ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名：
BRIDGE事業：高度なヒューマンマシンインタ ラクションのためのLLMによるマルチモーダル AIの強化

実施時期：2025年4月～2026年3月
所属機関名：産業技術総合研究所
代表者氏名：浅田 真生

成果概要：

AIが動画を見ながら人間と自然に会話する技術を向上させるために、大規模かつ高品質な学習データ 「VDAct 2.0」を新たに構築した。家事などの日常生活シーンを撮影した約3,000本の動画を対象に、AIが「事実に基づいた適切な受け答え」だけを学習できるよう、複数の基盤モデルを利用し不適切な質問・回答を自動で取り除く仕組みを開発した。その結果、従来の約2倍にあたる6,356件の対話データを整備でき、このデータで学習し たAIの応答精度が一貫して向上することを実験で確認した。

成果のポイント：

① 不適切な発言に関する自動チェック機構の構築
動画理解AIの学習には、事実に基づく正確なデータが不可欠である。一方、人手作成では「～と思う」「嬉しそう」といった、映像から確認できない推測や感情表現が混入する。これらは蓄積により性能へ影響を与え得る。
本研究では、「推測・憶測」「感情・心理の読み取り」「スラング・不適切表現」「空想・誇張表現」の4類型を定義し、3モデル（Qwen3-235B、Qwen3-32B、GPT-5-nano）の多数決による自動判定機構を構築した。専門家評価で較正し、高い一致率と大規模適用性を確保した。
② 高品質データセット「VDAct 2.0」の構築および公開
約3,000本の動画を対象に、6組のアノテータペアが対話を作成し、6,356件・63,958ターンの「VDAct 2.0」を構築・公開した。全データに自動チェックを適用し、対話内の70%以上が適切と判定されたもののみを収録した。さらに、割合に応じてDiamond（90%以上）、Gold（80%以上）、Standard（70%以上）の3段階に分類した。アノテータ間の品質差も確認され、本手法の有効性が示唆された。
③ 新データセットによる性能向上の検証
GPT-5、Claude Sonnet 4.5等の商用モデルと、Qwen2.5-VL、InternVL3等を含む計11モデルで評価を実施した。追加学習により、すべてのオープンソースモデルで正確性・適切性が向上した。
特にInternVL3-8Bでは正確率が約9ポイント、適切性が約8ポイント向上した。一方、最高性能モデルでも40～50%程度にとどまり、時間的文脈を踏まえた対話生成の難しさが示された。

Dialogue ID: 0203806
Q: Could you describe to me the man's first action and where he was?
A: He was tidying dirty glasses in the kitchen.
Q: Did the man do this activity for a long time?
A: The whole activity took about a minute.
Q: After the man washed the wine glasses, did he do anything else in the kitchen before moving on to his next activity?
A: I didn't see him wash the glasses, just put them in the sink. After that, he went to the living room to read a book.
Q: Where did the man read the book in the living room?
A: He read it on the sofa.
Q: Where did the man get the book from?
A: From a bookcase in the living room
Q: What did the man do after reading the book?
A: He turned on the TV and went to the kitchen to get a drink.
Q: What did the man get to drink from the kitchen?
A: A bottle of wine.
Q: Where did the man drink the wine?
A: He drank the wine straight from the bottle while sitting on the sofa.
Q: After drinking the wine, what did the man do?
A: He cleaned the TV after drinking.
Q: How did the man clean the television?
A: By wiping it with a towel he got out of the bathroom.

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：

【発表論文】 Wiradee Imrattanatrai, Masaki Asada, Kimihiro Hasegawa, Ken Fukuda, Teruko Mitamura. “VDAct 2.0: Scaling Video-Grounded Dialogue for Event-driven Activity Understanding with LLM-Assisted Filtering.” International Conference on Language Resources and Evaluation (LREC2026) にて発表予定
【データセット公開先】 https://github.com/aistairc/VDAct2.0