ABCI 3.0開発加速利用 （2024年度）成果概要（公開用）

課題名：動画基盤モデルの構築	実施時期：2025/1/20 ~ 2025/3/31 所属機関名：産業技術総合研究所 代表者氏名：原 健翔

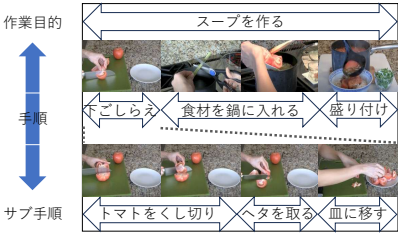
成果概要：本研究では動画を扱うAI基盤モデルとして，(1)手・物体インタラクションの詳細な理解を可能とする基盤モデルおよび(2)作業手順を理解可能な基盤モデルの構築を目的として研究開発を進めた． 手・物体インタラクション理解に関しては，本タスクにおいて本質的な要素を適切に評価するための新しいベンチマークを構築しABCI上で各種モデルの性能評価を実施した．手順理解に関しては，作業手順の階層的な理解を実現するためのモデルに関する研究開発を進めた．

成果のポイント：

本研究では，基盤モデルにより静止画内の内容を高精度に理解することは実現されつつあるコンピュータビジョン分野において，動画中の時系列的な変化を的確に理解するという未だ困難な課題に焦点を当てた．人が作業をする際には目的を持った一連の作業が行われ，各手順の中で扱う物体の状態は継続的に変化していくため，それらの手順を適切に切り分けつつその中で生じる手・物体インタラクションを理解することは難しいことから，手・物体インタラクションの理解および作業手順の理解の両側面から問題の解決に取り組んだ．

手・物体インタラクション理解に関しては，構築する基盤モデルの性能を適切に評価するためのベンチマークを構築しており，そのベンチマーク上で既存のモデルがどのくらい有効に機能するのかを評価するための計算資源としてABCI3.0を活用した．Qwen2.5-VLなど最新のマルチモーダル基盤モデルによる手・物体インタラクション理解能力を評価し現状の課題について分析した．

手順理解に関しては，右図に示すように，作業動画中に現れる手順を，大まかな粗い手順（例：食材の下処理）からより細かい1つ1つの詳細な手順（例：人参を乱切り）まで階層的に理解可能なモデルの研究開発を進めた．階層間の関係を捉えた上で適切な手順階層での表現を学習可能とするため，HT-Step，Ego4D Goal-Stepといったデータセットを利用しモデルの学習や表現手法の開発に取り組んだ．



成果についてより詳細な情報を提供しているWebページ、発表論文などの情報：

複数の国内/国際会議に現在論文を投稿中．