実世界の事物と紐づいた
対話機能を持つロボットを目指して

理化学研究所 ガーディアンロボットプロジェクト (GRP)
吉野 幸一郎

知識獲得・対話研究チーム
Knowledge Acquisition &
Dialogue Research Team

ガーディアンロボット
プロジェクト
Guardian Robot Project

奈良先端大
ロボット対話知能研究室
Intelligent robot dialogue
laboratory, NAIST

運営より

◆質疑は Slido を利用します

自己紹介

◆吉野 幸一郎 (Koichiro Yoshino)

- 2009　学士（環境情報学） 慶應義塾大
- 2014　修士（情報学） 京都大, 博士（情報学） 京都大
- 2014　学振PD 京都大 学術情報メディアセンター
- 2015　特任助教、助教 奈良先端大
- 2020　現職（**理研GRP知識獲得・対話研究チーム**リーダー）
  - 自然言語処理、音声言語処理、
    対話システム、対話ロボット（と機械学習）の研究

- 2016-2020　さきがけ研究員 JST
- 2017-2020　客員研究員 理研AIP
- 2019-2020　客員研究員 ハインリッヒハイネ大
- 2020-2022　客員准教授 奈良先端大
- 2022-　客員教授 奈良先端大（**ロボット対話知能研**）

Dialog System Technology Challenge

◆世界を巻き込んだ対話研究のオープンサイエンス化

- Google, Amazon, Meta, Microsoft, NTT, Mitsubishi などが shared task としてデータを公開・ベンチマーク化する活動を主導 (2015-)
- 毎年 AAAI 併設のワークショップを実施
- 研究のトレンドとしては実世界化、○○-grounded

DSTC10 (2021)

- Five Parallel Tracks
  - MOD: Internet Meme Incorporated Open-domain Dialog: Task Description, Resources, Overview Paper
  - Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations: Task Description, Resources, Overview Paper
  - SIMMC 2.0: Situated Interactive Multimodal Conversational AI: Task Description, Resources, Overview Paper
  - Reasoning for Audio Visual Scene-Aware Dialog: Task Description, Resources, Overview Paper
  - Automatic Evaluation and Moderation of Open-domain Dialogue Systems: Task Description, Resources, Overview Paper
- Challenge Organizers: Koichiro Yoshino, Yun-Nung (Vivian) Chen, Paul Crook, Satwik Kottur, Jinchao Li, Behnam Hedayatnia, Seungwhan (Shane) Moon
- Venue: DSTC10 Workshop @ AAAI-22

研究室のメンバー

◆特別研究員

- 湯口 彰重
- 河野 誠也
- Angel Garcia Contreras

◆テクニカルスタッフ

- 波部 英子

◆JRA

- 田中 翔平

◆客員研究員

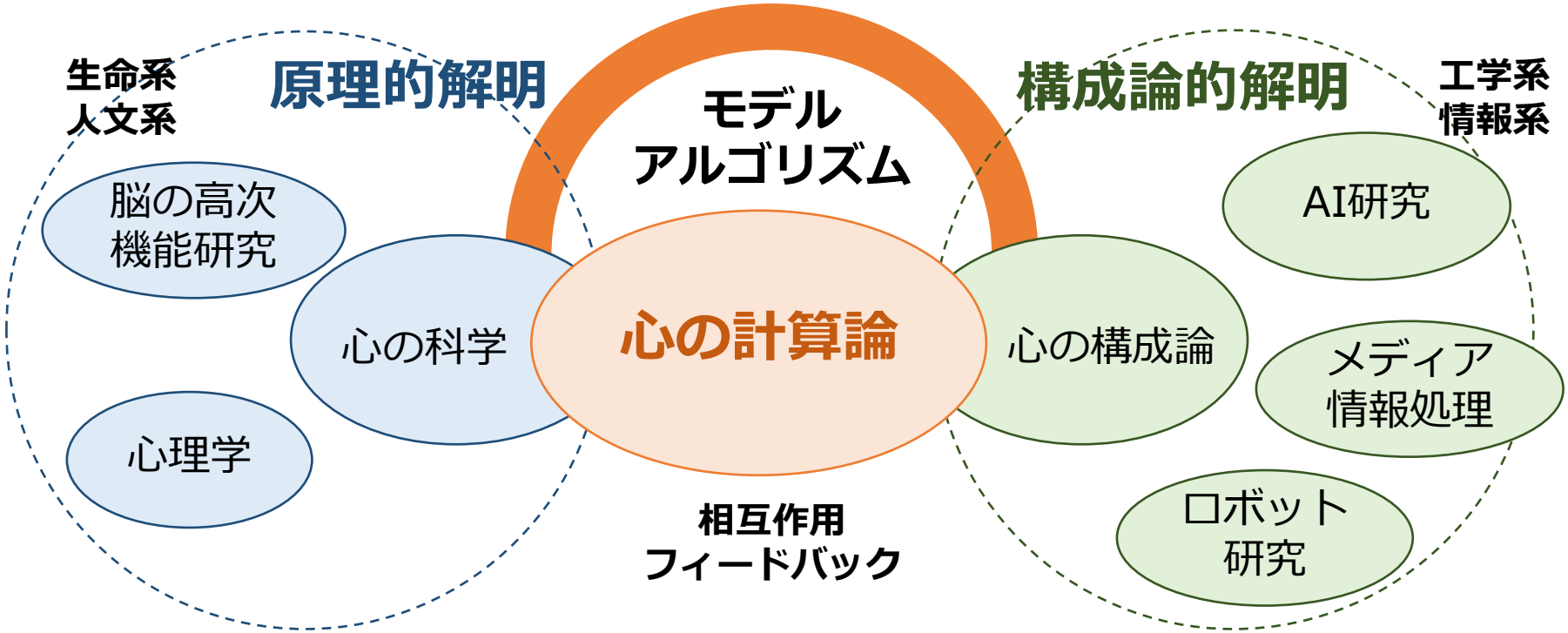
- 飯尾 尊優（同志社）
- 品川 政太朗（NAIST）

◆パートタイマー・研修生

- 中村 泰貴
- 金﨑 翔大
- 山本 賢太
- 植田 暢大
- 稲積 駿
- 山﨑 康之介
- 渡邉 寛大
- 大戸 康隆

理研ガーディアンロボットプロジェクト

◆人間の認知機能を中心とするこころのメカニズム（認識、記憶、思考、注意、感情、知能、動機）を計算論的に解明し、ロボット実装を通じて構成論的に実証する



生活空間におけるロボット

Astro, Amazon　　HSR, Toyota　　Pepper, Softbank　　Stretch, hello-robot

これらのロボットに次に期待されるのは
言語・対話インタフェースを備えつつ
我々の生活を豊かにすること

なぜ言語か？

◆情報システムに人間の意志を伝え、仕事をさせるために、プログラミング言語や種々の約束を持った指令、ハードウェア操作など、いわゆるヒューマンインタフェースが開発されて来ているが、**人間の意志を伝える道具としてもっとも大切なものが自然言語**であることは間違いない。言語はすべての人が特別な訓練を経なくても使え、自分の意志を非常に微妙な所まで伝えることができる道具である。
「**自然言語処理（編: 長尾 真） まえがきより**」

◆**これは真でもあり偽でもある**

- わざわざ言わなくても手伝って欲しい
- 言ったことは確実に手伝って欲しい

なぜ言語か

◆人、人のこころ、人の意識に興味がある

- こころ、意識とは？
- 人が「知能」と呼んでいるものは何か？

◆言語はコミュニケーション・知識構築のツールとして発展

- 高度な言語運用能力は知能の働きと明確な関係がある
- 他人の知恵を「言語」として積み重ねることが知能の本質
  - **巨人の肩に立つ**
- コミュニケーションを通じた情報の伝達・蓄積・推論

◆ロボットと自然言語

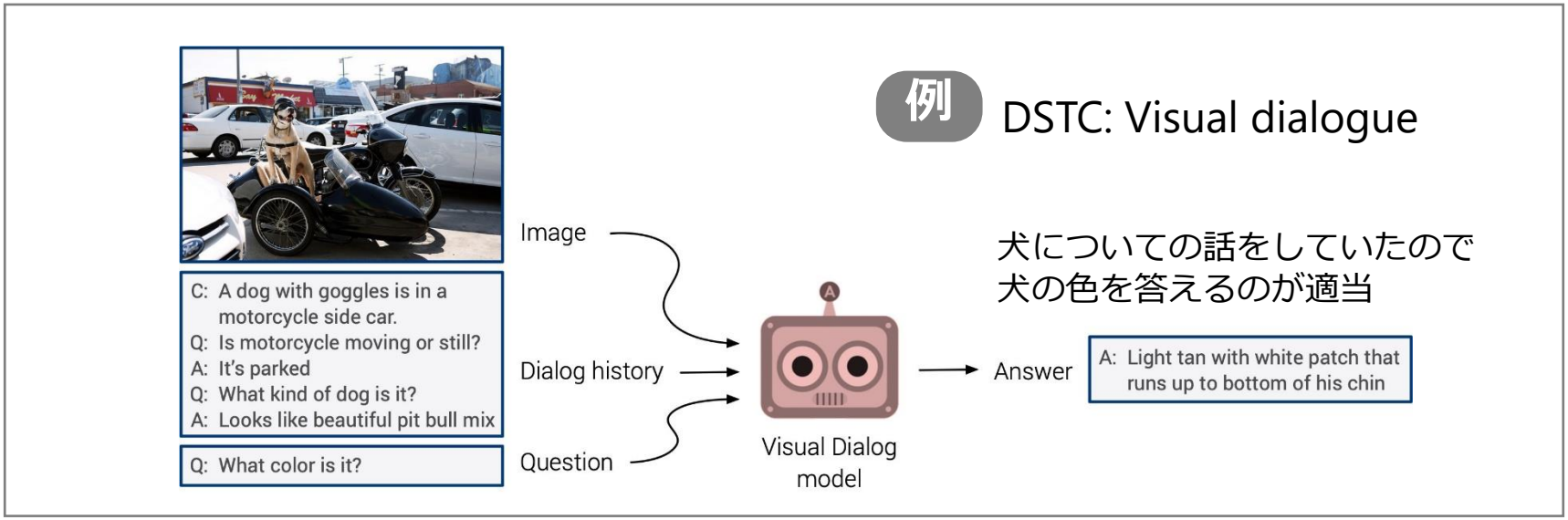
- 現在のロボットには「経験を蓄積して利用する能力」「他者の経験・知識を利用する能力」が足らない
  - 自身の経験を言語化（符号化）できていない
  - 他者から与えられた説明から学ぶことができない

言語処理と実世界処理

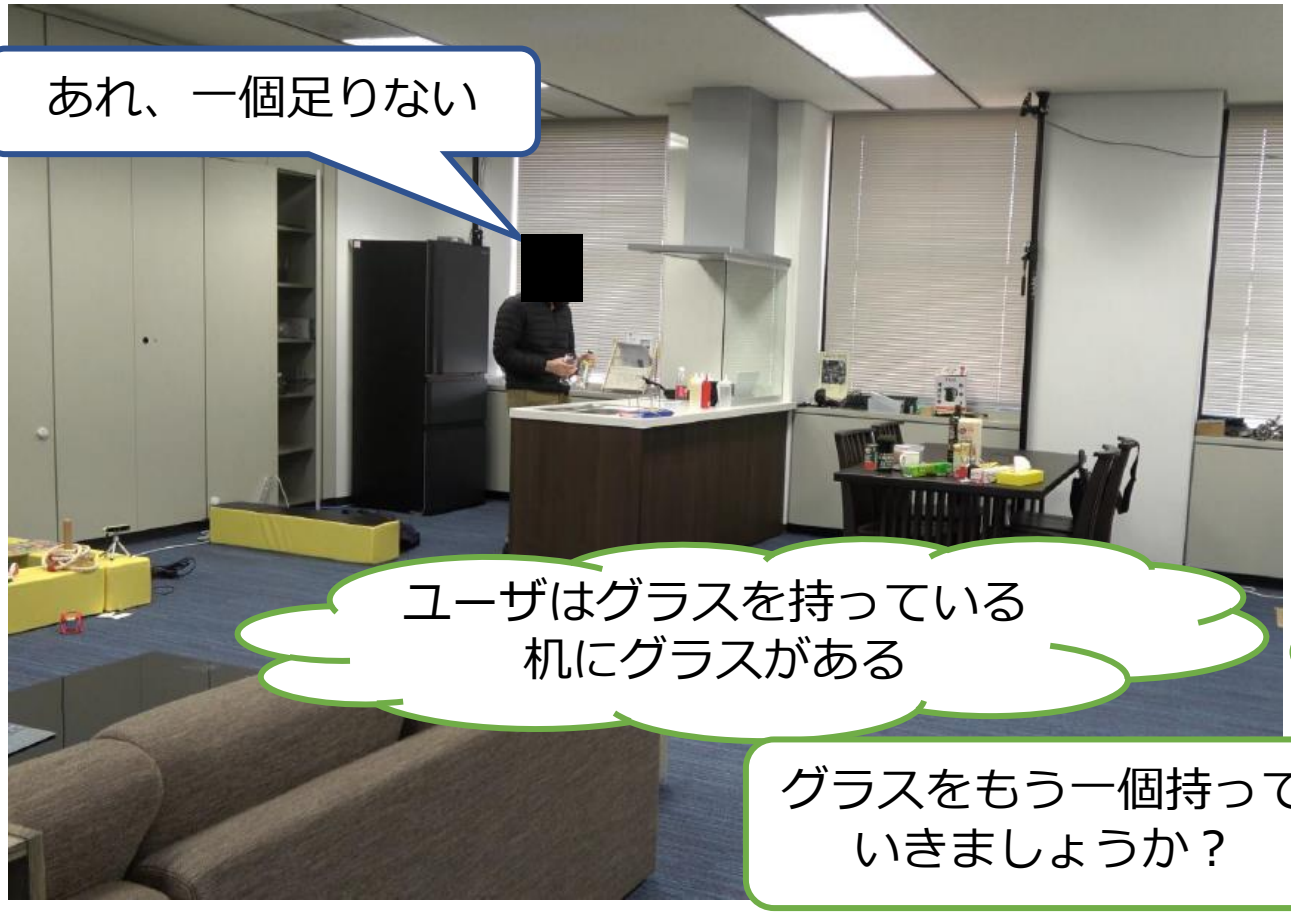
◆人間との対話でのコンテキストを理解しなければならない

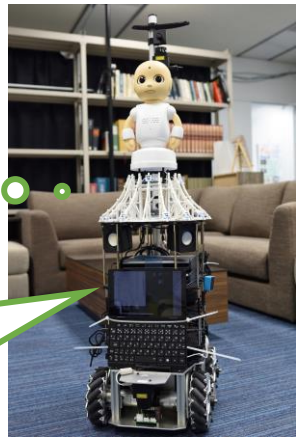
◆対話コンテキストとは「何を話したか」と「どういう状況にあるか」の両方を指す

- 話した内容に応じて答えるべき内容が変わる
- 見ているものに応じて話すべき内容が変わる



状況を理解する





×言語を理解して欲しい

○空気を読んで欲しい

気の利いた行動

◆Reactive

- behave in response to what happens to them, rather than deciding in advance how they want to behave
- ユーザに言われたことをする行動（従来の家庭内ロボット）

◆Proactive

- intended to cause changes, rather than just reacting to change
- ロボット・システムが主導権を持ってユーザに提案する

◆Reflective

- thinking deeply about something
- ユーザのことを慮ったユーザに言われた以上の内容を含む

関連研究: SayCan

◆Google が発表した家庭内ロボット [https://say-can.github.io/]

◆「テーブルを拭いて」などの要求に対して言語モデルと行動評価関数を用いてロボットの行動系列を導く

- スポンジを見つける -> スポンジを持ってくる -> テーブルを拭く

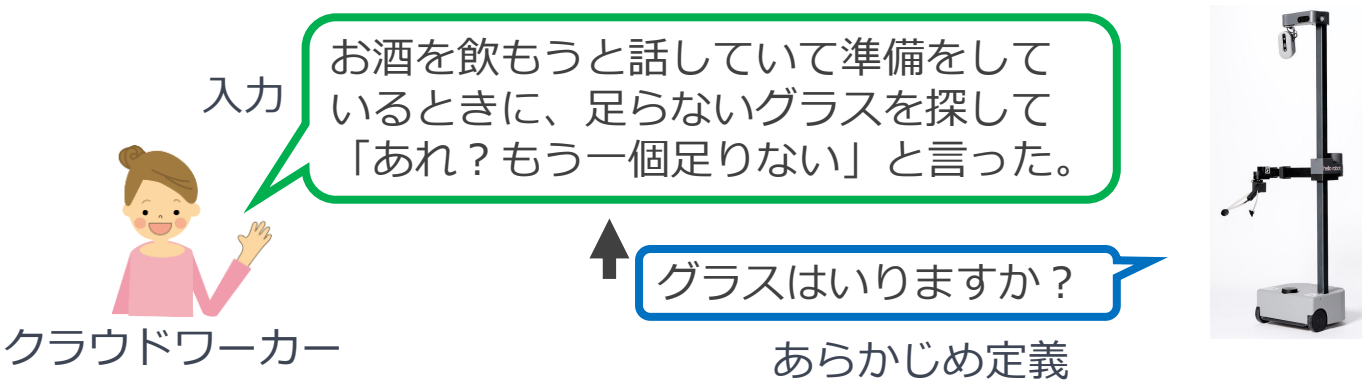
◆ユーザがロボットに対して明確に行動系列を命令しないというタスク設定は本研究と類似している

◆本研究はより曖昧な要求 (そもそも命令しない) に焦点を絞る

データ収集

◆どう気の利いた行動を収集するか？

◆ワーカーにロボットが定義された行動をとっている
動画を教示し、その行動が気が利いていると
みなせるような先行発話と状況を入力してもらう



ワーカーに見せた動画

◆**ロボットは Stretch RE1 を使用**

- カメラとロボットアームを備えている家庭用モバイルマニピュレータ
- ロボットアームの耐荷重は 1.5 kg

◆**ロボットの行動カテゴリは Stretch RE1 が実行可能な行動に基づいて定義**

カテゴリ	カテゴリ数
グラスを持ってくる、コップを片付ける、ゴミをゴミ箱に捨てる、など	40

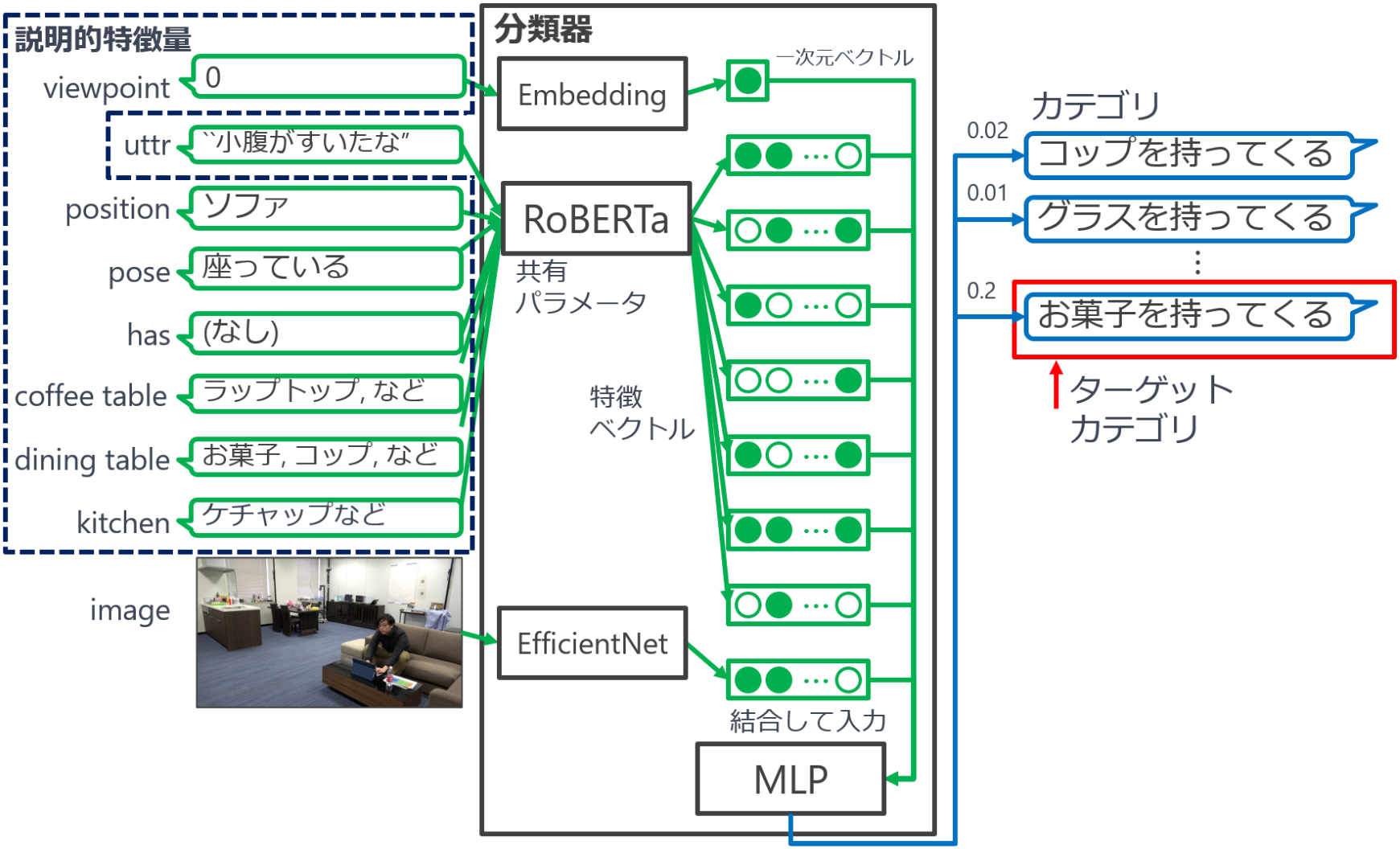
◆**カテゴリごとに10対話、合計400対話を収集**

作成したデータの例

状況の認識

◆ロボットの視点番号、ユーザの位置、ユーザが持っているもの、キッチンにあるもの、など状況を説明するラベルを人手で付与

気の利いた行動の選択


説明的特徴量
viewpoint
0
uttr
``小腹がすいたな"
position
ソファ
pose
座っている
has
(なし)
coffee table
ラップトップ, など
dining table
お菓子, コップ, など
kitchen
ケチャップなど
image
分類器
Embedding
一次元ベクトル
RoBERTa
共有
パラメータ
特徴
ベクトル
EfficientNet
結合して入力
MLP
カテゴリ
0.02
コップを持ってくる
0.01
グラスを持ってくる
0.2
お菓子を持ってくる
ターゲット
カテゴリ

実験設定

◆事前学習された日本語 RoBERTa と EfficientNet を利用

◆400本のデータを五分割交差検証にて評価

◆各分割データについて10回実験を試行

◆評価指標

- Accuracy
- R@5: 正解カテゴリが上位5位以内に含まれている割合
- Mean Reciprocal Rank (0 < MRR <= 1):

$$MRR = \frac{1}{|U|} \sum_{u \in U} \frac{1}{rank_u}$$

$rank_u$: 正解カテゴリの順位

$|U|$: テストデータ中のユーザ要求

実験結果

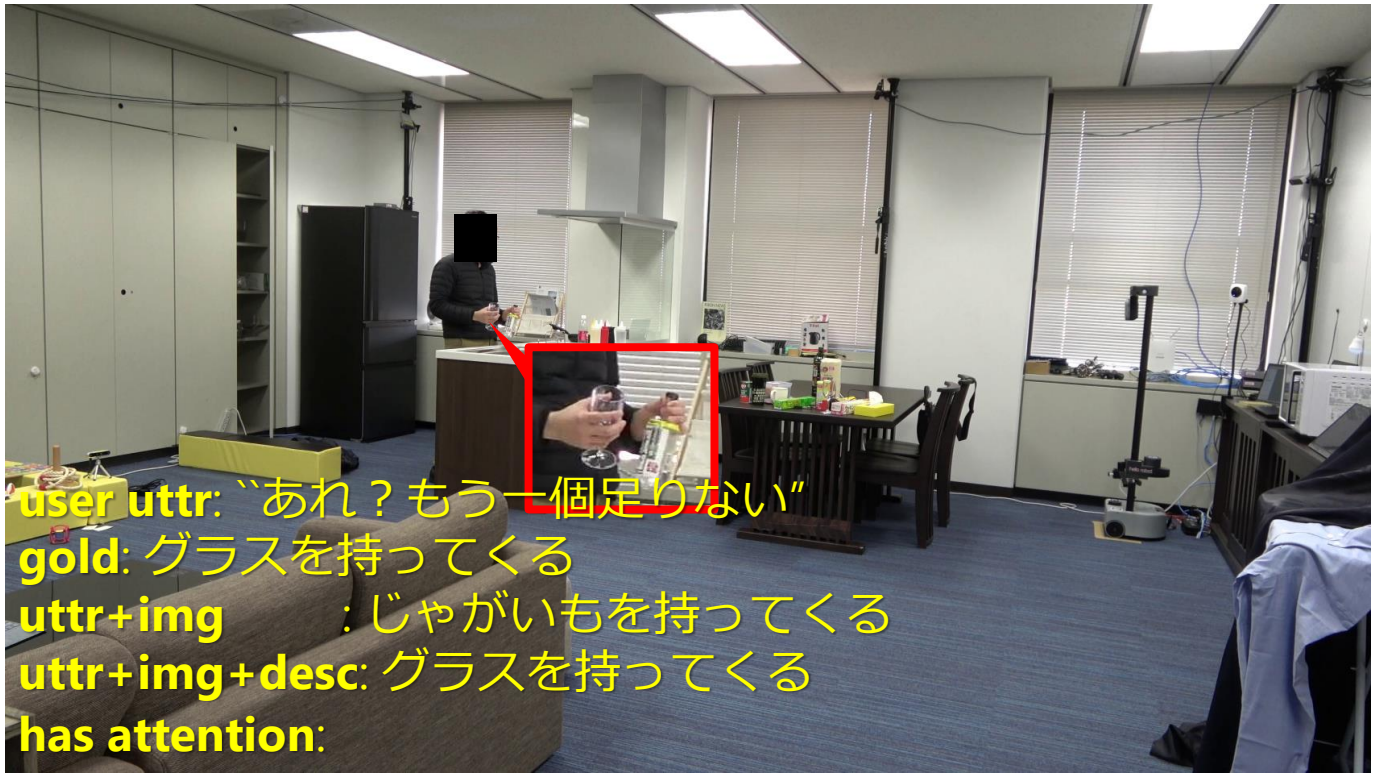
モデル	Accuracy (%)	R@5 (%)	MRR
uttr	**27.02	**53.85	**0.4054
uttr+img	**27.23	**54.50	**0.4064
uttr+img+desc	**63.58**	**87.12**	**0.7417**

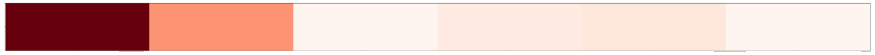
uttr: ユーザ発話のみを入力とした場合
uttr+img: ユーザ発話と画像を入力とした場合
uttr+img+desc: ユーザ発話と画像に加え説明的な特徴量も入力とした場合

- どの指標についても画像から得られる説明的な特徴量も利用したほうが性能が劇的に向上
- こうした気の利いた行動の選択に有効な特徴量ラベルを画像から抽出することが重要

→ 状況理解として何を用いるかが重要

分類成功の例と注視重み





◆説明的な特徴量を利用することで、ユーザ発話と画像のみでは選択できなかった気の利いた行動を選択可能

分類成功の例と注視重み

◆説明的な特徴量を利用することで、ユーザ発話と画像のみでは選択できなかった気の利いた行動を選択可能

分類成功の例と注視重み

◆説明的な特徴量を利用することで、ユーザ発話と画像のみでは選択できなかった気の利いた行動を選択可能

分類失敗の例

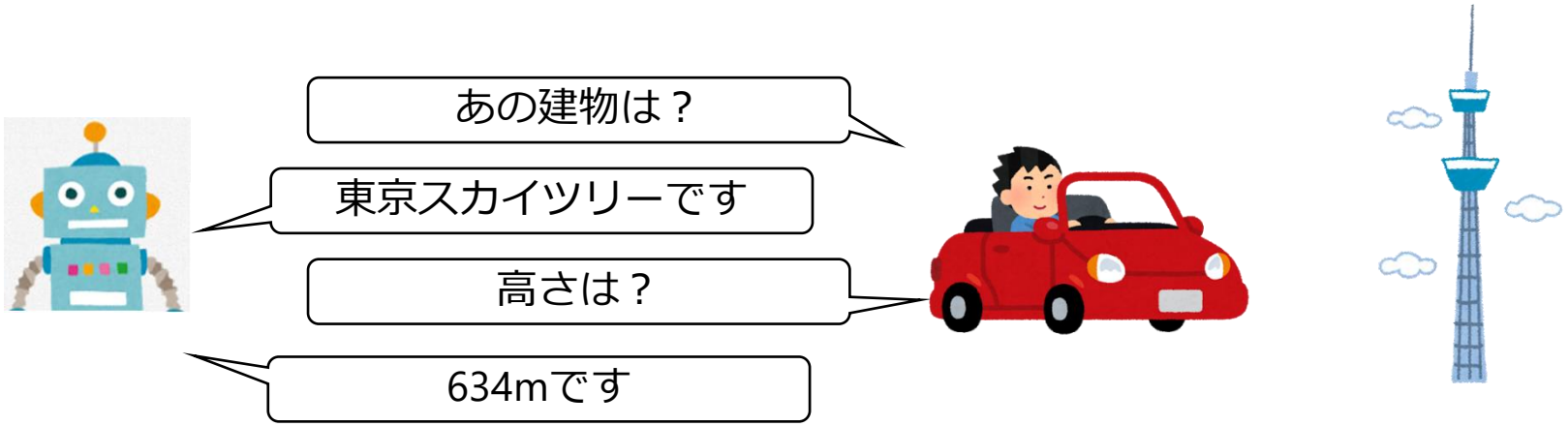
◆適切に注視を当てられていない場合もある

◆因果推論などの枠組みでより正確に状況と気の利いた行動を結びつける仕組みが必要

ロボットに求められること

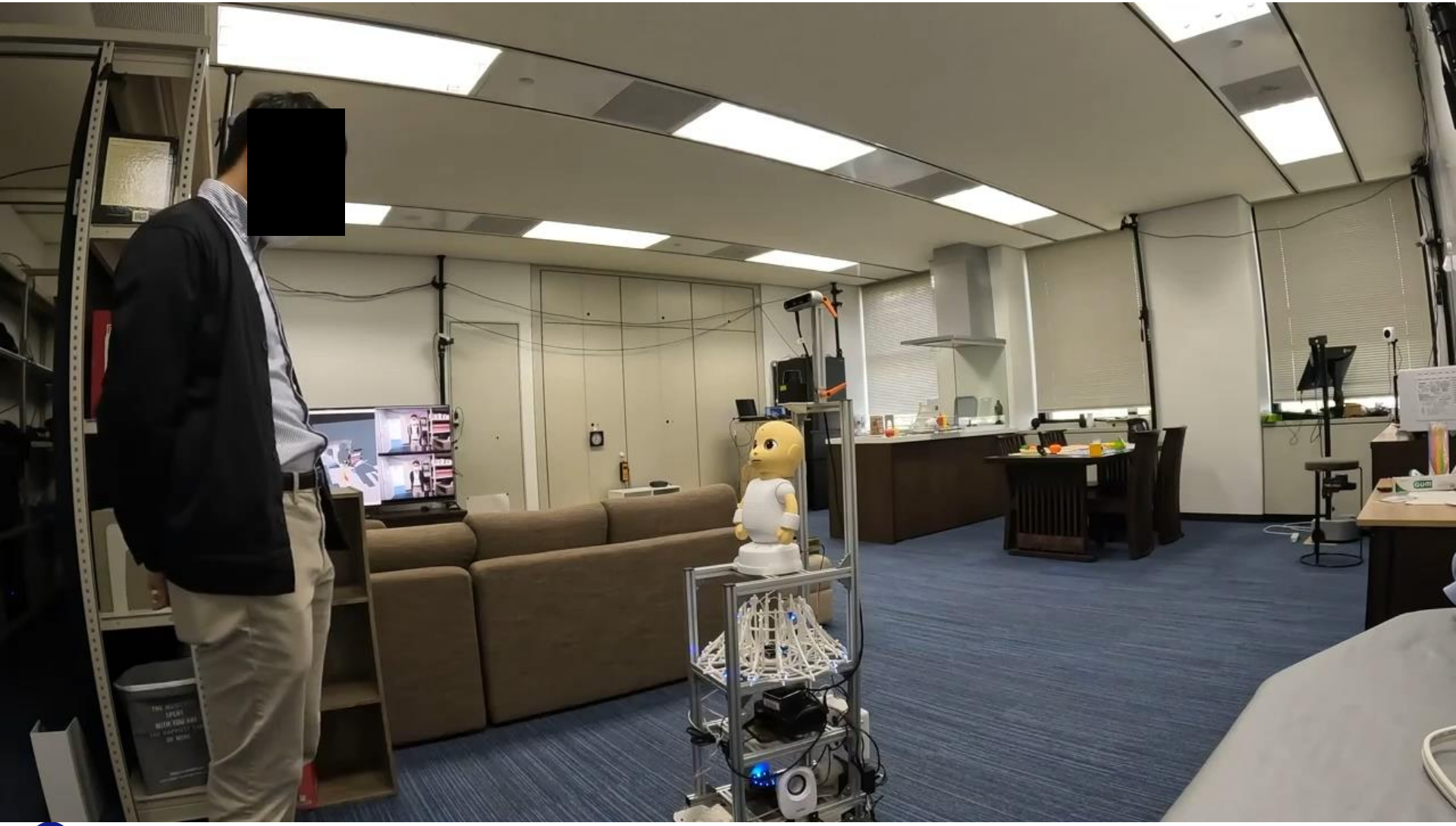
◆言語を含む周囲の状況を適切に理解すること

- 対話コンテキストと実世界の事物との「接地」が必要
- 観測したものの何について話しているか「選択」が必要

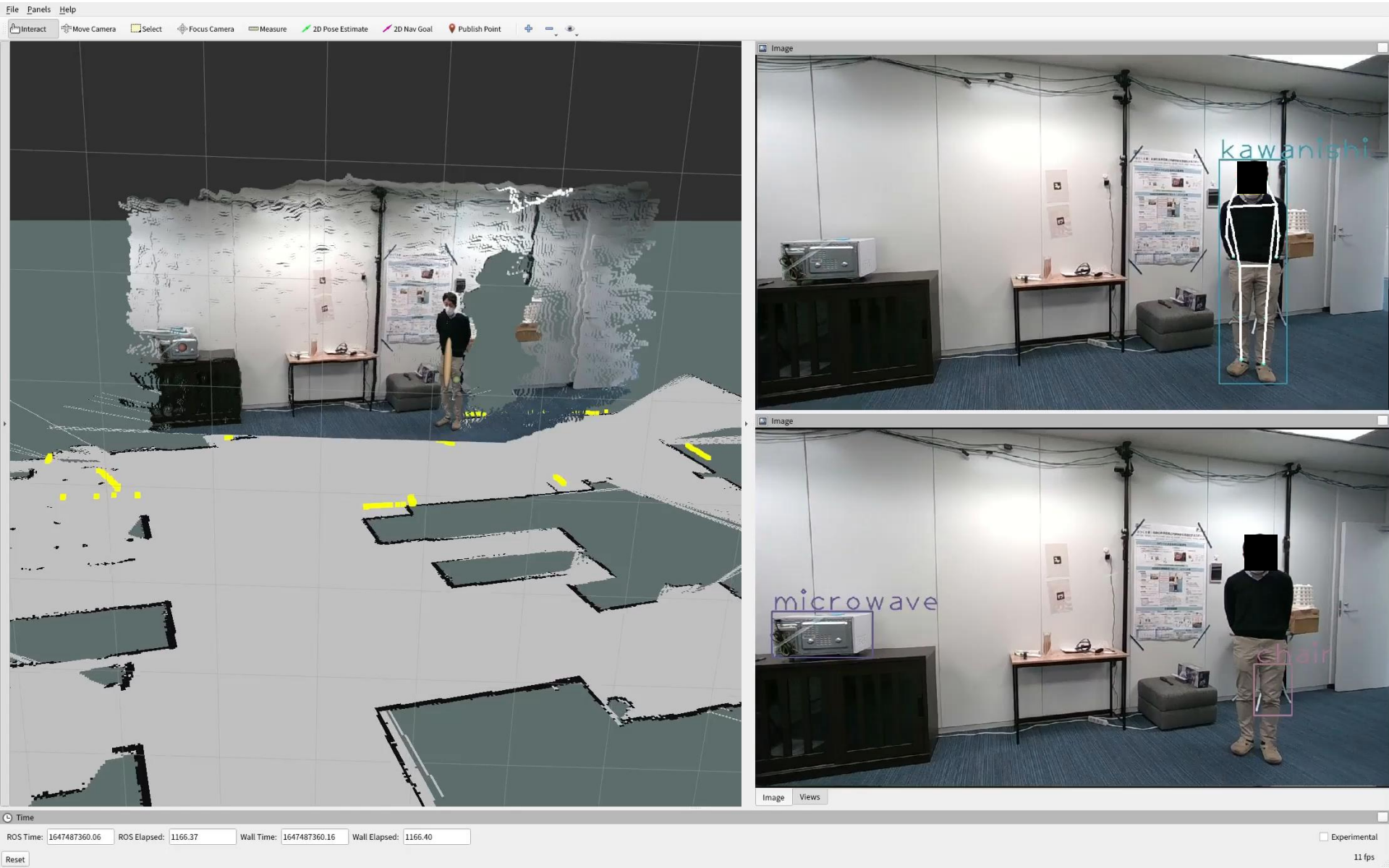


言語+マルチモーダル情報を使った状況の理解

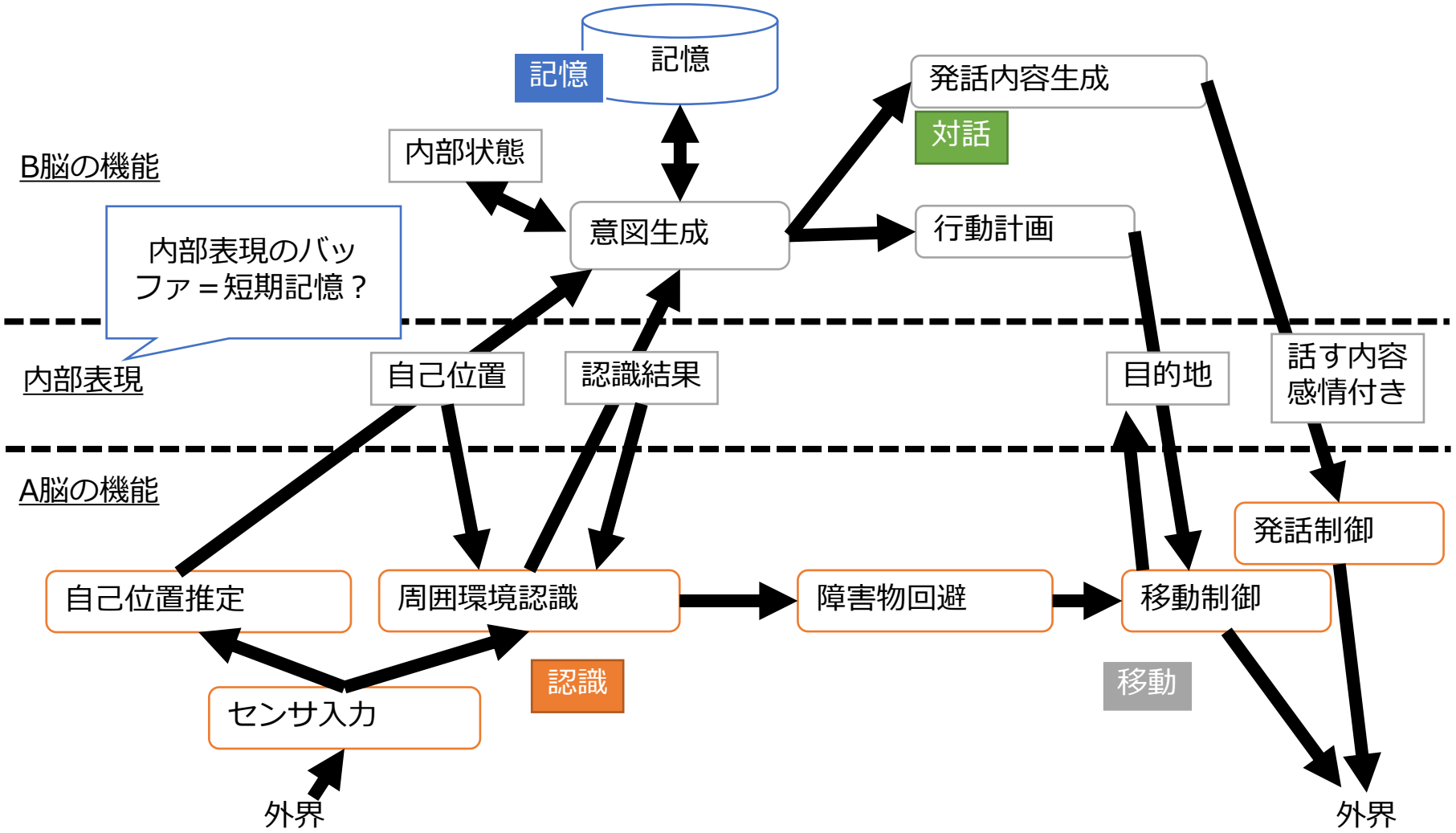
周囲の状況の理解



一人称視点での認識結果



ロボットの認識・記憶・行動



理解と言語

◆いかに理解結果を説明するか？

- 自身の動作や周囲の状況を言語表現で記述する

支援動作前の状態 (A)

支援動作後の状態 (B)

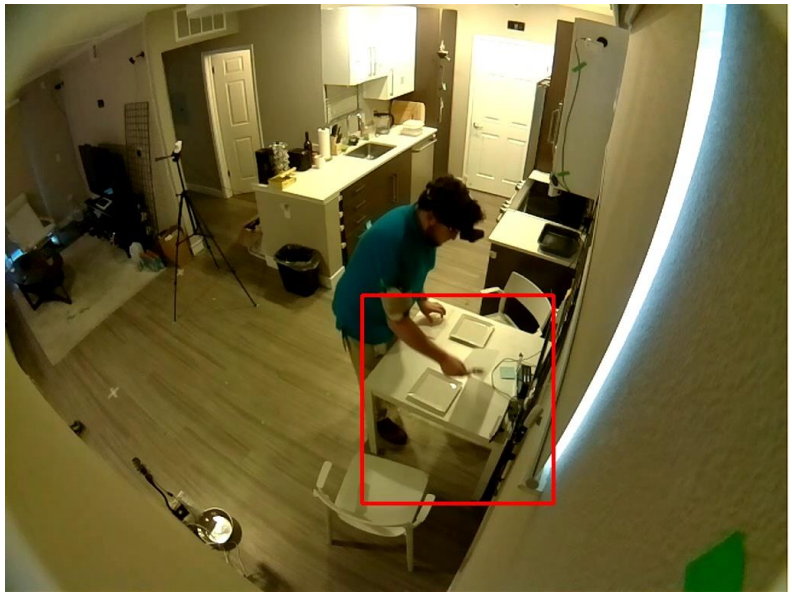
説明: **トイレの便座の蓋を開ける**

理解と言語

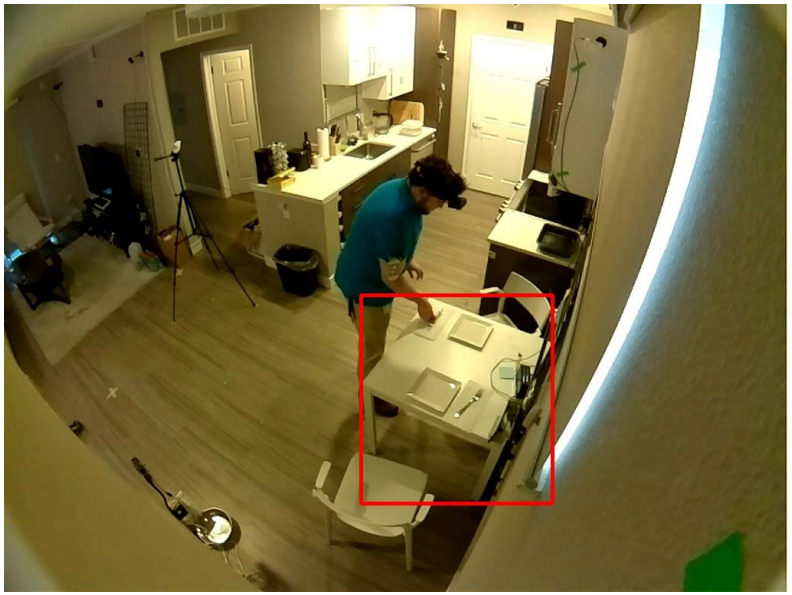
◆三人称視点でも同じ

- ロボットは家庭内環境システムの一部とみなしてよい

支援動作前の状態 (A)



支援動作後の状態 (B)



説明: **フォークをテーブルの上の皿の脇に並べる**

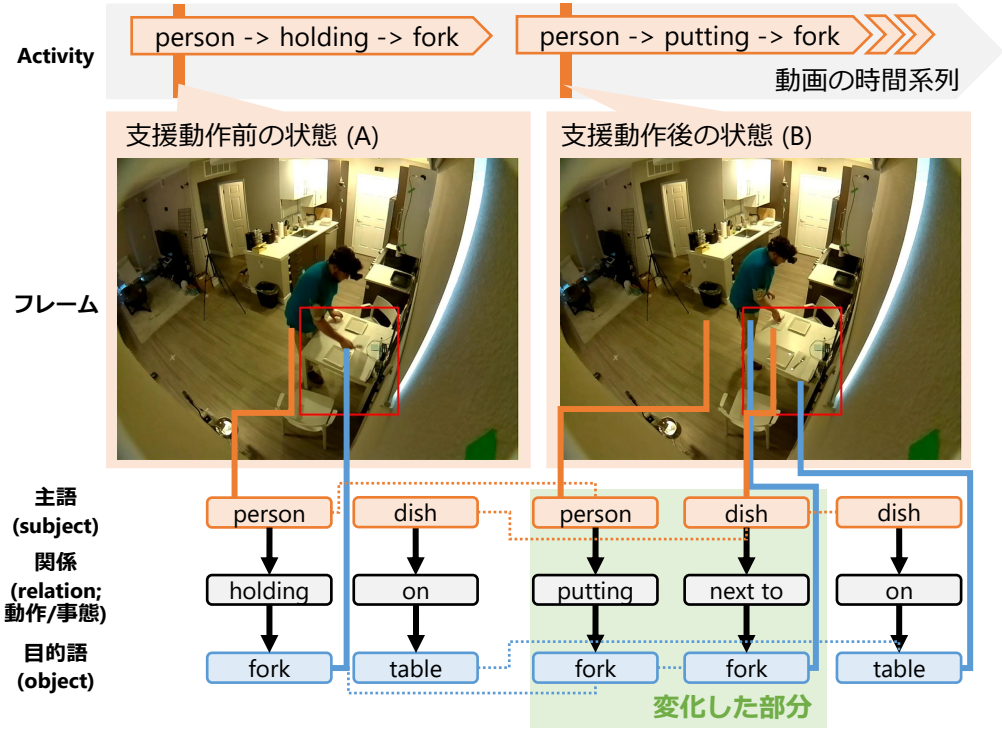
シーングラフの利用

◆画像中のイベント（事態）をグラフ形式で記述

- 画像中の情報を主語-関係-目的語 の triple で表現
- 言語とシーングラフの関係

◆どの状況を理解することが重要か？

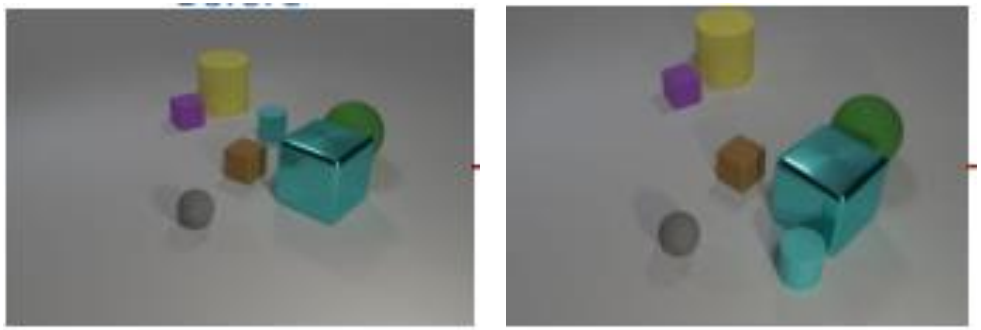
- おそらく**シーングラフが変化した点**が重要
- 事態に変化が生じた
- その変化を説明する



関連研究: Diff の説明

◆DUDAモデル

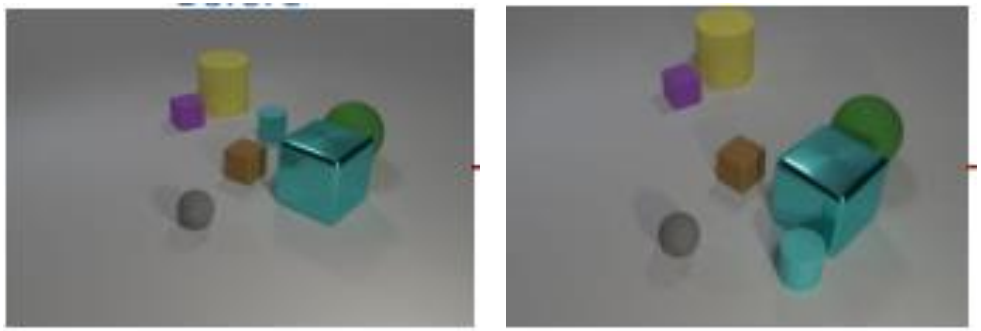
- 2枚の画像中のどこが変化したかを説明するモデル
- 変化点に着目した画像の埋め込み表現を作成



小さい円柱の位置が変化した.

◆本研究: 変化のイベントを捉える

- ロボットが捉えるべきは変化の事態
- DUDAと類似するモデルを利用可能

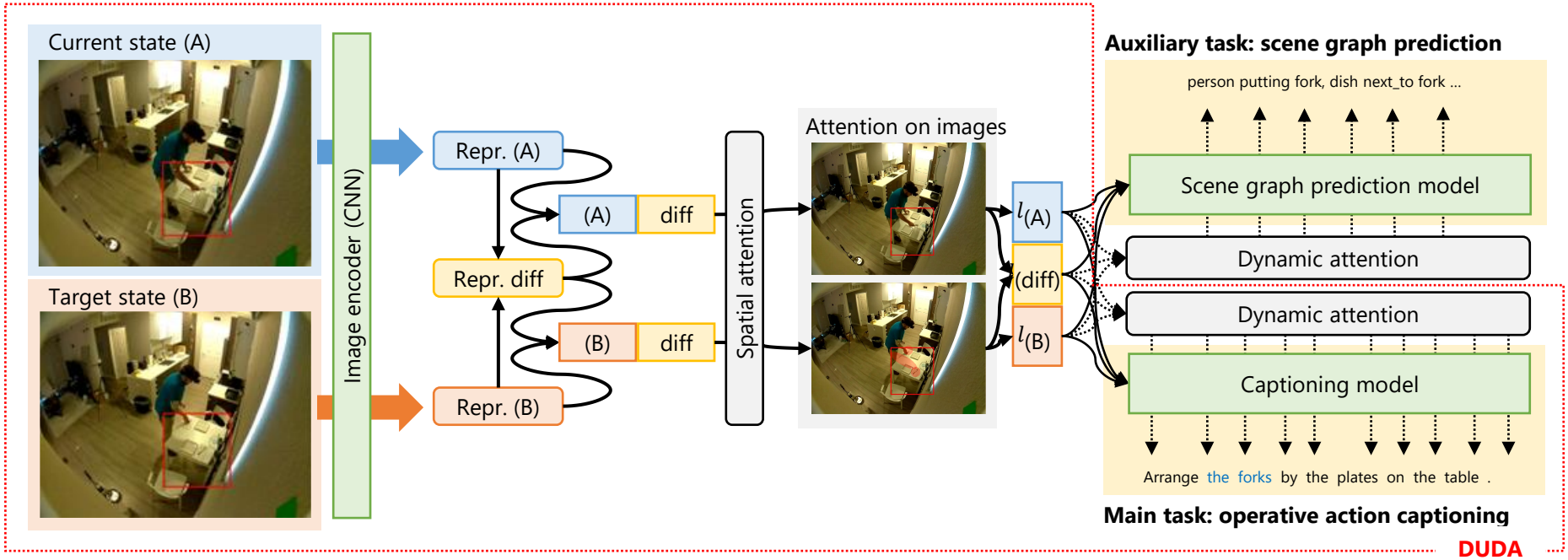


小さい円柱を立方体の前へ移動させる.

動作行動予測モデル

◆DUDA を拡張した変化に関連する事態を捉える仕組み

- 変化に関するのエンコーダを利用することで着目点を決定
- 補助タスクとしてシーングラフ予測を利用



動作行動予測の評価

◆ベースラインはシンプルなDUDA
- シーングラフを同時予測する場合を追加
- 交互に言語生成とシーングラフ予測を学習するか
- シーングラフ全体を予測するか差分を予測するか

◆自動評価: BLEU, ROUGE-L, CIDEr

モデル	実験条件		BLEU				ROUGE-L	CIDEr
			1	2	3	4		
ベースライン	-		0.389	0.238	0.151	0.0998	0.375	0.871
＋シーングラフ	交互学習 (1.0)	全体	0.350	0.194	0.113	0.0686	0.330	0567
		差分	0.359	0.205	0.126	0.0815	0.339	0.649
	交互学習 (0.9)	全体	0.396	0.244	0.156	0.105	0.383	0.921
		差分	0.392	0.245	0.158	0.107	0.389	0.913
	線形補完 (0.9)	全体	**0.405**	**0.260**	**0.167**	**0.114**	**0.392**	**1.001**
		差分	0.396	0.246	0.160	0.109	0.387	0.971

動作行動予測の人手評価

◆人手評価で内容の適切性だけでなく流暢性も向上

- ●事態の理解にシーングラフの予測が有効

モデル	実験条件		流暢性	内容
ベースライン	-		3.89	3.05
＋シーングラフ	線形補完 (0.9)	全体	4.42	3.45
		差分	**4.53**	**3.48**

支援動作前の状態 (A)

支援動作後の状態 (B)

ベースライン
バスケットの中の服を**バスケットに移した**
シーングラフ（全体）
バスケットに入っている服を**洗濯している**
シーングラフ（差分）
バスケットの服を**取り出している**

予測例

支援動作前の状態 (A)	支援動作後の状態 (B)	
		ベースライン ボウルに**野菜の入った食器**を入れる シーングラフ（全体） ボウルにサラダを**入れている** シーングラフ（差分） ボウルの中から食べ物を**出している**
		ベースライン 服を物干しに**かけてかけている** シーングラフ（全体） 服を干す シーングラフ（差分） 服を**物干しに**干す
		ベースライン **IH機**を洗っている シーングラフ（全体） 食洗機にコップを入れた シーングラフ（差分） 食洗機の**かごを引き出している**

動作軌跡と言語

◆入力: ロボットが一人称視点で観測可能なもの

- ●各関節の動作軌跡
- ●一人称視点からえられる動画情報

◆出力: 動作を説明するテキスト

- ●自然言語による動作結果の説明

正解	床のティーポットを取る	
A: E2E	てててててててて	e:9
B: 分節化	床のソースを取る	a:7, b:1, c:1
C: 注意機構	テーブルのソースを取る	b:1, c7, d:1
D: B+C	床にあるソースを取る	a:7, c:2

a: 正解、b: おおよそ正解、c: 間違いがある、
d: 文法は正しい、e: 意味不明

言語を使った状況の理解
現在地と課題

◆**様々な入力を言語で説明することが可能に**
- ロボットが持つマルチモーダル性の解釈
- 言語に結び付ける＝知識に結び付けるための入り口

◆**大規模事前学習モデルを用いた理解の拡張**
- 数少ない fine-tuning data から理解モデルを構築可能
- 大規模 LM の場合、そもそもテキストに書かれていないことは出てこないのでは？

◆**言語から動作（動作・観測から言語）**
- シンボル化・離散化の方がタスクとしては容易
- 言語からの生成タスクは**言明されない部分**がある
  - e.g., 音声認識と音声合成

真に言語と動作を結び付けるために

◆**言語に明に暗に含まれる部分の理解**

- ●言語が持つ裏の（あれば）意図の理解
- ●常識的知識の利用
- ●周囲の状況の利用
- ●特に言語を動作化するとき
  - ●どのように肉付けを行うのか
  - ●どのように個々のロボットの身体性に合わせるのか
  - ●**「もっと優しく持って」を実現できるか**

◆**物理世界における制約の利用**

- ●シミュレータだけでは物理的制約が満たせない
- ●実世界だけでは学習データが足りない
- ●タスクに必要なシミュレーションの粒度