# 仕　様　書

1. 件名

映像質問応答ベンチマークにおける多肢選択問題の人手評価

2. 研究の概要と背景

国立研究開発法人産業技術総合研究所人工知能研究センター（以下、「産総研」という）では、フィジカル領域の生成 AI 基盤モデルに関する研究開発の一環として、権利関係や倫理面での懸念が少ないマルチモーダル生成 AI モデルの構築に関する研究を行っている。この一環として、映像質問応答ベンチマーク HanDyVQA を開発した。

HanDyVQA は、人が様々な手作業を行う様子を記録した一人称視点映像に関する質問応答ベンチマークである。5 秒程度の短い動画クリップについて、映像中に映る手操作・物体および環境変化に関する推論能力を評価することを目的としている。

今回、本データセットに含まれる 11,668 問の多肢選択問題について人手評価を行う。

3. 用語の定義

- 「一人称視点映像」とは、行為者の頭部に装着した小型のカメラにより撮影した、行為者の視点に近い角度から記録した動画を指す。今回の作業で使用する動画は全て一人称視点映像である。
- 「動画クリップ」とは、産総研が提供する 5 秒程度の動画ファイルの内容を指す。
- 「質問」とは、1 つの動画クリップに対する質問文を指す。
- 「回答候補」とは、質問に対する回答の候補を示す文の組を指す。
- 「正答」とは、1 つの動画クリップと質問の組に対する正解を表す文を指す。
- 「質問回答ペア」とは、1 つの動画クリップに対応する質問・回答候補からなる組を指す。
- 「多肢選択問題」とは、1 つの動画クリップおよび質問が与えられた際、5 つの回答候補から正答を選ぶ問題を指す。
- 「質問カテゴリ」とは、ある動画クリップに対する質問文の種類を指す。カテゴリごとにアノテーションすべき情報が異なる。表 1 に各カテゴリの質問テンプレート及びその内容を示す。
- 「サンプル」とは、動画クリップ、付随情報および質問カテゴリからなる 1 問分のデータを指す。

受注者は各用語の定義を理解し、不明点があれば産総研に確認を行うこと。

4. 作業の概要

11,668 件の動画クリップに付与された 11,668 問の多肢選択問題について、1 問あたり 3 件、人間の作業者による回答の付与を行うこと。それと同時に、各質問回答ペアについて、原則と

してその回答をした作業者がその適格性に関する検証を行うこと。

本作業によって付与された回答および適格性の情報は、産総研が開発する動画像認識モデルの性能比較において、典型的な人間の作業者のパフォーマンスの報告に利用する。

5. 作業項目

(1) データの取扱い
(2) 作業の流れ
(3) 作業者の選定およびスクリーニング
(4) 多肢選択問題の回答および適格性の検証
(5) 人手評価結果のレビュー

5-1 データの取扱い

・評価用データは動画クリップおよび、各動画クリップに対して付与された質問および回答候補からなる。
・動画クリップ群は、産総研から受注者が指定するクラウドに、暗号化された通信手段(例えばSSL)によって転送するものとする。

5-2 作業の流れ

1) 産総研から評価用データ一式を受領する。
2) 少数のサンプル問題について、事前スクリーニングを実施し、作業者を選定する。
3) 各動画クリップについて、多肢選択問題の回答および適格性の検証を行う。
4) 3)の結果を産総研に納入する。
5) 産総研にて、4) の納品物をレビューして、修正点があれば差し戻しをする。
6) レビュー結果に問題なければ、回答結果を納入して納入の完了とする。

5-3 作業者の選定およびスクリーニング

作業者は、映像視聴が可能で英語を第一言語とする者、または第一言語と同等の言語能力を持つと判断される者であること。

本番に先立ち、問題および指示を適切に理解している作業者の事前スクリーニングを行うこと。事前に産総研が用意するサンプル問題(多肢選択問題 30 問)について、70%以上の正答率を達成した者のみを選任すること。回答基準については 5-4 に示すものと同様の基準で行うこと。

5-4 多肢選択問題の回答および適格性の検証

HanDyVQA ベンチマークに含まれる 6 カテゴリ、合計 11,668 問の多肢選択問題について、各問 3 人以上の作業者が下記 5-4-1.に示す回答基準に基づいて回答を行うこと。また、各選

択肢に関する適格性の検証を行うこと。図 1 に各カテゴリの設問例を示す。

各問題は (i) 5 秒間の一人称視点映像クリップ (ii) 質問文 (iii) 回答候補(5 つ、ただし Object カテゴリは最大 10 個)からなる。問題の種類別に 6 カテゴリ(Action, Process, Location, State Change, Object Part, Object)のいずれかに分類される。各カテゴリの質問テンプレート及びその内容を表 1 に、質問カテゴリ別の設問数を表 2 に、カテゴリ別の質問・回答例を表 3 に示す。

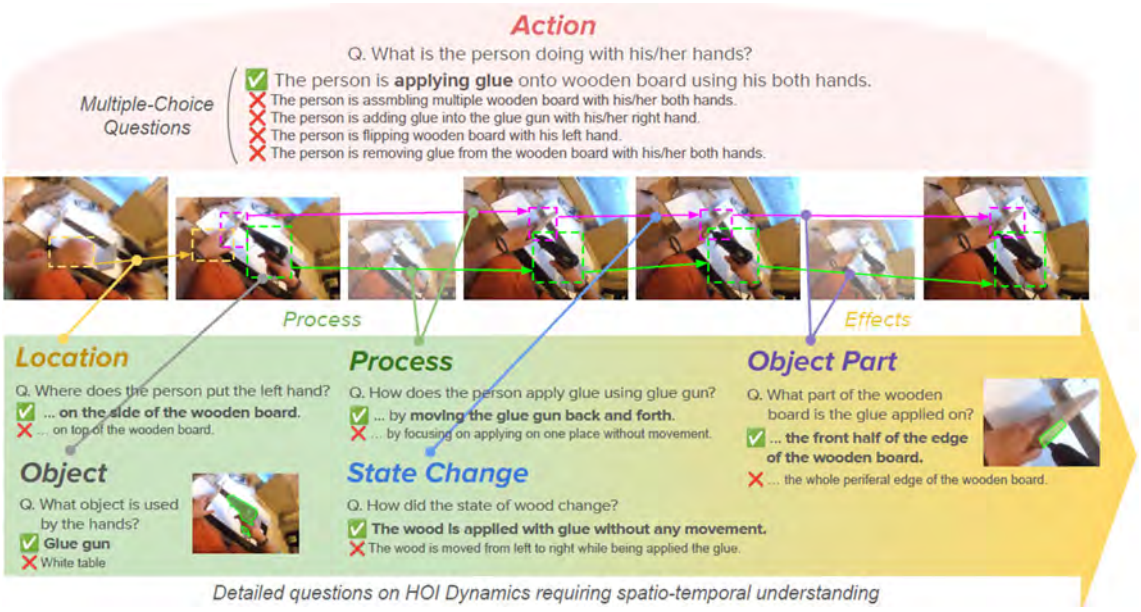



**図 1 HanDyVQA の各カテゴリの問題例。**

5-4-1 問題の回答

作業者は、各設問について、次の回答基準に従って正答と思われる選択肢を回答すること:

> 一人称視点映像を注意深く、特にイベントの原因および系列、物体の詳細や動き、人の行動や姿勢に注目して閲覧してください。提示された候補の中で動画の内容と照らし合わせて最も適当と思われる選択肢を 1 つだけ(Object カテゴリは当てはまるものすべてを)選択してください(Carefully watch the first-person view video and pay attention to the cause and sequence of events, the detail and movement of objects, and the action and pose of persons. Select only one option except Object category, which seems to be the most appropriate based on the contents of the video. For Object category, select all options that seem to be correct.)

5-4-2 適格性の検証

作業者は、設問の回答後、正解の選択肢を確認したうえで、下記項目について該当するも

のをチェックすること。

A) 正答が不適当である（質問文に対して適切に回答していない または 正答が動画クリップ内のイベントの内容と対応していない）

B) 正答以外に動画クリップの内容と照らし合わせて適当と思われる回答候補がある

Bについては、正答以外で適当と思われるすべての選択肢についてそれぞれチェックを行うこと。適格性の検証に関する例を表 4 に示す。

5-4-3 タスクの実施にあたっての留意点

- 作業者は、自身が有する知識に基づいて設問に回答すること。質問・回答候補の文章が示す内容を正確に理解することを目的とする場合を除き、動画内容の理解および設問の回答の補助または代替などを目的として第三者の補助またはインターネットを利用することを禁ずる。外部 AI ツールの使用および AI によって生成された情報の利用は作業中のいかなる場合にもそれを禁止する。
- 公平な評価のため、事前スクリーニングに合格した 3 名の作業者がそれぞれ全件の人手評価を行うことが望ましい。ただし、運用上の理由などにより、1 問に対してユニークな 3 名が評価を行える限りは、4 名以上の作業者により作業を分担しても差し支えない。

5-5 人手評価結果のレビュー

5-4 の人手評価結果について産総研が最終レビューを行う。結果が不適当と判断された場合は差し戻しを行い、目標数量（11,668 件）を達成するまで作業を繰り返すこと。

6. 支給品

・評価用データ（動画クリップおよび、各動画クリップに対して付与された質問および回答候補） 一式

7. 納入の完了

5-2 に示す項目をすべて行い、納入物品がすべて仕様通りである旨の最終レビューが済んでいることを産総研が確認して納入の完了とする。

8. 受注者の要件

受注者は以下の要件を満たすこと。各要件について、公開できる範囲においてそれを証明する書類のコピー等を事前に提出すること。

(a) 過去 5 年間に同規模（1 万件以上）の人手評価実績を 1 件以上有すること。

(b) 過去 5 年間に動画データを対象として、同規模（1 万件以上）の人手評価またはデータ作成実績を 1 件以上有すること。

(c) 過去 5 年間にコンピュータサイエンス分野の学術論文誌(和・英問わず)または査読付き国際会議に出版された論文等において使用された人手評価の実績を 1 件以上有すること。

9. 特記事項

(1) 評価用データの動画クリップはインターネット上に公開されているデータベースから取得したものであるが、その内容そのものは本課題の実施にあたっての非公開情報であるため、外部に漏洩しないよう、慎重に取り扱うこと。

(2) 受注者側では作業責任者を選任し、受注者および再委託先がある場合は再委託先の事業者に係る実施体制図(各法人名およびその役割を記したもの)を産総研に提出すること。実施体制図に記した事業者の外にデータを流通させることを厳禁とする。

(3) 受注者側のクラウドにおいては、先の実施体制図に記載された事業者内の人員以外のアクセスは禁止すること。

(4) 外部からの不正アクセスによって情報漏洩した疑いがある場合には、速やかに産総研に連絡し、対応を仰ぐこと。

10. 成果の取扱い

(1) 産総研は、受注者がアノテーションにより得られた技術上の成果のうち産総研が指示するもの(以下「成果」という。)についての利用及び処分に関する権利を専有するものとする。

(2) 受注者は、成果に係るアノテーションの著作権を産総研に無償で譲渡するものとし、著作者人格権を行使しないものとする。

(3) 受注者は、産総研に対し、納品した成果品が第三者の著作権を侵害しないことを保証するものとする。なお、納品した成果品について、第三者の権利侵害の問題が生じ、その結果、産総研又は第三者に費用や損害が生じた場合は、受注者は、その責任と負担においてこれを処理するものとする。

11. 納入物品

以下の納入物品を電子ファイルとして、ファイル転送サービスによって納品すること。xlsx 形式のシートまたは csv などの可読形式を想定するが、具体的なファイル形式およびフォーマットについては協議の上決定する。

・多肢選択問題 11,668 件に対する 3 名分の回答・検証結果(合計 35,004 回答)の電子ファイル　一式

12. 納入期限及び納入場所

納入期限: 2025 年 10 月 31 日(金)

納入場所: 〒305-8560 茨城県つくば市梅園 1-1-1
国立研究開発法人産業技術総合研究所　人工知能研究センター
中央事業所つくば本部・情報技術共同研究棟 4 階 4301 室

13. 付帯事項

・本仕様書の技術的内容及び知り得た情報については、守秘義務を負うものとする。
・本仕様書の技術的内容に関する質問等については、調達請求者と協議すること。また、本仕様書に定めのない事項及び疑義が生じた場合は、調達担当者と協議のうえ決定する。

表 1 質問テンプレートおよびその内容

| 質問カテゴリ | 質問文テンプレート | 質問内容 |
|---|---|---|
| Action | What is the person doing with his/her hands? | 行為者の手操作およびその周辺環境(物体など)への影響 |
| Process | How does the person [Action] [Object]? | どのように手や道具を使用しているか |
| Location | Where does the person [Action] [Object]? | 操作された物体がどこへ移動したか |
| State Change | How did the state of the [Object] change? | 対象物体の状態・構造・構成・配置などが動画内でどう変化したのか(あるいはしていないのか) |
| Object Part | What part of [Object] is being [Effect]? | 手操作の影響範囲がどこか |
| Object | What object is used by the hands? | 手で操作している物体の種類および位置 |

表 2 質問カテゴリ別の設問数

| 質問カテゴリ | 設問数 |
|---|---|
| Action | 1,978 |
| Process | 1,924 |

| | |
|---|---|
| Location | 1,974 |
| State Change | 1,940 |
| Object Part | 1,913 |
| Object | 1,939 |
| 合計 | 11,668 |

表 3 質問・回答例

以下に、各カテゴリ 2 問ずつ動画クリップの抜粋(1 秒ごとの静止画)、質問、回答候補および正答を示す。

カテゴリ:Action

Q. What is the person doing with his/her hands?

(A) The person is trimming the tape with the scissors in his right hand while holding the roll of tape with his left hand.

(B) The person's left hand holds the strap and cuts it with the cutter in his right hand.

(C) The person is tearing the tape by hand while keeping it taut using both hands.

(D) The person is measuring the tape with a measuring tape in his right hand while holding the cutter with his left hand.

(E) The person is applying tape to a box with the adhesive side down, using his right hand to press it while holding the box in his left hand.

正答:(B)

Q. What is the person doing with his/her hands?

(A) The person pries a strip of wood from the wall with the scraper in his right hand.

(B) The person peels a strip of paint off the wall with the scraper in his right hand.

(C) The person pushes a wood piece into the wall with a tool in his right hand.
(D) The person removes a strip of wallpaper from the wall using the scraper in his right hand.
(E) The person carves a groove into the wood on the table using the scraper in his right hand.
正答:(A)

カテゴリ:Process

Q. How does the person paint the ceiling with the paint roller?
(A) The person carefully applies paint onto the ceiling using both hands.
(B) The person gently paints the ceiling horizontally using the right hand.
(C) The person paints the ceiling vertically while holding the roller with the left hand.
(D) The person cautiously raises the roller above the head, allowing it to rest on the ceiling.
(E) The person smoothly rolls the roller around in circles on the ceiling using the right hand.
正答:(B)

Q. How does the person wipe the piece of white cloth with her left hand?
(A) The person wipes the piece of cloth in a circular motion with her left hand around the surface.
(B) The person moves the piece of cloth back and forth quickly with her left hand.
(C) The person wipes the cloth down along the surface using her left hand in a downward motion.
(D) The person wipes the piece of cloth from left to right using her left hand.
(E) The person wipes the piece of cloth from right to left using her left hand.
正答:(E)

カテゴリ:Location

Q. Where does the person put the bucket?

(A) The person placed the bucket on top of the other bucket.

(B) The person placed the bucket beside the other bucket.

(C) The person placed the bucket underneath the other bucket.

(D) The person placed the bucket in front of the other bucket.

(E) The person placed the bucket behind the other bucket.

正答:(A)

Q. Where does the person put the serveware?

(A) The person put the serveware into the top left of the refrigerator.

(B) The person put the serveware onto the bottom shelf of the refrigerator.

(C) The person put the serveware into the bottom right of the refrigerator.

(D) The person put the serveware into the middle section of the refrigerator.

(E) The person put the serveware onto the top right of the refrigerator.

正答:(C)

カテゴリ:State Change

Q. How did the state of a stool change?

(A) The stool was rotated 45 degrees clockwise while remaining on the ground.

(B) The stool was rotated 90 degrees clockwise while tilting slightly.

(C) The stool was rotated 90 degrees counterclockwise while remaining on the ground.

(D) The stool was rotated 180 degrees counterclockwise while lifted briefly.

(E) The stool was rotated 90 degrees counterclockwise while then moving slightly forward.

正答:(C)

Q. How did the state of a part of the tent change?

(A) The zipper of the tent was moved from the top to the bottom, causing the flap to close.

(B) The zipper of the tent was moved back and forth, causing the flap to remain partially open.

(C) The zipper of the tent was moved from the bottom to the top, causing the flap to open.

(D) The zipper of the tent was moved halfway, causing the bottom half of the flap to stay shut.

(E) The zipper of the tent was moved slightly upwards, causing a small opening at the bottom of the flap.

正答:(C)

カテゴリ:Object Parts

Q. What part of the bicycle cassette is cleaned?

(A) The bottom part of the bicycle cassette is cleaned.

(B) The top part of the bicycle cassette is cleaned.

(C) The inner part of the bicycle cassette is cleaned.

(D) The middle of the bicycle cassette is cleaned.

(E) The back part of the bicycle cassette is cleaned.

正答:(B)

Q. What part of the Jenga tower is removed?

(A) The topmost piece of the Jenga tower is removed.

(B) The piece from the middle section of the Jenga tower is taken out.

(C) The piece from the uppermost layer of Jenga blocks is extracted.

(D) The piece from the left side of Jenga is removed.

(E) The bottom block of the jenga blocks is removed.

正答:(E)

カテゴリ:Objects

Q. What object is used by the hands?

(A) Bottle.

(B) Hand drill.

(C) Wire.

(D) Cutter plier.

(E) Tool box.

正答:(C)

Q. What object is used by the hands?

(A) Liquid sprayer.

(B) Pedal.

(C) Chain.

(D) Frame.

(E) Empty container.

(F) Paint brush.

(G) Pressure sprayer.

正答:(A), (F), (G)(複数あることに注意)

表 4 適格性の検証の例

(i) 正答が不適切(動画内のイベントと不一致)な例

カテゴリ:Object Parts

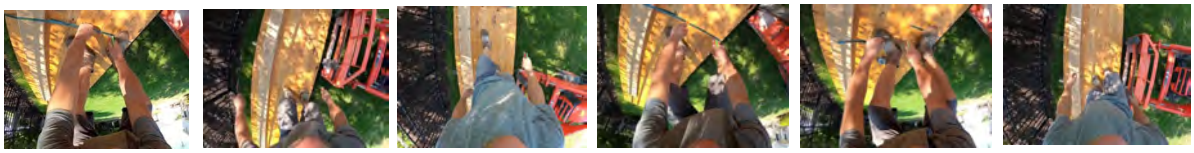
Q: What part of the bicycle cassette is cleaned?

正答 : The **bottom** part of the bicycle cassette is cleaned.

→ 動画ではカセットの上の部分が掃除されているため、正答として不適切。

(ii) 正答の他に正しい選択肢がある例

カテゴリ:Action



Q. What is the person doing with his/her hands?

(A) The person is trimming the tape with the scissors in his right hand while holding the roll of tape with his left hand.

(B) The person's left hand holds the strap and cuts it with the cutter in his right hand.

(C) The person is tearing the tape by hand while keeping it taut using both hands.

(D) The person is measuring the tape with a measuring tape in his right hand while holding the cutter with his left hand.

(E) The person is cutting the tape using both hands on the wood.

正答: (B)

→ (B)の他に、(E)も正解と言える。