ABCI 3.0開発加速利用 （2025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：2025/4/15 ~ 2026/3/31 所属機関名：産業技術総合研究所 代表者氏名：八木 拓真

成果概要：本研究では動画向けのマルチモーダル基盤モデルの開発に向けて，(1) 詳細手・物体インタラクション理解のためのベンチマーク・モデル開発 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．(1) について，手操作の過程・作用の理解を要求する動画質問応答ベンチマークを構築，ABCI上で各種モデルの性能評価を実施した．その知見をまとめたベンチマーク論文がトップ国際会議（CVPR2026）に採択された．(2) については，作業手順の粒度を考慮した行動検出の手法開発に取り組み，国内会議で発表を行った．

成果のポイント：

(1)詳細手・物体インタラクション理解のためのベンチマーク・モデル開発：

手操作の動作・対象物体だけでなく，その過程や作用に関して総合的な理解を要する動画質問応答ベンチマーク（HanDyVQA）を構築し，既存の動画用の基盤モデル（InternVideo，Qwen，GPT-4oなど）が手操作の理解において人間と比較して動きや手操作によるシーンの変化の理解などに関して改善の余地を残していることを示し，その成果がコンピュータビジョン分野のトップ会議であるCVPR2026にて採択された [1]．また，手操作中の手および操作物体を明示的に扱った動画基盤モデルの開発に取り組んだ [2]．モデルの訓練・推論に加え，ベンチマーク構築のための質問生成などにおいてもABCI3.0を利用した．

(2)作業手順を理解可能な基盤モデルの構築：

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行動検出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，入力動画および検出したい行動のテキスト情報からその行動の粒度（長さ）を予測するモジュールを使う手法を提案し，国内学会にて発表を行った [3]．

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：

[1] Masatoshi Tateno, Gido Kato, Kensho Hara, Hirokatsu Kataoka, Yoichi Sato, and Takuma Yagi. HanDyVQA: A Video QA Benchmark for Fine-Grained Hand-Object Interaction Dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'26). 2026.
https://masatate.github.io/HanDyVQA-project-page/
[2] 加藤義道, 舘野将寿, 原健翔, 片岡裕雄, 森島繁生, 八木拓真. 手物体の位置情報を考慮した視覚言語モデルによる微細な一人称視点HOI理解. 第28回画像の認識・理解シンポジウム（MIRU2025, 一般論文）. 2025.
[3] 田中僚真, 木林佑太, 八木拓真, 片岡裕雄, 青木義満, 原健翔. 手順ラベル記述に基づく持続時間推定を用いた作業動画における手順検出, 第28回画像の認識・理解シンポジウム（MIRU2025, 一般論文）. 2025.