文脈を LLM ベースの対話モジュールに渡す。対話のやり取りから得られた手がかりは追跡器に戻され（オンライン適応），文脈理解を継続的に補正する。

この構成は，センサを用いた行動認識に基づくタスク支援のスケーラビリティという課題に応えるために設計されている。**スケーラビリティを阻む二つの要因——(1) 新規タスクごとの高い学習コストと，(2) 不完全なセンシングに起因する誤った支援——に対し，PrISM は二つの中核機構で応える：(A) 一度の実演から追跡器を効率的に学習する方法（図 2A，第 4 節）と，(B) 混合主導対話を通じて文脈理解をリアルタイムで更新するオンライン適応機構（図 2B，第 6 節）である**。各モジュールは疎結合に設計されており，HAR モデルやセンサ・提示デバイス（カメラ・スマートウォッチ・環境センサなど）を差し替えても枠組みが成り立つ。

以降の各節では，この流れを構成する三層を順に詳述する。本稿の方針として，中核である適応（第 6 節）の評価を詳細に述べる一方，追跡・支援の各研究については枠組みの理解に必要な範囲で技術と代表的成果のみを簡潔に示す。詳細な実装と評価は各論文 [4–6] を参照のこと。また実装はオープンソースとして公開されている (`https://github.com/cmusmashlab/prism`)。

4　追跡：手続きのステップ推定

手続き支援の出発点は，利用者が手続きのどの工程（ステップ）にいるかを推定することである。最大の障害は，前述のとおり，どのセンサを用いても HAR の精度が限られる点にある。以下では一例としてスマートウォッチ（音響・運動）を用いて説明するが，PrISM の追跡はカメラや環境センサにも同様に適用できる。微細な動作や工程境界の曖昧さのため，フレーム単位の HAR を手続き追跡へ素朴に適用すると精度に問題が生じる。PrISM の追跡器（PrISM-Tracker [4]）は，この生の HAR 出力を手続きの構造的なタスクの事前知識で補正することでこの問題に対処する。手続きには「ある工程の後には特定の工程が続きやすい」「各工程にはおおよその所要時間がある」といった順序・時間の制約があり，これらを用いれば，フレーム単位では曖昧な予測を大域的に絞り込める。

4.1　状態遷移グラフと拡張ビタビ法

図 3 に追跡アルゴリズムの概要を示す。タスクが S 個のステップからなり時系列が T フレームのとき，PrISM の追跡器は $S \times T$ 個の隠れ状態 $h(i,\tau)$ を用意する。ここで $h(i,\tau)$ は「時刻 t において利用者が第 i ステップに少なくとも τ フレーム留まっている」ことを表す。各時刻で，フレーム単位の HAR 出力 $\boldsymbol{p}_t^{har} \in \mathbb{R}^S$ と，ステップ間の遷移確率 $P(j \to i)$ および各ステップの滞在時間分布（平均・標準偏差）を保持する遷移グラフ G とを用いて，全隠れ状態の確率 $\boldsymbol{p}_t^{hidden} = \{p_t(i,\tau)\}$ を逐次更新する。具体的には，滞在時間分布から「第 i ステップに τ フレーム留まった後さらに 1 フ

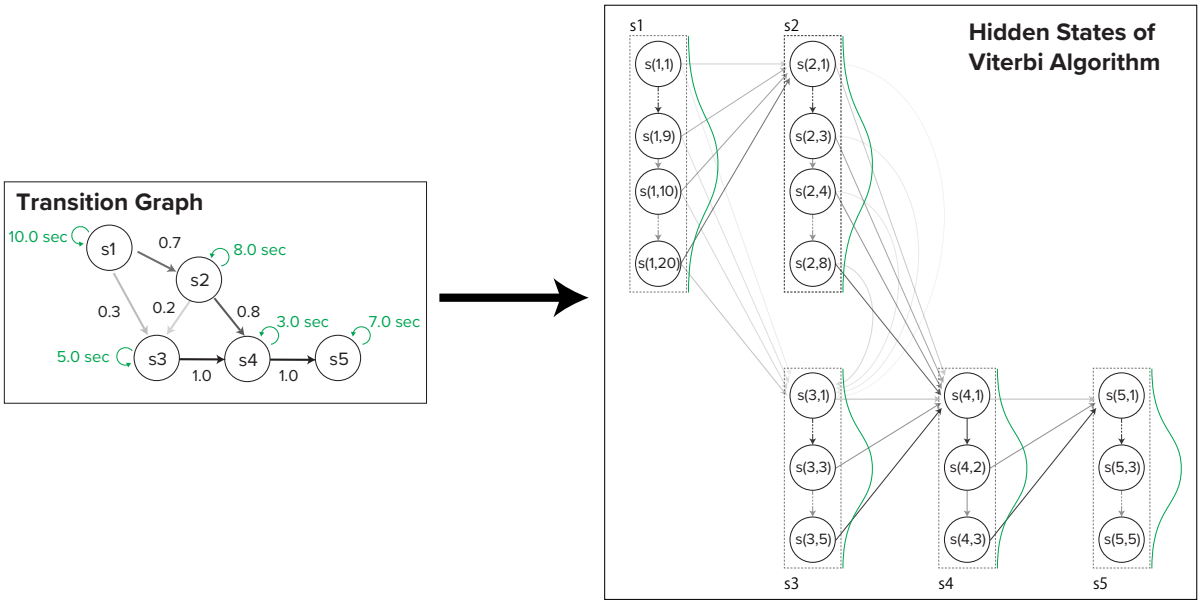


図 3: 拡張ビタビ法による追跡。各隠れ状態 $h(i,\tau)$ は「第 i ステップに少なくとも τ フレーム留まっている」ことを表す。フレーム単位の HAR 出力 $\boldsymbol{p}^{har}$ を，状態遷移グラフ G（遷移確率と滞在時間分布）と組み合わせて前向きに更新することで，ノイズの多い瞬間的な予測を手続きの構造で補正する。

レーム継続する確率」$c_i(\tau)$ を定め，前向き再帰

$$\tilde{p}_{t+1}(i,1) = \Big[\textstyle\sum_{j\neq i}\sum_{\tau'} p_t(j,\tau')\,\big(1-c_j(\tau')\big)\,P(j\to i)\Big]\,p^{har}_{t+1}(i), \tag{1}$$

$$\tilde{p}_{t+1}(i,\tau) = p_t(i,\tau-1)\,c_i(\tau-1)\,p^{har}_{t+1}(i) \quad (\tau>1), \tag{2}$$

により更新し，$\boldsymbol{p}^{hidden}_{t+1} = \tilde{\boldsymbol{p}}_{t+1}/\sum\tilde{\boldsymbol{p}}_{t+1}$ と正規化する。第 1 式は新たに第 i ステップへ遷移してきた成分，第 2 式は同一ステップ内で滞在を続ける成分を表す。この定式化は，状態の滞在時間を明示的に確率分布でモデル化する隠れセミマルコフモデル（HSMM）[33] の考え方に基づき，「あるステップに長く留まるほど次へ遷移しやすい」という手続き特有の事前知識を，ノイズの多い HAR 出力 $\boldsymbol{p}^{har}$ の補正に組み込む。計算量は $O(T^2(S+E))$（E は遷移グラフの辺数）であり，リアルタイム推論が可能である。**このマルチモーダル追跡器そのものが PrISM の第一の貢献 [4] であり，さらに本稿は，それを一度の実演から立ち上げられるよう拡張する（第 4.2 節）**。従来この種の追跡器は，遷移グラフの構築に複数回・複数人の実演を要したが，後述する信号フィルタリングと遷移グラフの自動生成により，単一の実演から実用的な精度で初期化できるようにし，新規タスクへの展開コストを大きく下げる。なお，フレーム単位の HAR 出力 $\boldsymbol{p}^{har}$ は，一般の HAR データセットで事前学習した畳み込みニューラルネットワークで特徴量を抽出し（スマートウォッチの場合は音響・運動），その上の分類器が各ステップの確率を出力することで得る [15]。複数のモダリティを融合することで，一方では捉えにくい工程をもう一方が補う。

4.2　一度の実演からの効率的な学習

ステップ追跡を素朴に学習するには，遷移グラフ G の構築に複数回・複数人の実演を要し，新規タスクへの展開コストが高い。さらに，この追跡器を一度の実演から学習できるよう，二つの工夫を導入する（図 2A）。

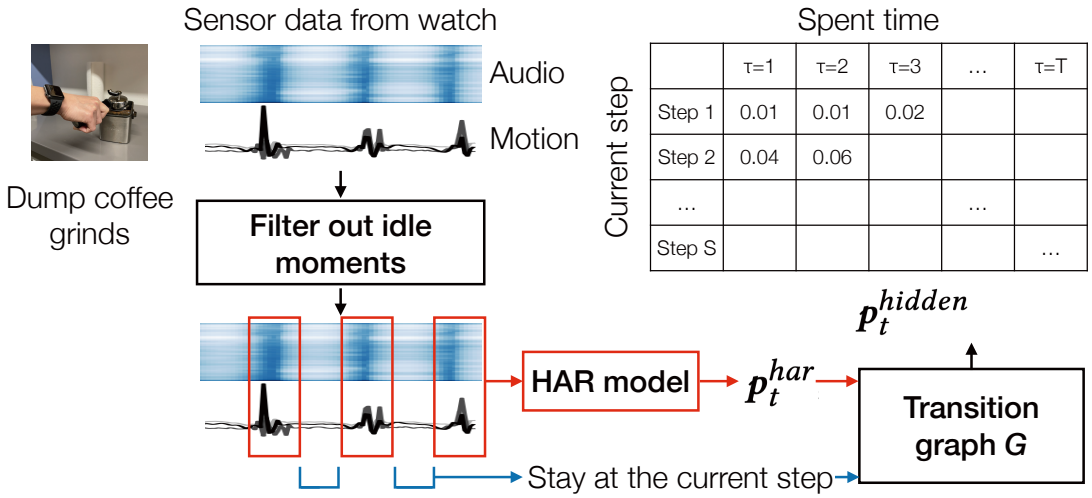


図 4: 学習効率化のための信号フィルタリング。教師なし異常検知により「行動に関連する重要な瞬間」と「待機・無関係な瞬間」を区別し，重要フレームにのみ HAR モデルを適用する。これにより少量の学習データでも頑健な追跡を実現する。

(1) 待機状態のフィルタリング（図 4）。手続き上の能動的な動作はセンサ信号を変化させるのに対し，待機や無関係な動作は信号が乏しい。そこで，逐次・教師なしの異常検知 [34] により各フレームを「重要／非重要」に分類し，重要フレームにのみ HAR を適用して隠れ状態を更新する。非重要フレームでは現在のステップ継続とみなす（$p_{t+1}(i,\tau) = p_t(i,\tau-1)$）。これにより，ほぼランダムな HAR 出力が追跡を混乱させることを防ぐ。このフィルタリングは追跡だけでなく HAR モデルの学習にも寄与する。異なるステップ間で似通った信号をもつ待機フレームを多く含めると誤分類を招きやすいが，それらを除くことで，少量の学習データからでも識別性の高いモデルが得られる。実際には，タスクごとに閾値 β を経験的に定め，学習データのおよそ 10% を非重要と判定した。

(2) 遷移グラフの自動生成。遷移グラフ G のパラメータは，(a) ステップ間の遷移確率と，(b) 各ステップの滞在時間の平均・標準偏差からなる。このうち遷移確率は手続きの論理的な順序関係を反映するため，LLM で推定できると考え，LLM に妥当なステップ順序を K 回（K=20）生成させ，その頻度を遷移確率の代理として用いる。滞在時間は実演中の実測値を平均，その α%（α=10）を標準偏差とする単純な経験則を採る。素朴な手法ながら，学習サンプルが少ない状況でとくに有効に働く。これは，従来の PrISM-Tracker が，グラフ G を構築するための実演データの不足によって性能を制約されていたことの裏返しでもある。

これらにより，後述する Study #1 では一度の実演のみで，従来法に対し平均約 17.6 ポイントのフレーム精度向上を達成した（第 7.1 節）。ただし重要なのは，こうして得られる追跡器は**なお不完全であり，その絶対精度は手続き追跡の本質的難しさを反映して低い**点である。PrISM が採るのは，この不完全な追跡器を完成形ではなく**出発点となる暫定的な追跡とみなし，対話を通じて継続的に補っていく**という立場である。

5　支援：受動的な質問応答と能動的な観察・介入

追跡された文脈は，二様式の支援を通じて利用者に価値をもたらす。一つは利用者の問いに答える**受動的**な質問応答（PrISM-Q&A [6]），もう一つはシステムが先回りして働きかける**能動的**な観察・

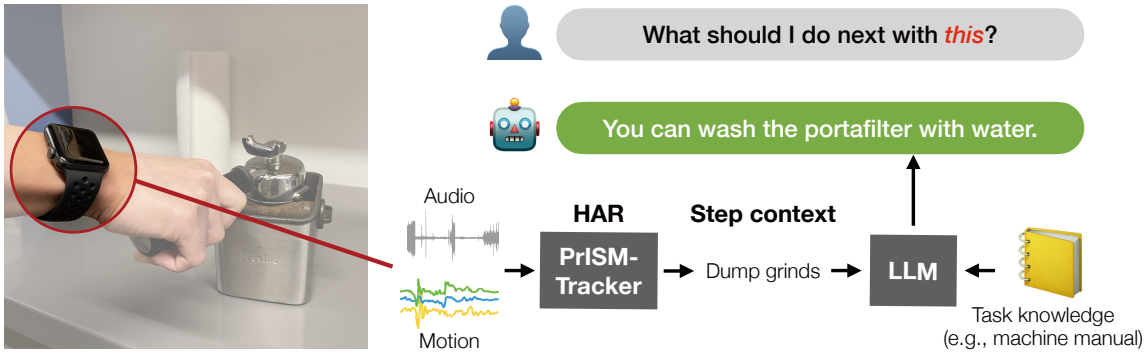


図 5: PrISM-Q&A の概要。手を使う手続きタスク中の利用者が，道具の名前が分からず代名詞を用いて「これは次どうする？」のように尋ねても，PrISM-Tracker が推定した工程文脈を LLM に与えることで，曖昧な問いを解決し，状況に即した応答を返す。手続きタスク支援に特化した，ステップ認識型の音声質問応答インタラクションである。

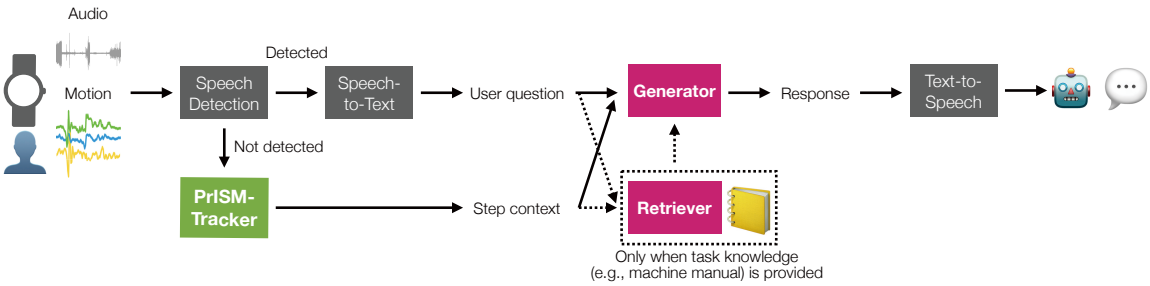


図 6: PrISM-Q&A の処理パイプライン。スマートウォッチの音響・運動信号を PrISM-Tracker が処理して工程文脈を推定し，その情報を LLM のプロンプトに注入することで，「次は何をすればよい？」のような曖昧で文脈依存の問いに正確に応答する。

介入（PrISM-Observer [5]）である。以下では一例としてスマートウォッチ上の音声支援を述べるが，提示デバイスやセンサも差し替え可能である。両者は補完的である。質問応答は利用者が自ら気づいた疑問に応える一方，能動的介入は利用者が気づかない失念や安全上の誤りを未然に防ぐ。いずれも同じステップ追跡器の出力を文脈として消費しており，したがって追跡の誤りはどちらの支援にも誤りとして波及する。この事実が，次節（第 6 節）で論じる適応の動機を直接に与える。

5.1 文脈に応じた質問応答（PrISM-Q&A）

手を使う作業中の利用者は，作業を中断して状況を言葉で説明する余裕がなく，「次は何をすればいい？」「これはもうやった？」といった**曖昧で文脈依存の問いを発しがち**である。通常の音声アシスタントはこうした問いを解決できない。PrISM-Q&A [6] は，PrISM-Tracker が推定した工程文脈を LLM に与えることでこれを解決する（全体像を図 5 に，処理パイプラインを図 6 に示す）。具体的には，現在の推定ステップ（確信度に応じて「おそらく X」「X かもしれない」と表現を変える），遷移グラフから推定される次の候補ステップ，および曖昧性を文脈で解決するよう促す指示をプロンプトへ注入する。これにより，利用者の問いに含まれる代名詞や省略を，センシングされた文脈で補って解釈できる。

技術的には，二つの構成要素が文脈を注入する。第一は**ステップ認識リトリーバ**で，知識源（例：機器のマニュアル）から関連箇所を検索する前に，利用者の問いに推定ステップを付加する。例えば

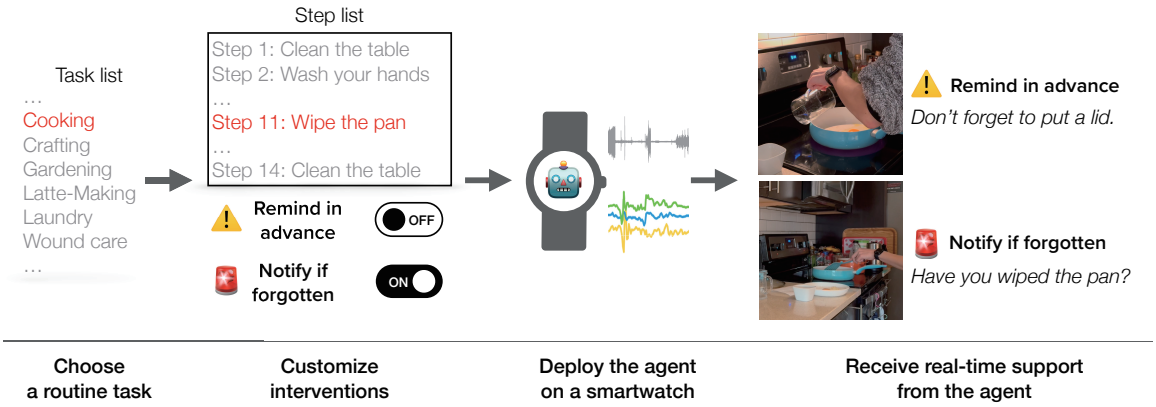


図 7: PrISM-Observer の概要。料理などの日常的な手続きタスクにおける利用者の誤りを軽減するために，介入を設計・発火するフレームワークである。PrISM-Tracker が推定した行動の確率モデルから介入の好機を予測する。重要な工程を最適なタイミングで先回りリマインドし，工程を失念した場合にはリアルタイムで通知する。同じデバイスを介入の提示にも用いることができる。

「次は何をすればいい？」という問いを「次は何をすればいい？　いま手を洗っている」と補ってから埋め込み検索を行うことで，検索の精度が高まる。第二は**ステップ認識ジェネレータ**で，現在ステップ（確率に応じた表現の調整つき）と次候補ステップを LLM のプロンプトに与え，曖昧性を文脈で解決させる。リアルタイムシステムでは，利用者が発話すると，端末上の音声認識で文字化し，上記の文脈付きで応答を生成し，スマートウォッチ（あるいはイヤホン）で読み上げる，という一連のパイプラインが動作する。応答は，まず「コーヒーを淹れ終えているなら，次は…」のように**システムの文脈理解を明示してから本題に入る形をとり，利用者が誤解を即座に訂正できる透明性**をもたせている。この透明性は，後述する適応機構と相補的に働く。すなわち，システムが自らの文脈理解を口に出すことで，利用者の訂正（次節の適応の入力信号）を誘発しやすくなる。

5.2　先回りした観察・介入（PrISM-Observer）

質問応答が利用者の能動性に依存するのに対し，重要な工程の失念（とりわけ安全に関わる手順）は，システム側が先回りして防ぐべきである。PrISM-Observer [5] は，「いつ介入すべきか」を確率モデルで決定する（全体像を図 7 に，介入タイミングの決定を図 8 に示す）。時刻 t において目標ステップ $\hat{s}$ までの残り時間を確率変数 $D_t^{\hat{s}}$ とみなし，その期待値 $E[D_t^{\hat{s}}]$ とエントロピー $H[D_t^{\hat{s}}]$ を

$$E[D_t^{\hat{s}}] = \sum_{s \in S} P(s) \sum_{\tau \in T_s^{\hat{s}}} P(\tau)\, \mathrm{Time}(s \to \hat{s}, \tau), \tag{3}$$

$$H[D_t^{\hat{s}}] = -\sum P(D_t^{\hat{s}}) \log P(D_t^{\hat{s}}) \tag{4}$$

として求める。ここで $T_s^{\hat{s}}$ は現在ステップ s から $\hat{s}$ へ至る経路の集合，$P(s)$ は時刻 t に利用者が s にいる確率（追跡器の出力），$P(\tau)$ は経路 τ を辿る確率，$\mathrm{Time}(s \to \hat{s}, \tau)$ は経路 τ に沿った平均所要時間である（$P(\tau)$ と Time は遷移グラフ G から計算される）。$E[D_t^{\hat{s}}]$ の二重和は，**現在状態の不確実性**（$P(s)$）と**将来経路の不確実性**（$P(\tau)$）という二種類の不確実性を同時に織り込んでいる点が要諦である。経路の列挙は組合せ的に増えるため，これらはモンテカルロ法（10,000 サンプル）でリアルタイム推定する。エントロピー $H[D_t^{\hat{s}}]$ が十分小さく（ステップ依存の閾値 $h^{\hat{s}}$ を下回り）期待値が安定した時点を「介入の好機」と捉え，目的に応じて二種類の介入を発火する：重要工程の手

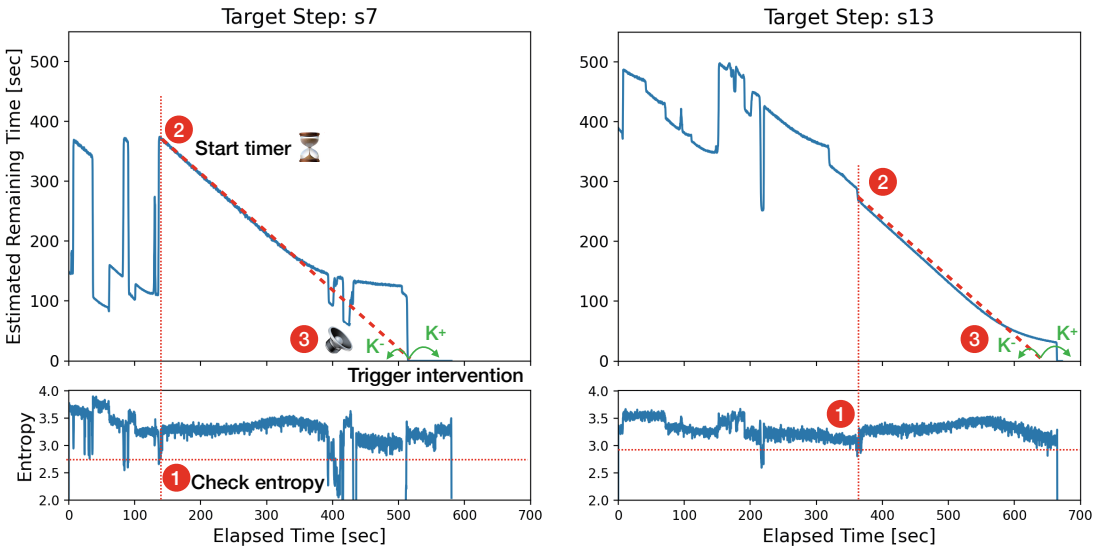


図 8: PrISM-Observer の介入タイミング決定。目標ステップ $\hat{s}$ までの残り時間の期待値 $E[D_t^{\hat{s}}]$ とエントロピー $H[D_t^{\hat{s}}]$ を逐次推定し，不確実性が十分下がった時点を捉えて，先回りリマインド（K^- 秒前）または失念通知（K^+ 秒後）を発火する。

前で促す**先回りリマインド**（$E[D_t^{\hat{s}}]$ の K^- 秒前）と，工程が検出されないまま時間が過ぎた場合の**失念通知**（K^+ 秒後）である。

この二種類は，要求するセンシング能力が異なる点が設計上重要である。先回りリマインドは目標ステップの前に発火するため，そのステップ自体を高精度に検出する必要はなく，残り時間の推定さえできればよい。一方，失念通知は「そのステップが実行されなかった」ことを判定して初めて発火するため，当該ステップを信頼して検出できる必要がある。そこで，各ステップの HAR 精度に応じて，どのステップにどちらの介入を割り当てるかを選択する。また，モンテカルロ推定に由来するエントロピーの揺らぎを抑えるため移動平均で平滑化し，タイマー始動後に期待値が大きく変動した場合はタイマーを破棄するなど，誤発火を抑える工夫を施している。特筆すべきは，この枠組みが**不完全なセンシングを前提に設計されている**点である。フレーム単位の HAR 精度が 50% 程度でも，残り時間を確率的に推論することで，実用に足るタイミングの介入を生成できる。これは「知覚の不完全さを，確率的推論で吸収する」という PrISM 全体の思想を，介入のタイミング決定において体現している。質問応答が利用者の主導に応える受動的な様式であるのに対し，観察・介入はシステムが主導する能動的な様式であり，両者を併せもつことで，利用者が気づく疑問にも気づかない誤りにも対応できる。**そして次節で述べるとおり，これら二様式の対話は，利用者を助けるだけでなく，システム自身の文脈理解を補正する入力としても機能する——この双方向性が PrISM の核心である。**

6　適応：対話を通じたオンライン文脈適応

6.1　コアとなる着想：不完全なセンシングを信頼性高く運用する

従来の文脈認識アシスタント研究の多くは，Wizard-of-Oz 法を用い「もしセンシングが完璧なら」という仮想的に完璧な支援を探究してきた。これに対し本研究は，センシングは原理的に不完全であるという現実を受け入れ，その不完全さを支援の設計で補うという立場をとる。具体的には，利用者

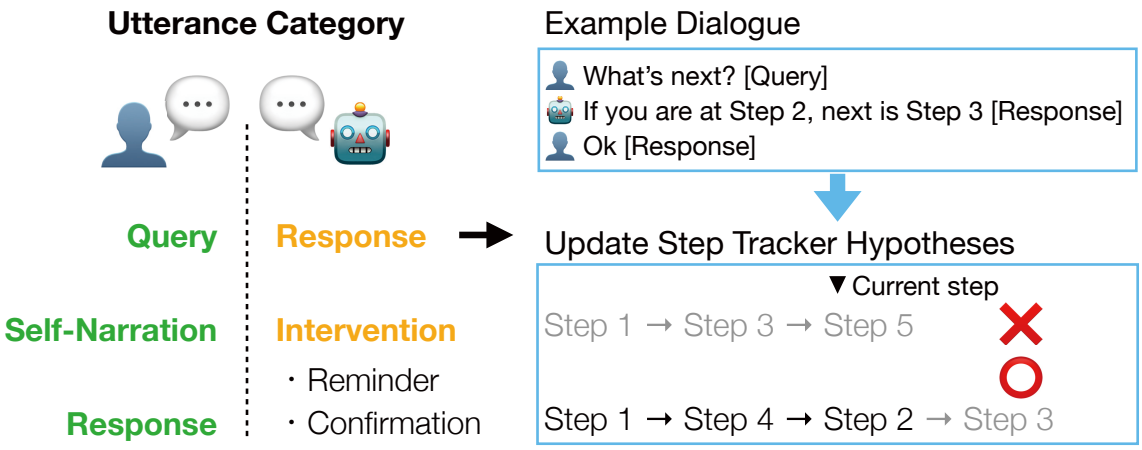


図 9: 混合主導対話に基づくオンライン文脈適応。利用者の発話（質問・自己ナレーション・応答）とアシスタントの発話（応答・介入）を Allen ら [37] の混合主導設計に基づいて分類し，各やり取りから工程文脈を抽出して追跡器の仮説集合 H を更新する。図の例では，利用者が「OK」と応答したことで，現在のステップが 3 であることが示唆されるため，システムは文脈に整合する仮説に確率を集中させている。

の質問・自己ナレーション・システムの確認への応答は，いずれもその時点での利用者の工程を示唆する。**これらを安価な教師信号として捉え，追跡器の内部状態をリアルタイムに更新すれば，再学習なしに文脈理解を改善できる**。これは，言語モデルへの人間フィードバック（RLHF [35]）が学習時にオフラインで作用するのに対し，推論時にオンラインで，かつ自然な対話の副産物として作用する点で対照的であり，物理世界で人間と協調する AI に適した適応様式である。

この立場は，現代の AI におけるいくつかの潮流と接続する。第一に，言語モデルの**接地（grounding）**である。LLM は膨大な手続き的知識をもつが，それを「いま・ここ」の物理状態に結びつける手段をもたない。PrISM は，センシングと対話の双方からその場の文脈を LLM へ供給することで，この接地を実現する。第二に，**テスト時適応**である。本機構は，配備後の各時刻に観測（対話）を取り込んでモデルの振る舞いを更新する点で，テスト時にモデルを適応させる枠組み [36] と軌を一にするが，更新信号が勾配ではなく記号的な制約（対話から抽出した工程情報）である点に特徴がある。第三に，**不確実性の明示的な取り扱い**である。追跡器は確率分布として仮説を保持し，そのエントロピーが高いときにのみ利用者へ問いかける。不確実性を定量化し，それを能動的な情報獲得（active sensing）の引き金とすることで，最小限の対話で最大の文脈改善を得る。これらの観点から，**PrISM は単なるシステム統合ではなく，「不完全な知覚をもつ AI が，人間との対話を通じて不確実性を縮減しながら協調する」という問題に対する一つの計算論的枠組み**を提示している。

6.2　混合主導対話の設計

PrISM は，利用者とアシスタントのあいだの発話を Allen ら [37] や Horvitz [38] の**混合主導（mixed-initiative）**支援の枠組みで整理する（図 9）。利用者の発話は三種に分けられる：(i) **質問**（タスクに関する問いや依頼），(ii) **自己ナレーション**（応答を期待せず自分の行動を口に出す），(iii) **応答**（アシスタントの発話への反応）。アシスタントの発話は，(i) 質問への**応答**と，(ii) 文脈理解に基づいて自ら開始する**介入**（重要工程のリマインドや，不確実性が高いときの「いま何をしています

か？」という確認）に分けられる。これらは，混合主導設計における「下位対話の開始（subdialogue initiation）」や「自発的報告（unsolicited reporting）」といった行為に対応づけられる。

アシスタントがいつ・何を話すかは，この観点から次のように設計する。利用者が質問したときは応答する。利用者が自己ナレーションしたときは，それを受け止め，次の工程の示唆などの追加情報を提供する。利用者が応答したときは，さらなるやり取りが必要かを判断する。利用者の現在ステップが不確かなときは，明確化のための質問を開始する。利用者があらかじめ定めた重要ステップに近づいたときは，リマインドを発火する。利用者を過度に煩わせないよう，介入の最大回数はハイパーパラメータで制御する。なお，本アシスタントの文脈認識はステップ情報に焦点を当てているため，視覚情報を要する問い（例：「これは何色？」）には答えられず，その場合は「分かりません」と応答する。対話全体は LLM ベースのパイプラインで，上記の方針に沿って振る舞うようプロンプトされて生成される。

6.3 対話からの文脈抽出と追跡器の更新

各対話のやり取りから，アシスタントは工程文脈を推定する。例えば「手を洗った後，次は何をすればいい？」という問いからは「現在は手洗いの工程」が，「手は洗いましたか？」「いいえ，まだです」というやり取りからは「その工程は未完了」が推定できる。LLM に，タスク固有情報と対話全体を与え，やり取りが (i) 現在ステップ，(ii) 過去ステップ，(iii) 未来ステップ，(iv) 文脈情報なし，のいずれを含むかを判定させ，該当する場合は具体的なステップを特定させる。重要なのは，この文脈が利用者の発話だけでなく，アシスタントの発話と利用者の応答からも抽出される点である。図 9 の例では，アシスタントが「ステップ 2 にいるなら，次はステップ 3 です」と応じ，利用者が「OK」と答えたことから，システムは「利用者はステップ 2 を終え，いまステップ 3 にいる」と推定し，これに整合する仮説へ確率を集中させる。リマインドや確認への応答も同様に，その時点の工程を強く示唆する手がかりとなる。

抽出された文脈は，追跡器の仮説集合の更新に用いられる。ここで仮説系列 $H(i,\tau)$ とは，第 4 節の隠れ状態 $h(i,\tau)$ に至る経路（すなわちそれまでに辿ったステップ列）を指し，$\boldsymbol{P}_t^{hypo} = \{P_t(i,\tau)\}$ はその経路確率である。追跡器は，利用者が現在 $h(i,t)$ にいるという各仮説とその確率を保持している。対話から得た文脈を制約 $\mathcal{C}$（例：「現在は第 i^* ステップ」「第 k ステップは未完了」）として表すと，更新は両立しない仮説のマスキングと再正規化

$$P_t(i,\tau) \;\leftarrow\; \frac{P_t(i,\tau)\mathbb{1}[H(i,\tau) \models \mathcal{C}]}{\sum_{i',\tau'} P_t(i',\tau')\mathbb{1}[H(i',\tau') \models \mathcal{C}]} \tag{5}$$

で与えられる（$H \models \mathcal{C}$ は仮説が制約と整合することを表す）。この更新は空間計算量を高々 $S \times T$ しか増やさず，リアルタイムで動作する。

■ハード棄却の頑健性　上式は整合しない仮説の確率を 0 にする**ハードな**棄却であり，文脈抽出が誤れば正しい仮説を消しかねない。実際，Study #2 では対話の 92% が正しく解釈されたが（第 7.2 節），残り 8% は誤抽出であった。しかし系は二つの理由でこの誤りから回復する。第一に，棄却後も各時刻で HAR 尤度 $\boldsymbol{p}^{har}$ が注入され続けるため，誤って 0 にされたステップも遷移を通じて再び確率を得られる（不可逆な消失ではない）。第二に，後続の対話が文脈を上書き的に訂正する。我々はこのハード棄却を，リアルタイム性と実装単純さの観点から採用したが，抽出確信度で重み付けする**ソフトな**更新（$\mathbb{1}[\cdot]$ を $[0,1]$ の整合度に置き換える）は，誤抽出への頑健性をさらに高める自

然な拡張であり，今後の課題である。要するに本機構は，自然な対話の副産物を構造化された追跡状態へ写像することで，知覚の誤りを継続的に訂正する仕組みである。

6.4 対話ポリシーと実装

混合主導対話を実現するため，システムは三つの並列プロセスを動作させる。第一は**応答プロセス**で，利用者の発話を受けて LLM で返答を生成する（PrISM-Q&A）。第二は**観察プロセス**で，PrISM-Observer のアルゴリズムを実行する。重要なのは，このプロセスがステップ追跡器を監視しているため，対話のやり取り（状態更新器を介した適応）が介入の振る舞いに動的に影響する点である。例えば，当初は利用者が目標ステップから遠いとシステムが考えていても，対話から実は近いと判明すれば，ただちにリマインドを発火するよう適応する。第三は**エントロピー監視プロセス**で，ステップ仮説の確率分布のエントロピーを定期的に計算し，閾値 E を超えた（不確実性が高い）ときに「いま何をしていますか？」という確認の問いを開始させる。利用者を過度に煩わせないよう，1 セッションあたりの確認回数は最大 M 回に制限する。

音声インタラクションは，**利用者発話中・アシスタント発話中・応答待ち・センサ読取中**の四状態で管理する。ステップ追跡器は，音声のやり取りが起きていない（センサ読取中の）あいだだけセンサデータを処理する。ウェイクワードは用いず，発話区間検出（VAD）により利用者の発話開始・終了を判定して状態遷移を制御する。

PrISM フレームワークは様々なモダリティのセンサを用いることができるが、以下のユーザ実験ではスマートウォッチ型のプロトタイプを用いた。プロトタイプは Apple Watch Series 7 と MacBook Pro（M1, 16GB）で構成し，音響（16 kHz）・運動（50 Hz）データを腕時計から逐次送信する。利用者は非利き手の手首に時計を装着し，Bluetooth イヤホンでシステムと対話する。HAR モデルには Mollyn ら [15] のものを用い，音声認識は端末上で局所的に（Whisper），読み上げは OS の機能で行う。LLM（GPT-4o-mini）は，遷移グラフ生成・Q&A 応答・対話からの文脈推定という三つの処理にそれぞれ専用プロンプトで用いる。主なハイパーパラメータは，フィルタリング閾値 β（学習データの約 10% を非重要と判定），確認を促すエントロピー閾値 E，確認の最大回数 M（$= 2$）である。利用者の発話終了から応答開始までの遅延は約 2 秒で，リアルタイムで動作する。PrISM システムが動作している様子はこちらの動画からも確認できる: `https://www.youtube.com/watch?v=q-Qh1ptKutQ`

7 ユーザ実験

PrISM フレームワークを，三つのユーザ実験と一つの応用探索で評価する。Study #1 では一度の実演からの学習効率を七つのタスクで検証し（第 7.1 節），Study #2 ではオンライン文脈適応が追跡精度と支援の質に与える効果を N=20 で，Study #3 では複数セッションでの利用者の自立支援を N=6 で評価する（第 7.2 節，第 7.3 節）。最後に，エンドユーザが自らのタスク向けにアシスタントを構築できることを確かめる（第 7.4 節）。なお以上の実験はすべて，所属する大学の倫理審査委員から事前に承認を得ている。

表 1: Study #1：一度の実演で学習した追跡器のフレーム単位精度（Macro-F1, %）。七つの日常タスクで従来法 [4] と比較。

タスク	従来の学習法 [4]	第 4.2 節の手法
コーヒー作り	21.8	44.9
紅茶作り	21.5	47.1
シリアル作り	27.6	34.8
サンドイッチ作り	23.5	44.0
ラテ作り	13.1	49.7
ステンシル作り	24.6	34.4
創傷ケア（臨床）	12.8	14.0

7.1 Study #1：一度の実演からの学習効率

スマートウォッチで記録した七つのタスク（合計 190 セッション）で，第 4.2 節 の効率的学習を評価した。これらのタスクは，台所での日常作業（コーヒー・紅茶・シリアル・サンドイッチ作り），機器を伴う複雑な手続き（ラテ作り・ステンシル作り），患者が臨床で行う医療タスク（創傷ケア）にわたり，分岐構造や所要時間，センサ信号の性質が大きく異なる多様な手続きを網羅する。各タスクの遷移グラフ（図 15）と追跡器の混同行列（図 16）を付録に示す。収集したこれらのデータセットと実装は，再現性と再利用のためオープンソースとして公開している [7]。表 1 のとおり，一度の実演のみで従来法を全タスクで上回り，改善幅は平均約 17.6 ポイント（タスクにより 1.2〜36.6 ポイント）であった。アブレーションでは，フィルタリングと遷移グラフ自動生成がそれぞれ 6.3%，12.7% ポイントの寄与を示した。遷移グラフ自動生成の寄与が大きいのは，前述のとおり従来法がグラフ構築に十分なデータを得られていなかったためである。一方，絶対精度はなお低く（例：ラテ作り 49.7%），手続き追跡の本質的難しさと，対話による適応の必要性を裏づける。特に分岐構造をもたない創傷ケアでは自動生成の効果が限定的で，フィルタリングの効果も小さかった。これは，センサ信号の変化が乏しいタスクが，学習データが限られる状況では本質的に追跡困難であることを再度示している。ここから二つの知見が得られる：(1) タスクは追跡可能であるために識別性のある信号を生む必要があり，(2) センシングの不完全さは残るため，暫定的な初期推定の上に頑健に積み上げるアシスタント設計が不可欠である。後者こそが，本稿の中核である対話を通じた適応（第 6 節）の出発点となる。

7.2 Study #2：オンライン文脈適応の効果（N=20）

最も複雑なラテ作りタスクを対象に，提案する**協調（collaborative）**アシスタント（混合主導対話によるオンライン適応あり）と，質問にのみ答える**受動（passive）**アシスタント（PrISM-Q&A 相当，適応なし）を，被験者間計画で比較した（N=20，いずれもラテ作り未経験者）。受動条件では，状態更新器と能動的介入を無効化し，Q&A 機能のみを残した。すなわち，システムは質問には答えるが，状態を適応させず，リマインドや確認の問いかけは行わない。測定指標は，タスク完了時間，追跡精度，および NASA-TLX [39] と System Usability Scale [40] による主観的評価に加え，応答の信頼性・知覚された追跡精度・協調感を問う三つの 7 件法質問とした。重要工程として，著者

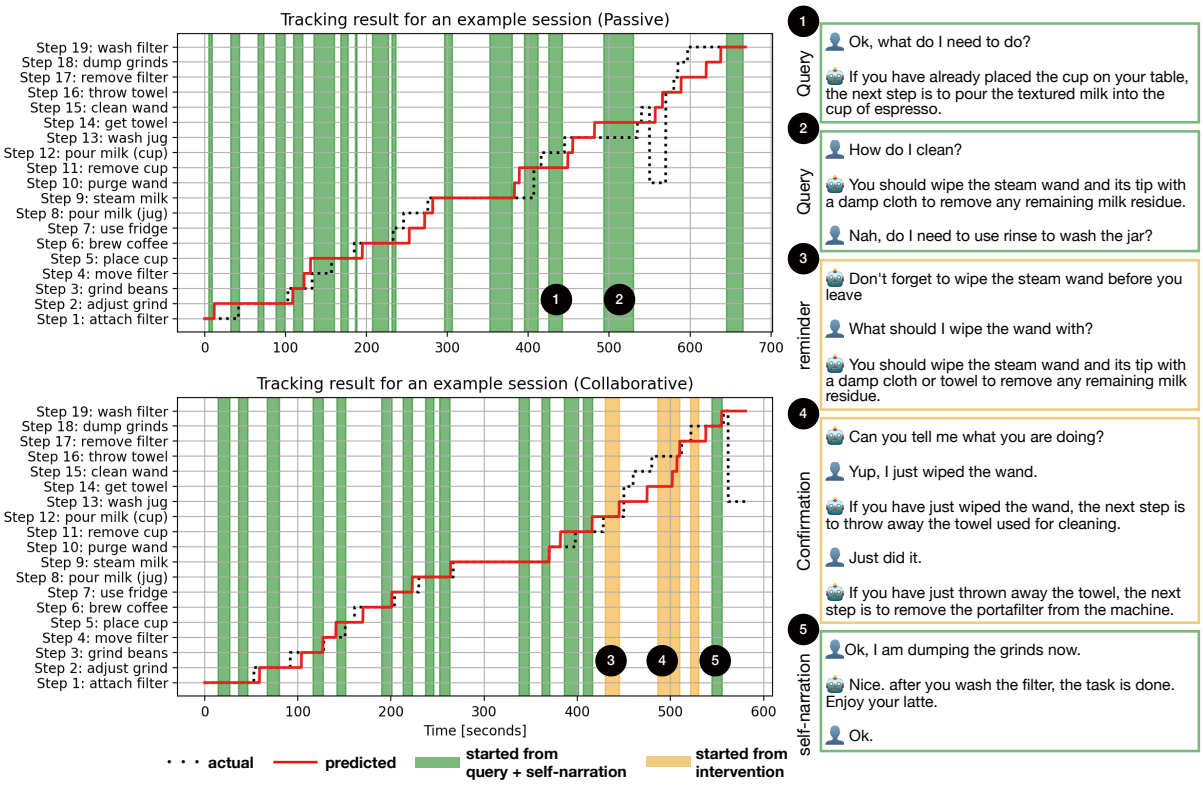


図 10: Study #2 のセッション例。黒破線が正解，赤実線がアシスタントの追跡予測。受動条件（上）では分岐の多い後半で追跡を見失い，応答が的外れになる。協調条件（下）では，確認の問いかけ (4)・リマインド (3)・利用者の自己ナレーション (5) といった対話のたびに追跡が更新され，高い文脈整合性を保つ（同例で 69.0% 対 54.1%）。

らは安全や失敗に関わるステップ（スチームワンドの洗浄，コーヒー粕の廃棄）をリマインドの対象に定めた。教示用の動画が手順の順序を刷り込む可能性を避けるため，異なる順序の動画を 5 種類用意して無作為に割り当てた。

追跡精度。20 名のうち課題を完遂した 17 名（受動 7 名・協調 10 名）を対象に集計した。フレーム単位精度（Macro-F1）は，受動条件で 45.4%（SD 13.2），協調条件で **71.3**%（SD 12.8）であり，両条件の差は統計的に有意で（Welch の t 検定，$p < 0.01$），効果量はきわめて大きかった（Cohen's $d \approx 2.0$）。重要な点として，この精度は対話の発生区間そのものを除外して測定している。すなわち「対話で正解を注入した瞬間」を除き，注入後にセンシングのみで進む区間での追跡を評価しているため，改善は単なる正解の転写ではなく，適応が後続フレームへ持続していることを示す。図 10 のセッション例が示すように，受動条件では分岐の多い後半で追跡を見失うのに対し，協調条件では各対話のたびに追跡が補正され，その効果が持続する。システムログ上，協調条件の対話の 92% が文脈抽出モジュールで正しく解釈されていた。

支援の質。追跡精度の向上は支援の質に直結した。不適切な応答の割合は，受動条件の 27.1%（SD 11.0）から協調条件では **9.4**%（SD 5.3）へ有意に減少した（Welch の t 検定，$p < 0.01$，Cohen's $d \approx 2.2$，図 11b）。不適切応答の多くは「次のステップは？」のような曖昧な問いに対する現在ステップの誤推定に起因しており，適応機構による追跡改善が，支援の質の向上へ直接的に波及している。

タスク遂行と対話行動。タスク完了時間は協調条件で短い傾向（482.0 対 633.9 秒）を示したが，個人差による分散が大きく，統計的に有意ではなかった。1 セッションあたりの利用者発話数は受動

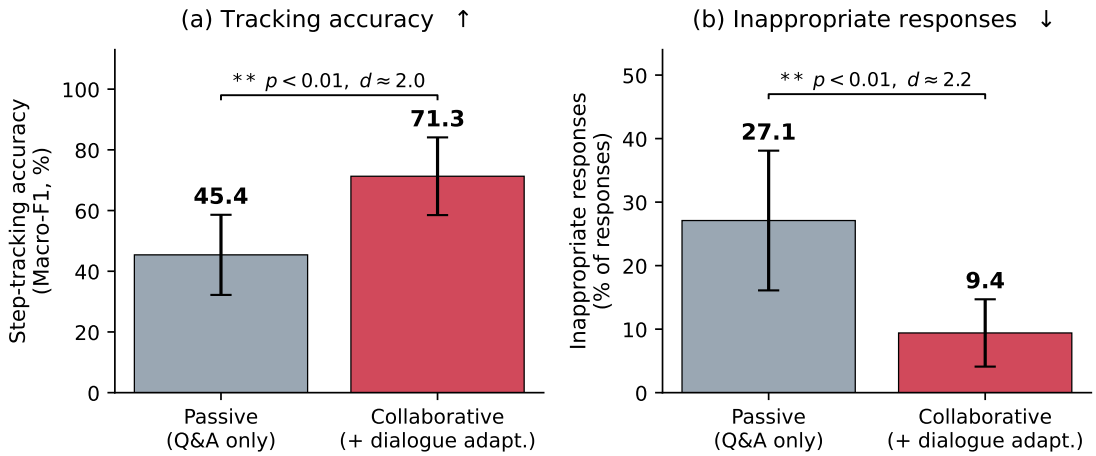
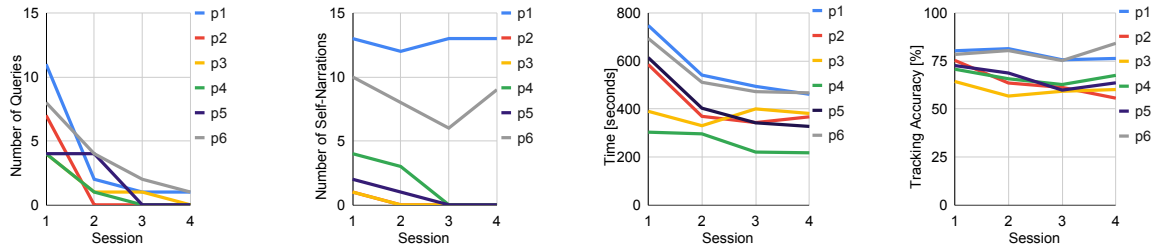


図 11: Study #2 の主要結果（被験者間，N=20，うち課題完遂者 17 名を解析）。混合主導対話によるオンライン適応を備えた協調条件は，受動（質問応答のみ）条件に対し，(a) 追跡精度を 45.4% → 71.3% へ向上させ，(b) 不適切応答を 27.1% → 9.4% へ低減した（いずれも Cohen's $d \approx 2$ の大きな効果量）。エラーバーは標準偏差。

19.6，協調 12.7 で，これも有意差はなかったが，受動条件では不適切応答を受けた利用者が問いを言い換えて再質問する傾向があり，発話数を押し上げたと考えられる。協調条件では，アシスタントから 1 セッションあたり平均 3.9 回の能動的介入（二つのステップへのリマインドと，不確実時の確認）があった。これらの介入は，後述するように利用者から適時かつ有用と受け止められ，協調感の醸成にも寄与した。

利用者体験。NASA-TLX では必要努力・遂行品質・総合スコアで，System Usability Scale（受動 68.5 対 協調 83.3）でも協調条件が有意に優れた。これは，質の高い Q&A と適時の能動的介入が，タスク遂行に伴う認知的負担を軽減したことを示す。受動条件の利用者が，不適切な応答を補うために問いを丁寧に言い換える必要があったのに対し，協調条件の利用者は大雑把な問いでも済んだことが，この差の一因と考えられる。さらに，追加質問への回答（7 件法）では，「アシスタントが工程を正確に把握している」という「知覚された追跡精度」と「協調感」において協調条件が有意に高かった（Mann-Whitney U 検定，$p < 0.05$）。インタビューでは，先回りリマインドが「ちょうど使い終わった直後にスチームワンドの洗浄を促してくれて驚いた，正確だった」と評価され，確認の問いかけも「センシングが完璧でないと感じたが，どこまで進んだか聞いてくれることで，ついてきてくれている感覚があった」と協調感に寄与したことが示された。すなわち，能動的な介入は，それ自体が文脈適応の機会であると同時に，利用者に「協調している」という感覚を与える二重の役割を果たす。

定性的知見。インタビューをオープンコーディング [41] で分析した結果，三つのテーマが浮かび上がった。第一に，**知覚精度の体感は対話の質を通じて形成**されていた。受動条件で曖昧な問いに不適切な応答を受けた参加者は「物理世界を私ほどには理解していない」と感じ，逆に協調条件では適時のリマインドによって「正確だ」と評価した。第二に，**アシスタントを助けるための労力が認知負荷を左右**した。受動条件のある参加者は「曖昧な問いはよくないと感じ，『コーヒーを淹れ終えた，次は？』のように補ったらうまくいった」と述べ，別の参加者は「行動を毎回ナレーションする必要があり少し面倒だった」と語った。一方，協調条件では「『次は？』のような大雑把な問いに答えてくれたので，余計な労力は要らなかった」との声があり，これは定量的な認知負荷の差を裏づける。第三に，**協調感を左右する要因として，システムの能動的な振る舞いが鍵**であった。受動条件の参加者



(a) 質問数　**(b)** 自己ナレーション数　**(c)** タスク完了時間　**(d)** アシスタントの追跡精度

図 12: Study #3（N=6，4 セッション）。全利用者が習熟とともに質問数を減らし（a），完了時間も短縮した（c）。自己ナレーションには個人差が現れ（b），頻繁に自己ナレーションした利用者（p1, p6）では追跡精度が高く保たれ（∼75%），対話量が減った利用者でも最終セッションの精度（∼62%）は受動条件（∼45%）を上回った（d）。

は「協調しようとしたが，どの段階にいるか正確に分からず，もどかしかった」「助けを求めたときだけ応じるので，協調というより補助に感じた」と述べた。対照的に，協調条件で確認の問いかけを受けた参加者は「センシングが完璧でないと感じたが，どこまで進んだか聞いてくれることで，ついてきてくれている感覚があった」と語り，混合主導設計の確認とリマインドが，利用者を過度に煩わせることなく協調感を醸成したことが示された。

7.3　Study #3：複数セッションでの自立支援（N=6）

Study #2 は単一セッションでの明確な効果を示したが，日常タスクでは利用者は同じシステムを繰り返し利用し，習熟していく。そこで，利用者とアシスタントの支援が時間とともにどう変化するかを掘り下げるため，複数セッションの研究を行った。同じラテ作りを 6 名が 1 週間に 4 セッション行い，各セッション前にはリマインドの希望（どのステップを，先回りで促すか失念時に通知するか）を利用者が設定できるようにした。本実験は，**統合されたシステムが複数セッションにわたって利用者をどう支えるか**という，これまで未解明だった側面に経験的な証拠を与える。全利用者が習熟とともに質問数を減らし（図 12a），完了時間も短縮した（図 12c）。自己ナレーションには個人差が現れ，4 名は次第に話しかけなくなった一方，2 名（p1, p6）はアシスタントとの対話自体を楽しみ，一貫して自己ナレーションを続けた（図 12b）。この差は追跡精度に反映され，頻繁に話しかけた利用者では精度が ∼75% と高く保たれた（図 12d）。重要なのは，対話量が減った利用者でも最終セッションの精度（∼62%）が，Study #2 の受動条件（∼45%）を上回った点である。これは，先回り介入への応答が追跡を補い，かつ利用者の習熟により行動が学習データに近づいたためと考えられる。リマインドの設定にも個人差が現れ，「念のため常に先回りで促してほしい」利用者と，習熟に応じて「失念通知」へ切り替える利用者がいた。混合主導設計は，こうした**時間とともに変化する個人の好みに柔軟に適応**し，利用者の自立を促しつつ必要な安全網を残すことができた。

定性的知見。インタビューのオープンコーディングからは，利用の仕方が時間とともに変化する様子が具体的に見えた。第一に，**自発的発話の進化**である。当初は逐次的な指示や詳細な回答を求めていた参加者も，習熟すると重要なリマインドや誤り防止の合図のみを望むようになった。「最初は多く質問したが，自信がついて聞く必要がなくなった」（p2）という声がある一方，p1・p6 は「各工程の後に自然とナレーションすることで，正しくできているか確認でき，また応答を聞くと誰かと一緒

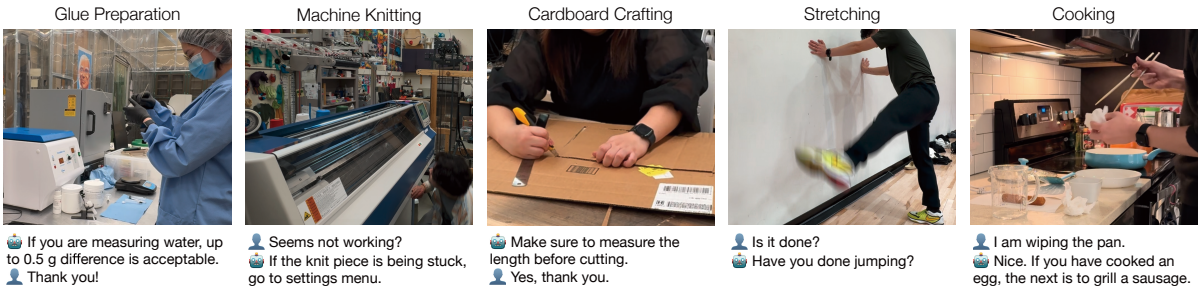


図 13: STEM 大学院生が PrISM フレームワークと一度の実演で自作した支援アシスタントの例。物理実験用の接着剤調製 (Glue Preparation)，編み機の操作 (Machine Knitting)，段ボール工作 (Cardboard Crafting)，運動前のストレッチ (Stretching)，料理 (Cooking) など多様なタスクへ適用された。

にいるようで楽しかった」と，対話自体を好んで続けた。第二に，**リマインド設定の戦略**にも違いが現れた。3 名は習熟後も「念のため先回りで促してほしい，重要な工程に気づける」(p6) と一貫した先回りリマインドを望み，残る 3 名は習熟に応じて「失念通知」へ切り替えつつ，「もし 1 か月後に同じ作業をするなら，リマインドを戻したい」(p3) と，長期の中断後には再び支援を受けられる選択肢を求めた。第三に，**確認の問いかけの受容性**である。習熟後でも 1〜2 回の確認は許容され，むしろ「いま何をしているか聞かれると，注意がタスクに戻って助かる」(p4) と歓迎された。頻繁すぎる確認は煩わしくなりうるが，回数を上限で制御する設計が，限定的な能動的介入を概ね好意的なものにしていた。

7.4　エンドユーザによる応用探索

フレームワークの汎用性を確かめるため，6 名の理工系大学院生に，各自のタスク向けの支援アシスタントを作成してもらった。**各自がステップを定義して一度実演し，手順書とリマインドの希望を与えるだけで，追跡器の学習とアシスタントの設定が完了する**。結果として，物理実験用の接着剤調製，編み機の操作，段ボール工作，運動前のストレッチ，定期的なオフィス清掃，朝食準備など多様なアシスタントが構築された（図 13)。利用者は，自分自身のための支援も，他者を助けるための支援も設計した。個人利用では，日常の決まった作業での先回りリマインド（ストレッチ・清掃・朝食準備）や，複雑なタスク（編み機操作）での Q&A が価値を見いだされた。他者支援では，初学者にとって混合主導対話がとくに有益だと指摘された（例：接着剤調製を不慣れな人に教える場面)。このように，一度の実演からの学習と対話による適応の統合は，**専門家でないエンドユーザが自らの文脈で支援 AI を構築することを可能にし，文脈認識アシスタントの民主化**を促す。物理実験から運動，家事まで広がる応用の幅は，本フレームワークが特定タスクに限定されない汎用的な基盤であることを示している。

8　議論：不完全なセンシングを信頼性高く使うために

一連の研究を通じて得られた中心的な知見は，物理世界の支援 AI において重要なのは知覚精度の極大化そのものではなく，不完全さに信頼性を付与する支援設計だという点である。これは，人間と知的エージェントが個々よりも協働でより良く振る舞う「人間–計算機パートナーシップ」や，双方が互いの振る舞いに適応する**相互的協適応** [42] の考え方と軌を一にする。本研究の文脈では，シス

テムが対話から学んで文脈理解を更新する一方，利用者もまた，システムの能力に合わせて問いの仕方やナレーションの頻度を調整する（Study #2・#3 で観察された）。この双方向の適応こそが，不完全な知覚という制約のもとで協調を成立させる原動力である。PrISM は，この原理を次の四点で具体化した。

不確実性へのロバスト性。多くのセンサにとって手続き中の行動を完璧に把握することは極めて困難であり，利用者は想定外の振る舞いも見せる。混合主導対話——とくにアシスタントの能動的な問いかけ——をオンラインのフィードバックとして用いることが，この不確実性に対処する鍵となる。重要なのは，システムが**自らの不確実性を自覚し，それが高いときにだけ利用者の助けを求める設計**である。これにより，常に問いかけるのでも，誤った確信のまま突き進むのでもない，中庸の振る舞いが実現する。

透明性と信頼。アシスタントが応答の冒頭で自らの文脈理解を述べることで，利用者は誤りを対話的に訂正でき，協調感が高まる。不透明な「ブラックボックス」的振る舞いは信頼を損ない，誤判断を招きうるが，PrISM は理解を可視化し，訂正可能性を担保する。Study #2 で応答の信頼性に条件間で差が出なかったのは，この透明性により，利用者が最初の応答が不適切でも問いを言い換えて正答に到達できたためと考えられる。透明性は，誤りを致命的な失敗ではなく回復可能なやり取りへと変える。

主導権の共有。制御を奪いすぎるシステムは利用者を無力化する。手続きタスクでは，利用者が自分の流儀で進め，必要なら無視し，異議を唱えられる自由が重要である。混合主導設計は，固定的な逐次指示ではなく，利用者の主導性を尊重しながら必要な時のみ介入する。Study #3 で参加者がセッションごとにステップ順を柔軟に変えた事実は，順序を強制しない設計判断の妥当性を裏づける。

個人差への適応。対話の様式は個人ごと・時間とともに変化する（Study #3）。ある利用者は対話自体を楽しんで自己ナレーションを続け，別の利用者は習熟とともに沈黙した。リマインドの好みも，常時先回りを望む者と失念通知に切り替える者に分かれた。介入頻度の調整や自己ナレーションの受容など，単一の支援バランスに固定しない柔軟さが，長期的な有用性を支える。これは，画一的な「最適方策」を求めるのではなく，**利用者ごとに異なる協調点へ収束できる適応性が，実世界の支援 AI に求められる**ことを示唆する。

9　将来研究と社会実装

医療分野への展開。PrISM の社会的インパクトが最も顕在化しつつあるのが医療・アクセシビリティ分野である。本フレームワークは，皮膚がんの手術後に患者が在宅で行う創傷セルフケアの支援に向けて，皮膚科外科医との連携のもと，臨床現場および患者の自宅で展開が進められている [8, 9]。創傷ケアは工程の抜けや誤りが感染や治癒遅延に直結する一方，患者は医療者の常時監督なしに自宅で手続きを正確に遂行しなければならず，まさに本フレームワークが対象とする課題である。皮膚科外科の患者を対象とした調査 [9] は，術後創傷ケアにおけるデジタル支援への需要と，その導入を阻む障壁の双方を明らかにしており，現場のニーズに根ざした技術開発の必要性を裏づけている。高齢化に伴う在宅医療の拡大と医療従事者の不足という社会的背景のもとで，こうした自立的なセルフケアを安全に支える技術の意義は大きい。

医療応用は創傷ケアにとどまらない。著者らは，認知症患者の在宅キッチンでの作業を見守り，火



(a) 皮膚がん術後の在宅創傷ケア支援
（プライバシに配慮したスマートウォッチベース）



(b) 認知症患者の在宅キッチンでの安全支援
（スマートウォッチ＋ LiDAR 等の環境センサベース）

図 14: PrISM の医療分野への展開。(a) 皮膚がん手術後の在宅創傷ケア支援。(b) 認知症患者の在宅キッチンでの作業を見守り，安全に関わるリアルタイムアラートを提供するプロジェクト。後者はスマートウォッチに加え LiDAR 等の環境センサを併用しており，PrISM がマルチモーダルなセンサ構成に対応することを示す。

や刃物の取り扱いなど安全に関わる場面でリアルタイムにアラートを出すプロジェクトにも，本フレームワークを展開している（図 14）。ここでは，常時装着が難しい場合も想定し，スマートウォッチに加えて LiDAR 等の環境センサを併用する。すなわち，利用者や生活環境に応じてセンサ構成を選べる PrISM のマルチモーダル性が，安全性が最優先される医療現場でこそ活きる。

さらに，追跡情報を医療者と共有すれば，自宅での回復過程を遠隔モニタリングする新たな可能性も開ける。患者がいつ・どのように手続きを行ったかの記録は，合併症の予兆を早期に捉えたり，不要な通院を減らしたりするための客観的な情報となりうる。同様の遠隔見守りは，認知症ケアなど，自立支援と安全確保の両立が求められる領域へも展開しうる。現時点では臨床的有効性の定量評価は進行中であり，今後の重要な検証課題だが，PrISM は単なる利便性向上にとどまらず，医療リソースの制約を補い在宅医療の質と安全を高める知能情報処理技術として，実社会での価値を生み始めている。

デバイスとセンサの拡張。本稿ではスマートウォッチに焦点を当てたが，追跡手法はモジュール式に他のプラットフォームへ拡張できる。介入の提示には視覚ディスプレイや，誤りを即座に防ぐ筋電気刺激 [43] なども有効であり，追加のセンサによる知覚能力の補完も今後の方向である。ここで，基盤となる HAR の精度と利用者体験の関係を定量的に解明することは，不確実性下の人間– AI 協調を設計するうえで重要な知見を与える。

利用者行動への仮定。現在の実装は，利用者が複数のステップを同時並行で行わない，および手順リストにない予期せぬ行動を取らない，という仮定を置いている。並行行動の認識はカメラを用いる場合でも困難な研究課題であり，また予期せぬ行動への対応には遷移グラフの拡張が必要となる。利用者の主導性をより広く許容するには，これらの仮定を緩める方向の研究が求められる。

オーサリングと安全性。新規タスクのアシスタントを作る際の実演・設定の手続きをさらに容易にすることも重要である。既存のタスク知識を活用したり，一度の実演中に利用者がナレーションで補足できるようにしたりすることで，効率的なオーサリングが期待できる。加えて，応答は LLM が生成するため，誤情報を防ぐ安全機構が不可欠であり，システム設計者やエンドユーザが振る舞いを直感的に設定できる仕組みが，とくに医療のような高リスク領域では重要となる。

評価上の限界。本稿の中心評価（Study #2・#3）は，分岐が多く難易度の高いラテ作りタスクに焦点を当てた。見出しとなる改善（追跡 45.4% → 71.3%）の一般性は，学習効率については 7 タスク・190 セッションで確認したものの（第 7.1 節），オンライン適応の効果については他タスクでの再現が今後の課題である。また比較対象は主に自己のアブレーション（協調 対 受動）であり，対話適応と「都度ユーザに直接質問する」素朴な方策との切り分けや，外部手法との比較を進めることで，貢献の位置づけがさらに明確になる。

10 結論

本稿は，フレームワーク PrISM を提案し，追跡 → 支援 → 適応という一貫した枠組みのもとに，四つの貢献として評価した。物理世界の手続き支援 AI が直面する本質的課題——少量データからの知覚学習と，不可避な不確実性下でのロバスト性——に対し，本フレームワークは，一度の実演からの効率的な追跡器学習と，混合主導対話を通じたオンライン文脈適応という二つの機構で応える。中核となる思想は，完璧な知覚を追い求めるのではなく，人間と AI の自然な対話を安価な適応信号として活用することで，不完全なセンシングを信頼性高く運用する点にある。スマートウォッチを用いた評価は，適応的対話によって，追跡精度（45.4% → 71.3%）と支援の質（不適切応答 27.1% → 9.4%，いずれも効果量 Cohen's $d \approx 2$）の有意な改善，および複数セッションにわたる利用者の自立支援を実証し，エンドユーザによる多様なアシスタントの自作と，医療分野での実展開がその実用性とインパクトを裏づけている。

AI が物理世界へと進出していくなかで，その成否を分けるのは，個々の知覚モデルの精度だけではなく，不完全な知覚をもつシステムが，リアルタイムに人間とどのように協調して不確実性を乗り越えるかという設計である。PrISM は，センシングと対話を統合し，言語モデルを物理世界へ接地させながら，自然な対話を安価なオンライン教師信号として不確実性を縮減する一つの計算論的枠組みを示した。現在は，皮膚がんの患者の術後の自宅での創傷ケアや，認知症の患者の自宅での日常生活支援に PrISM を適用し，その有効性を検証している。これは，**利用者を置き換えるのではなく，その活動を支援をするための実世界 AI への一歩であると同時に，知覚・言語・人間協調を横断する知能情報処理技術の今後の重要な研究方向を指し示す**ものである。

参考文献

[1] Xin Wang et al. HoloAssist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[2] Anind K. Dey. Understanding and using context. *Personal and Ubiquitous Computing*, 5(1):4–7, 2001. doi: 10.1007/s007790170019.

[3] Ryo Suzuki et al. Everyday ar through ai-in-the-loop. *arXiv preprint arXiv:2412.12681*, 2024.

[4] Riku Arakawa, Hiromu Yakura, Vimal Mollyn, Suzanne Nie, Emma Russell, Dustin P. DeMeo, Haarika A. Reddy, Alexander K. Maytin, Bryan T. Carroll, Jill Fain Lehman, and Mayank Goel. PrISM-Tracker: A framework for multimodal procedure tracking using wearable sensors and state transition information with user-driven handling of errors and uncertainty. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)*, 6(4):156:1–156:27, 2022. doi: 10.1145/3569504.

[5] Riku Arakawa, Hiromu Yakura, and Mayank Goel. PrISM-Observer: Intervention agent to help users perform everyday procedures sensed using a smartwatch. In *Proc. 37th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 15:1–15:16. ACM, 2024. doi:

10.1145/3654777.3676350.

[6] Riku Arakawa, Jill Fain Lehman, and Mayank Goel. PrISM-Q&A: Step-aware voice assistant on a smartwatch enabled by multimodal procedure tracking and large language models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)*, 8(4):180:1–180:26, 2024. doi: 10.1145/3699759.

[7] Riku Arakawa, Prasoon Patidar, Will Page, Jill Fain Lehman, and Mayank Goel. Scaling context-aware task assistants that learn from demonstration and adapt through mixed-initiative dialogue. In *Proc. 38th Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 2025. doi: 10.1145/3746059.3747700.

[8] Megan V. Ha, Emma Russell, Haarika A. Reddy, Alexander K. Maytin, Dustin P. DeMeo, Riku Arakawa, Mayank Goel, Jill F. Lehman, and Bryan T. Carroll. Self-narration for patient monitoring with smartwatch technology in post-operative wound care after dermatologic surgery. *Archives of Dermatological Research*, 316(7), 2024. doi: 10.1007/s00403-024-03149-z.

[9] Annalise Vaccarello, Alexander K. Maytin, Yash Kumar, Toluwalashe Onamusi, Haarika A. Reddy, Mayank Goel, Riku Arakawa, Jill Fain Lehman, and Bryan T. Carroll. Barriers to use of digital assistance for postoperative wound care: a single-center survey of dermatologic surgery patients. *Archives of Dermatological Research*, 316(7), 2024. doi: 10.1007/s00403-024-03025-w.

[10] Gaoping Huang, Xun Qian, Tianyi Wang, Fagun Patel, Maitreya Sreeram, Yuanzhi Cao, Karthik Ramani, and Alexander J. Quinn. AdapTutAR: An adaptive tutoring system for machine tasks in augmented reality. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems*, pages 417:1–417:15. ACM, 2021. doi: 10.1145/3411764.3445283.

[11] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys*, 54(4):1–40, 2021. doi: 10.1145/3447744.

[12] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. Ubicoustics: Plug-and-play acoustic activity recognition. In *Proc. 31st Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 213–224, 2018. doi: 10.1145/3242587.3242609.

[13] Karan Ahuja, Yue Jiang, Mayank Goel, and Chris Harrison. Vid2Doppler: Synthesizing doppler radar data from videos for training privacy-preserving activity recognition. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems*, pages 292:1–292:10. ACM, 2021. doi: 10.1145/3411764.3445138.

[14] Prasoon Patidar, Mayank Goel, and Yuvraj Agarwal. VAX: Using existing video and audio-based activity recognition models to bootstrap privacy-sensitive sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)*, 7(3):117:1–117:24, 2023. doi: 10.1145/3610907.

[15] Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. SAMoSA: Sensing activities with motion and subsampled audio. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)*, 6(3):132:1–132:19, 2022. doi: 10.1145/3550284.

[16] Ekaterina H. Spriggs, Fernando De la Torre, and Martial Hebert. Temporal segmentation and activity classification from first-person sensing. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, Washington, DC, 2009. IEEE Computer Society. doi: 10.1109/CVPRW.2009.5204354.

[17] Yasushi Nakauchi, Takuo Suzuki, Akira Tokumasu, and Sho Murakami. Cooking procedure recognition and inference in sensor embedded kitchen. In *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 593–600, New York, NY, 2009. IEEE. doi: 10.1109/ROMAN.2009.5326050.

[18] Jeremy Warner, Ben Lafreniere, George W. Fitzmaurice, and Tovi Grossman. Electrotutor: Test-driven physical computing tutorials. In *The 31st Annual ACM Symposium on User Interface Software and Technology, UIST 2018, Berlin, Germany, October 14-17, 2018*, pages 435–446. ACM, 2018. doi: 10.1145/3242587.3242591.

[19] Reiko Hamada, Jun Okabe, Ichiro Ide, Shin'ichi Satoh, Shuichi Sakai, and Hidehiko Tanaka. Cooking navi: assistant for daily cooking in kitchen. In *Proceedings of the 13th ACM International Conference on Multimedia, Singapore, November 6-11, 2005*, pages 371–374. ACM, 2005. doi: 10.1145/1101149.1101228.

[20] Daisuke Uriu, Mizuki Namai, Satoru Tokuhisa, Ryo Kashiwagi, Masahiko Inami, and Naohito Okude. Panavi: Recipe medium with a sensors-embedded pan for domestic users to master professional culinary arts. In *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems*, pages 129–138, New York, NY, 2012. ACM. doi: 10.1145/2207676.2207695.

[21] Ayaka Sato, Keita Watanabe, and Jun Rekimoto. Mimicook: A cooking assistant system with situated guidance. In *Proceedings of the 8th International Conference on Tangible, Embedded, and Embodied Interaction*, pages 121–124, New York, NY, 2014. ACM. doi: 10.1145/2540930.2540952.

[22] Jackie (Junrui) Yang, Leping Qiu, Emmanuel Angel Corona-Moreno, Louisa Shi, Hung Bui, Monica S. Lam, and James A. Landay. AMMA: adaptive multimodal assistants through automated state tracking and user model-directed guidance planning. In *IEEE Conference Virtual Reality and 3D User Interfaces, VR 2024, Orlando, FL, USA, March 16-21, 2024*, pages 892–902. IEEE, 2024. doi: 10.1109/VR58804.2024.00108.

[23] Javier Serván, Fernando Mas, José Luis Menéndez, and José Ríos. Assembly work instruction deployment using augmented reality. *Key Engineering Materials*, 502:25–30, 2012. doi: 10.4028/www.scientific.net/KEM.502.25.

[24] João Bernardo Alves, Bernardo Marques, Carlos Ferreira, Paulo Dias, and Beatriz Sousa Santos. Comparing augmented reality visualization methods for assembly procedures. *Virtual Reality*, 26 (1):235–248, 2022. doi: 10.1007/s10055-021-00557-8.

[25] Yuanzhi Cao, Xun Qian, Tianyi Wang, Rachel Lee, Ke Huo, and Karthik Ramani. An exploratory study of augmented reality presence for tutoring machine tasks. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM, 2020. doi: 10.1145/3313831.3376688.

[26] Alexandra Vtyurina and Adam Fourney. Exploring the role of conversational cues in guided task support with virtual assistants. In *Proc. 2018 CHI Conference on Human Factors in Computing Systems*, pages 208:1–208:7. ACM, 2018. doi: 10.1145/3173574.3173782.

[27] Razan Jaber, Sabrina Zhong, Sanna Kuoppamäki, Aida Hosseini, Iona Gessinger, Duncan P. Brumby, Benjamin R. Cowan, and Donald McMillan. Cooking with agents: Designing context-aware voice interaction. In *Proc. CHI Conference on Human Factors in Computing Systems*. ACM, 2024. doi: 10.1145/3613904.3642183.

[28] Justin Chan, Solomon Nsumba, Mitchell Wortsman, Achal Dave, Ludwig Schmidt, Shyamnath Gollakota, and Kelly E. Michaelsen. Detecting clinical medication errors with ai enabled wearable cameras. *npj Digital Medicine*, 7(1), 2024. doi: 10.1038/s41746-024-01295-2.

[29] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R. Cowan, and Heinrich Hussmann. Eliciting and analysing users' envisioned dialogues with perfect voice assistants. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems*, pages 254:1–254:15. ACM, 2021. doi: 10.1145/3411764.3445536.

[30] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.*, 135:364–381, 2022. doi: 10.1016/J.FUTURE.2022.05.014.

[31] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. Interactively picking real-world objects with unconstrained spoken language instructions. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 3774–3781. IEEE, 2018. doi: 10.1109/ICRA.2018.8460699.

[32] Vaibhav V. Unhelkar, Shen Li, and Julie A. Shah. Decision-making for bidirectional communication in sequential human-robot collaborative tasks. In *Proc. 2020 ACM/IEEE International Conference*

on Human-Robot Interaction (HRI), pages 329–341. ACM, 2020. doi: 10.1145/3319502.3374779.
[33] Shun-Zheng Yu. Hidden semi-markov models. *Artificial Intelligence*, 174(2):215–243, 2010. doi: 10.1016/j.artint.2009.11.011.
[34] Kenji Yamanishi, Jun ichi Takeuchi, Graham Williams, and Peter Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, 8(3):275–300, 2004. doi: 10.1023/B:DAMI.0000023676.72185.7c.
[35] Long Ouyang et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744, 2022.
[36] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.
[37] James E Allen, Curry I Guinn, and Eric Horvitz. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23, 1999.
[38] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 159–166. ACM, 1999. doi: 10.1145/302979.303030.
[39] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, volume 52, pages 139–183. Elsevier, 1988. doi: 10.1016/S0166-4115(08)62386-9.
[40] John Brooke. SUS: A 'quick and dirty' usability scale. In Patrick W. Jordan et al., editors, *Usability Evaluation in Industry*, pages 207–212. Taylor & Francis, 1996.
[41] Juliet Corbin and Anselm Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory.* SAGE Publications, 3rd edition, 2008.
[42] Michel Beaudouin-Lafon, Susanne Bødker, and Wendy E. Mackay. Generative theories of interaction. *ACM Trans. Comput.-Hum. Interact. (TOCHI)*, 28(6):1–54, 2021. doi: 10.1145/3468505.
[43] Romain Nith, Yun Ho, and Pedro Lopes. SplitBody: Reducing mental workload while multitasking via muscle stimulation. In *Proc. CHI Conference on Human Factors in Computing Systems*, pages 81:1–81:11. ACM, 2024. doi: 10.1145/3613904.3642629.

付録 A　収集したデータセットとオープンソース化

本研究の一つの貢献は，マルチモーダルセンサ（カメラ，スマートウォッチ）による手続きタスクの大規模なデータセットを収集・公開した点にある。Study #1 で用いた七つのタスクは合計 190 セッションにのぼり，台所での短い日常作業から，分岐の多い機器操作（ラテ作り・ステンシル作り），そして臨床現場で患者が行う創傷ケアまで，難易度と性質の異なる手続きを幅広く含む。図 15 に，各タスクの遷移グラフを示す。矢印の濃淡は遷移確率を，各ステップ下の数値は平均所要時間と分散を表しており，利用者がステップ順を一意に定めず多様な経路で手続きを遂行している様子（＝追跡の難しさ）が見て取れる。これらのデータセットと実装一式は，再現性の確保と他の研究者による再利用のため，オープンソースとして公開している [7]。非視覚センサのみで多様な実世界手続きを記録した本データセットは，実世界 AI の研究基盤として広く活用されうる。

付録 B　追跡器の混同行列

図 16 に，一度の実演で学習した追跡器の混同行列を，Study #1 の各タスクについて示す。対角成分が各ステップの正解率に対応する。分岐が多く所要時間のばらつきが大きいタスクや，創傷ケアのようにセンサ信号の変化が乏しいタスクほど混同が増える傾向が確認でき，本文で述べた「センサ

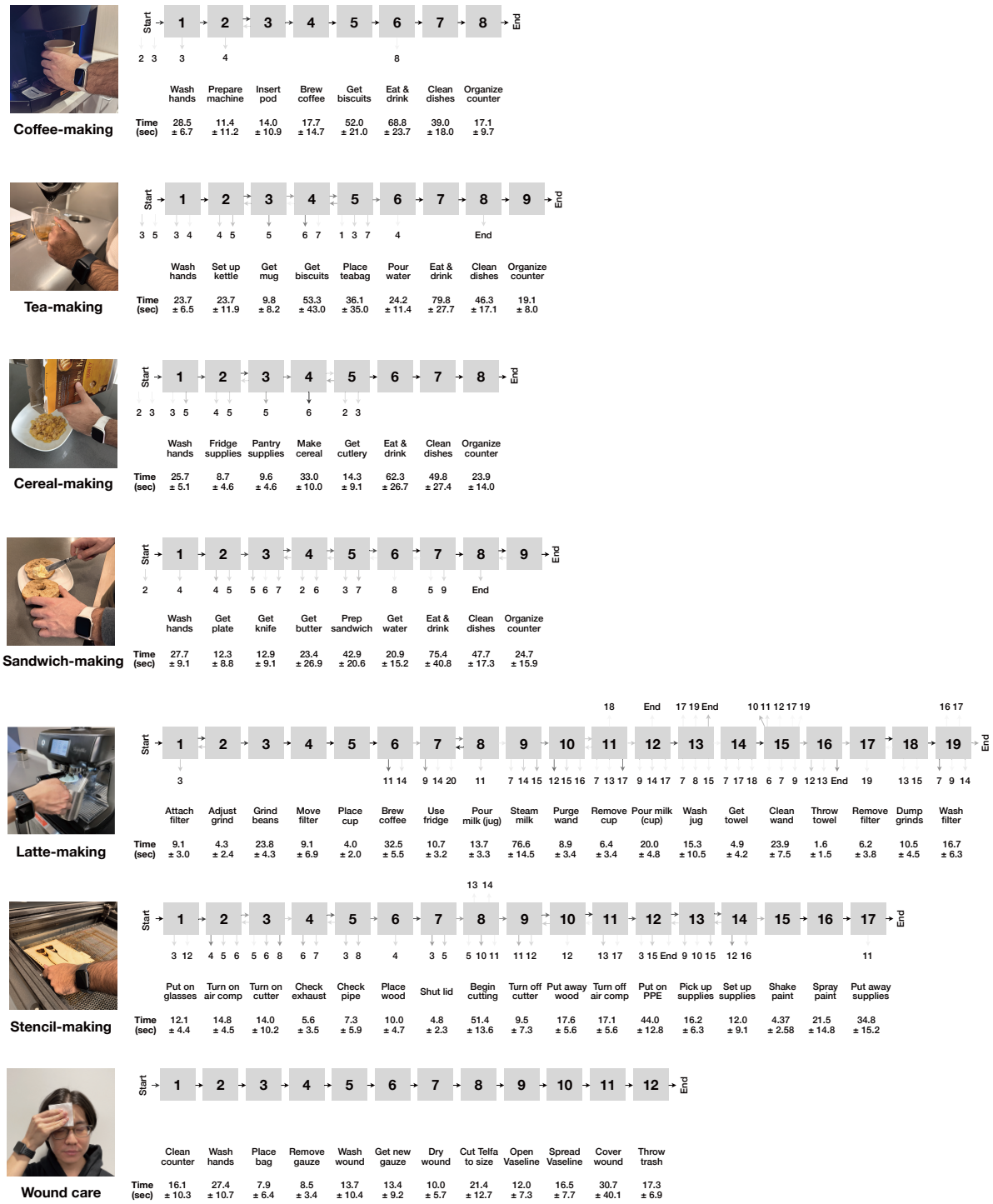


図 15: Study #1 で用いた七つのタスクの遷移グラフ。矢印の濃淡が遷移確率を表す（あるステップから出る矢印の確率和は 1.0）。各ステップには平均所要時間と分散が付記されている。コーヒー・紅茶・シリアル・サンドイッチ作りは短い日常作業，ラテ作り・ステンシル作りは分岐の多い複雑な手続き，創傷ケアは患者が臨床で行う単線的な医療タスクである。

で識別可能な信号を生む工程ほど追跡しやすい」という知見（第 7.1 節）を裏づける。こうした追跡器の不完全さこそが，本稿の中核である対話を通じたオンライン文脈適応（第 6 節）の動機である。

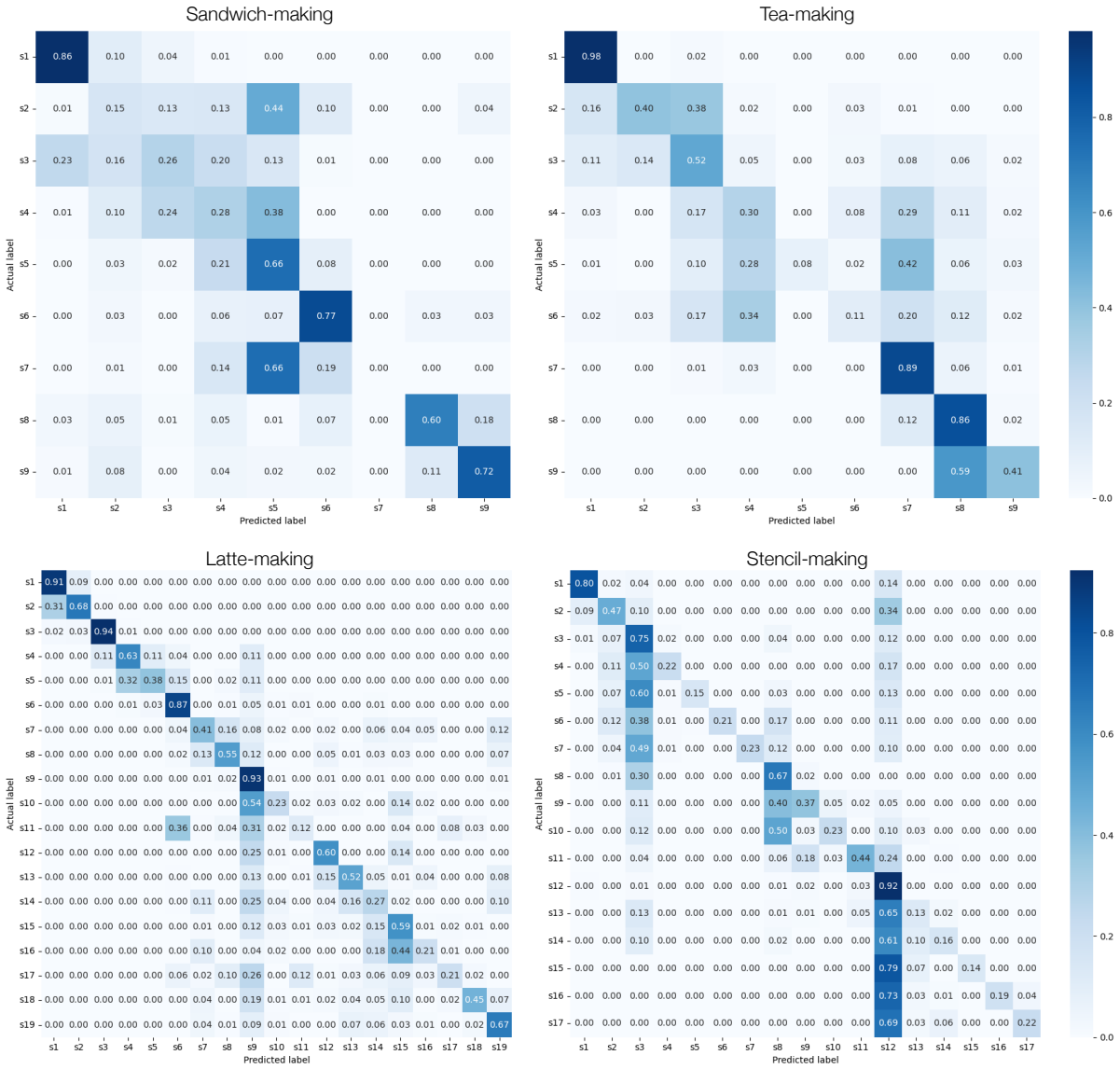


図 16: 一度の実演で学習した追跡器の混同行列（Study #1 の各タスク）。各行は実際のステップ，各列は予測されたステップを表し，対角成分が正解率に対応する。ステップのラベルは 図 15 を参照。