

2026年2月11日

AI最前線：生成AIからフィジカルAIへ

藤吉弘亘（中部大学・理工学部AIロボティクス学科）

<http://mprg.jp>



MPRG

MACHINE PERCEPTION AND ROBOTICS GROUP

学歴：

- 1988年 岐南工業高校電子科卒業
- 1992年 中部大学電子工学科卒業
- 1994年 中部大学大学院修士課程修了
- 1997年 中部大学大学院博士後期課程満期退学（博士）

研究活動：

- 1997年 米カーネギーメロン大学ロボット工学研究所ポスドク研究員（3年）
- 2000年 中部大学工学部講師
- 2004年 中部大学工学部准教授
- 2005年 米カーネギーメロン大学ロボット工学研究所客員研究員（1年）
- 2010年 中部大学工学部教授～
- 2014年 機械知覚ロボティクス研究グループ
- 現在に至る

学外活動：

- 日本ディープラーニング協会理事
- クロスアポイントメント（デンソー）

J-WAVE 81.3FM  INNOVATION WORLD



vol.162 藤吉弘亘（中部大学）「+AI」で変わる未来」

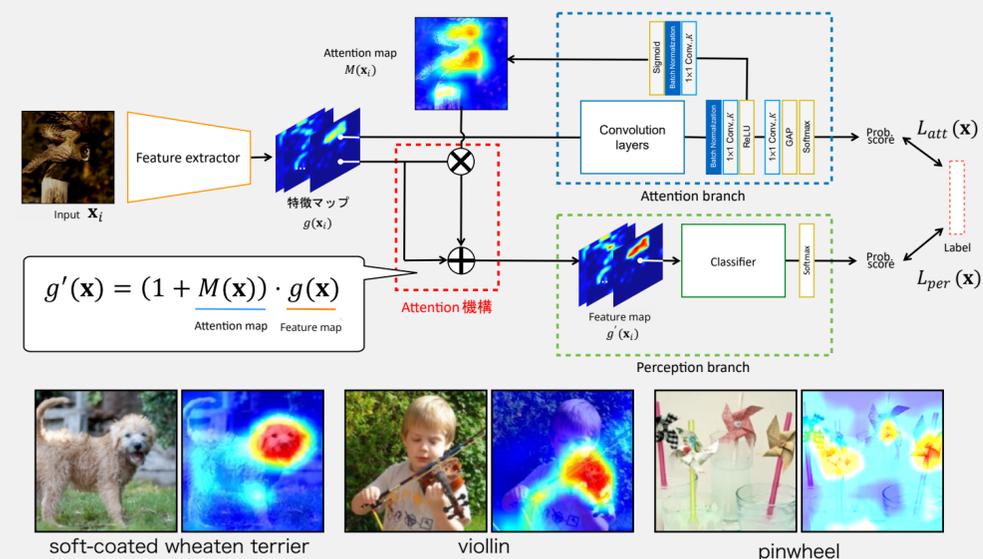


<https://www.youtube.com/watch?v=z9jEO9iVHU8>

説明可能なAI:

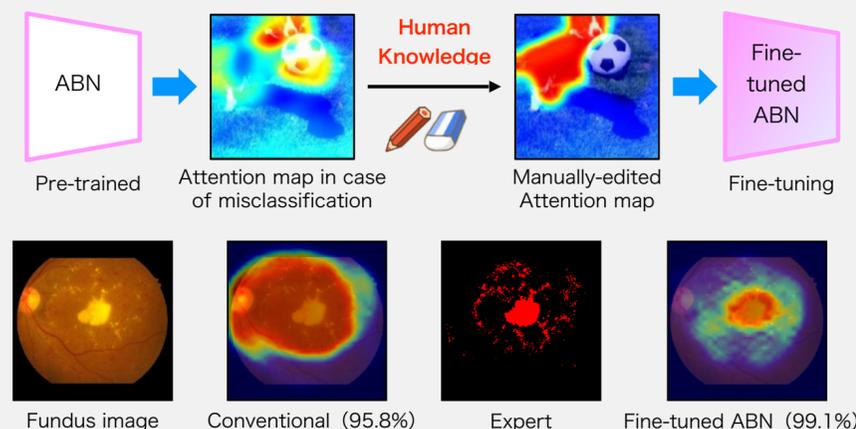
■ Attention Branch Network [Fukui+, CVPR'19]

- Attention Branchを統合して注意マップを生成することで、分類精度と説明可能性（解釈性）を向上



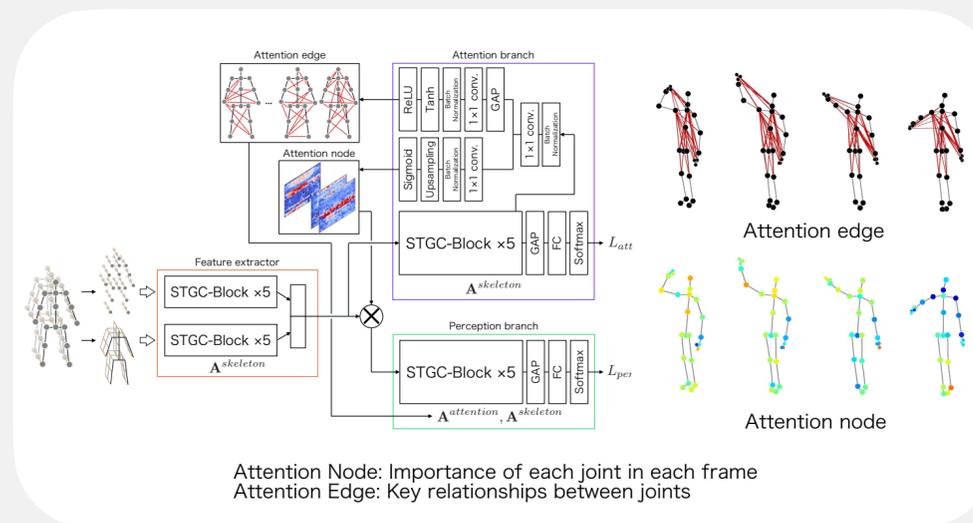
■ Embedding Human knowledge into DNN [Mitsuhashi+, VISSAP'21]

- 人の知見を深層学習モデルに組み込むことで、視覚的説明（可視化・解釈）の質と分類精度を向上



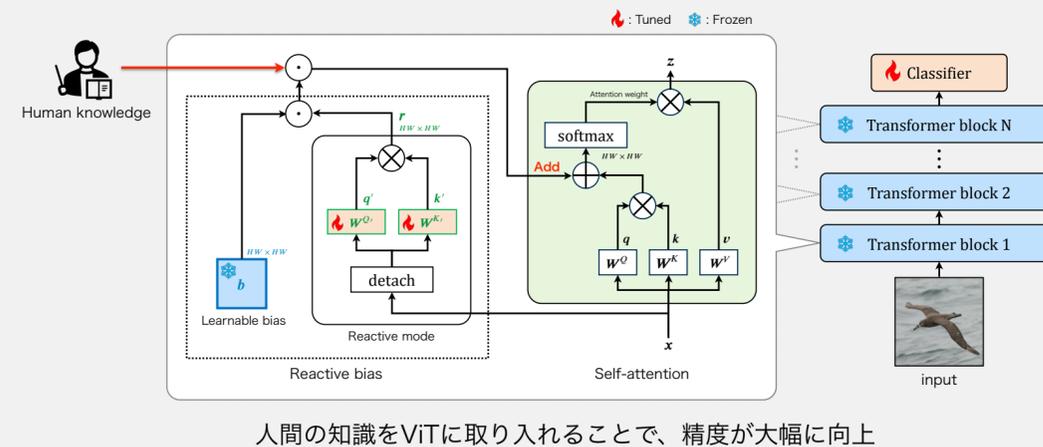
■ Spatial Temporal Attention Graph [Shiraki+, ACCV'20]

- 関節の重要度と関係性を考慮したグラフ畳み込みニューラルネットワーク



■ Reactive Bias for ViT

- Reactive Bias を通じて、人間の知識を Vision Transformer (ViT) に埋め込む

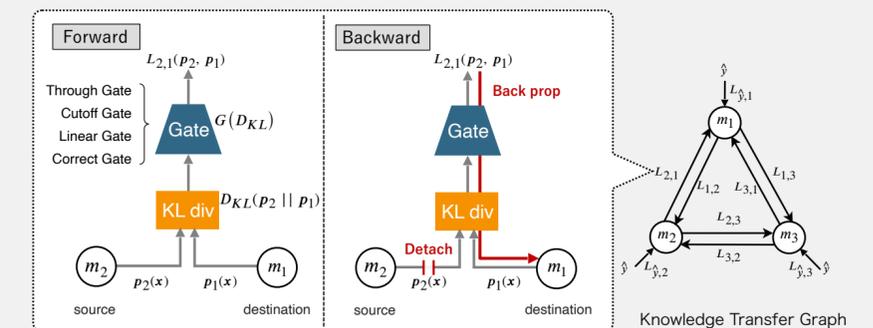


人間の知識をViTに取り入れることで、精度が大幅に向上

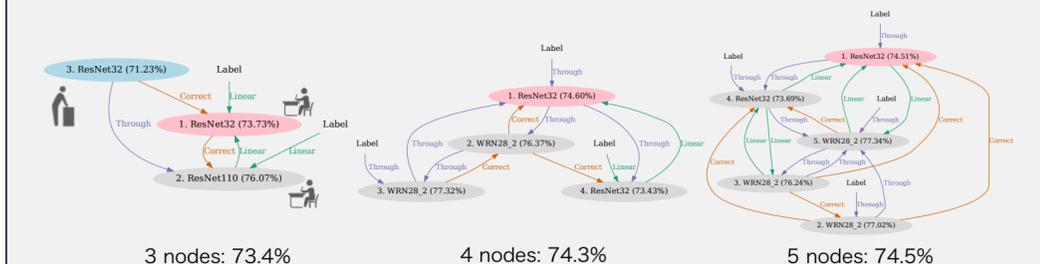
Deep Collaborative Learning:

■ Knowledge Transfer Graph [Minami+, ACCV'20]

- モデル間の知識転移の流れを可視化するグラフ構造表現であり、効率的な転移学習と協調的な深層学習を促進

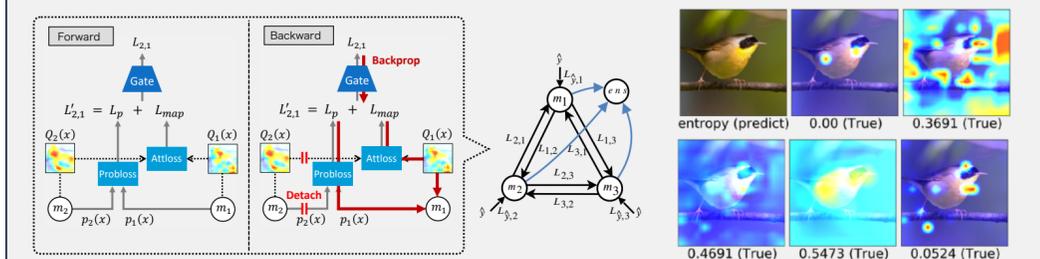


→ 知識転移グラフをハイパーパラメータ探索として最適化



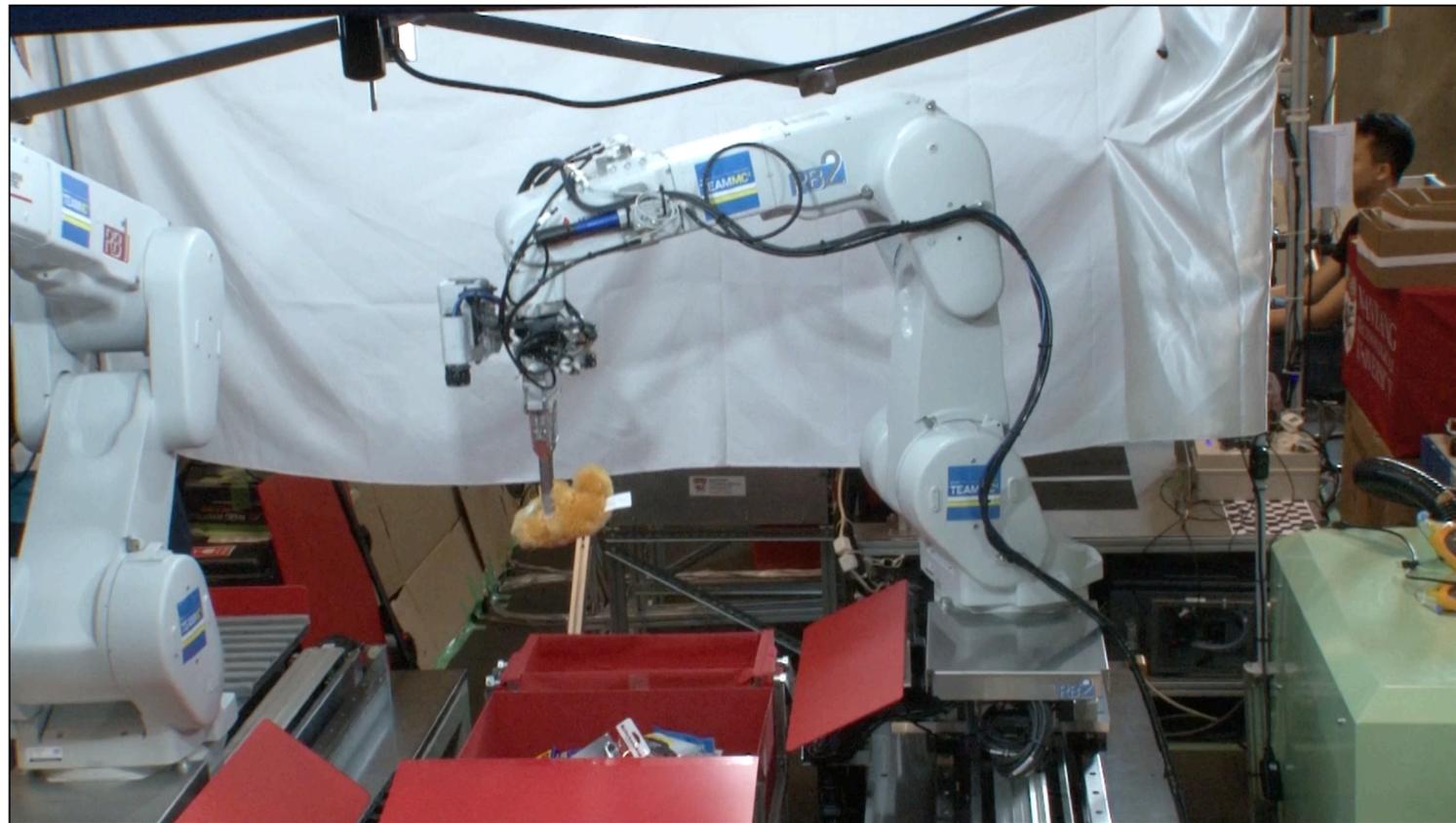
■ Ensembling Knowledge Transfer Graphs [Okamoto+, ECCV'22]

- 複数のネットワークを統合するためのアンサンブルノードを導入



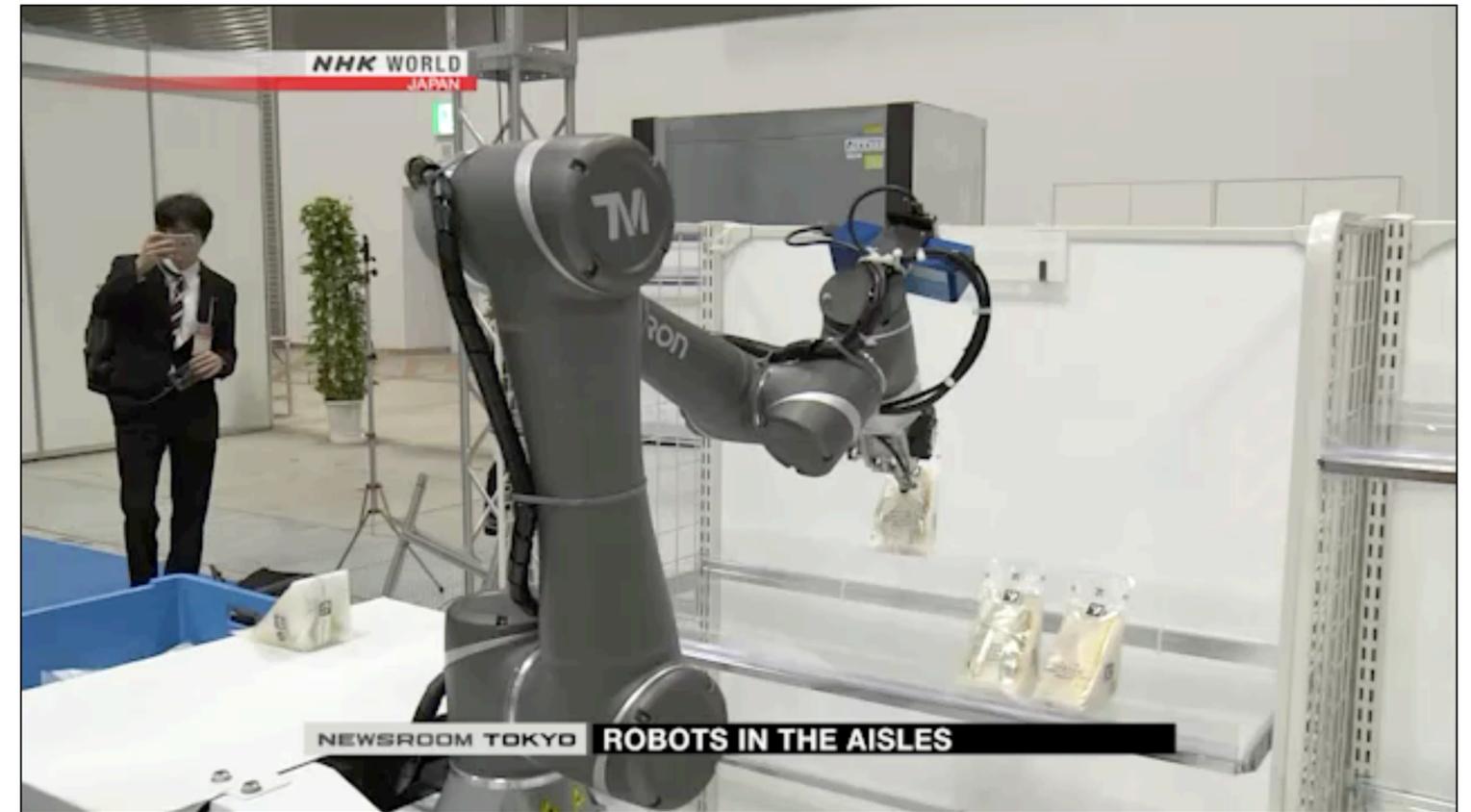
→ アンサンブル学習に適した多様な注意マップを獲得

Amazon Robotics Challenge



 Stow Task: 3rd place

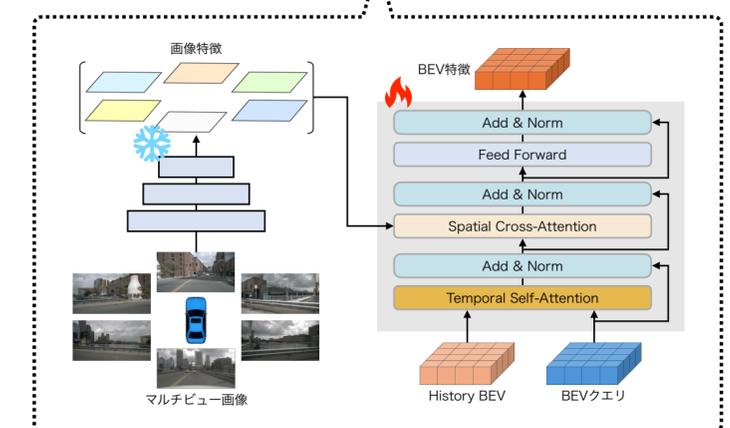
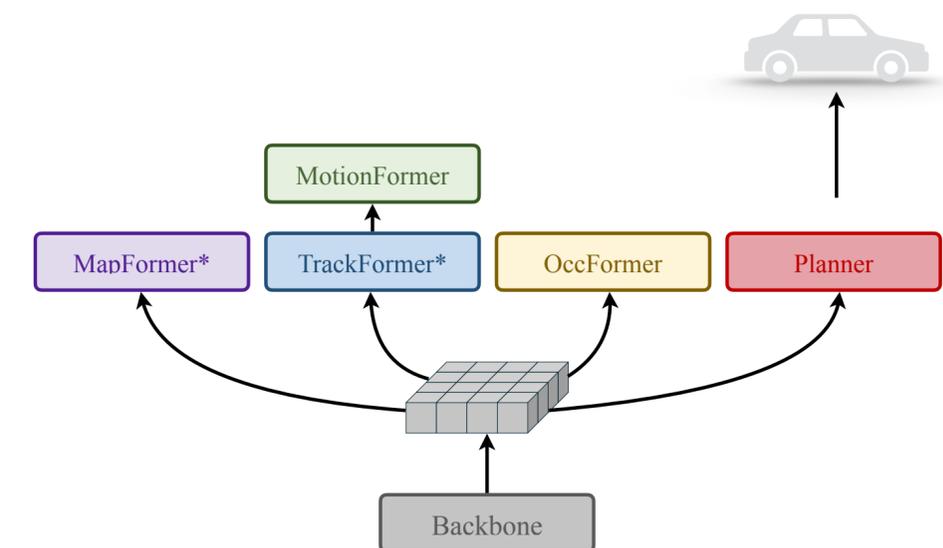
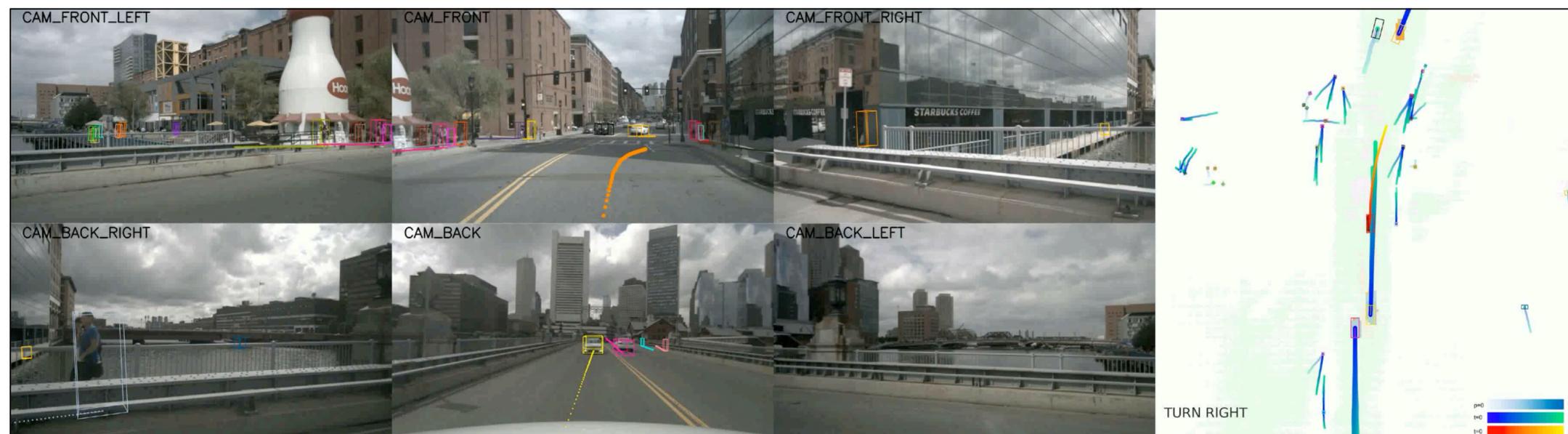
World Robot Summit



 WRS2018-FCSC: 2nd place

・ Planning-oriented 統合E2Eモデル

- BEV特徴・占有 (Occupancy) ・トラッキングなどの表現も学びつつ、最後は計画に最適化



マルチモーダルPARA-Drive

- ・ 次の文章を読んで、最後の()の中に入る単語を答えて下さい。

<文章 1> このリンゴは ()

<選択肢> ①アインシュタイン ②黄色い ③美味しい ④行く

- | | | |
|-----------|--------------------------|----------------|
| ①アインシュタイン | : × 文章としておかしい | → [正解の確率: 2%] |
| ②黄色い | : △ 文法的に正しいが違和感あり | → [正解の確率: 5%] |
| ③美味しい | : ○ 正しそう | → [正解の確率: 90%] |
| ④行く | : × 文章としておかしい | → [正解の確率: 3%] |

<文章 2> 農家の友達から普通とは違うリンゴをもらった。このリンゴは ()

<選択肢> ①アインシュタイン ②黄色い ③美味しい ④行く

- | | | |
|-----------|----------------------|----------------|
| ①アインシュタイン | : × 文章としておかしい | → [正解の確率: 2%] |
| ②黄色い | : ○ 正しそう | → [正解の確率: 55%] |
| ③美味しい | : ○ 正しそう | → [正解の確率: 40%] |
| ④行く | : × 文章としておかしい | → [正解の確率: 3%] |

言語モデル: 生成される単語・文章に確率を割り当てるモデル

- ・ 単語（トークン）列 y_1, \dots, y_T の生成確率 $P(y_i, \dots, y_T)$ を推定
 - 次単語（あるテキストに続く単語）を予測できる

$$y^* = \arg \max_{y \in V} P(y | \text{英国, の, 首都, は})$$

全単語の集合

$P(\text{東京} \text{英国, の, 首都, は}) = 0.00000043$	}	$y^* = \text{ロンドン}$
$P(\text{パリ} \text{英国, の, 首都, は}) = 0.00000082$		
$P(\dots \dots \text{英国, の, 首都, は}) = \dots$		
$P(\text{ロンドン} \text{英国, の, 首都, は}) = 0.00000103$		

計算された確率の最大値を
与える単語を選択

- ・ 翻訳前の文章を考慮することで機械翻訳への応用が可能
 - 次単語予測タスクにおける条件部分に、翻訳前の文章を追加
 - 英語→日本語： $P(\text{日本語の文章} | \text{英語の文章})$ が最大となるような日本語の文章を生成

- 「穴埋め問題」を解いて学習

<文章2> 農家の友達から普通とは違うリンゴをもらった。このリンゴは ()

<選択肢> ①アインシュタイン ②黄色い ③美味しい ④行く

①アインシュタイン : **×** 文章としておかしい

②黄色い : **○** 正しそう

③美味しい : **○** 正しそう

④行く : **×** 文章としておかしい



選択肢②が正解と予想するには:

「リンゴは通常赤色である」という文章には含まれていない知識が必要

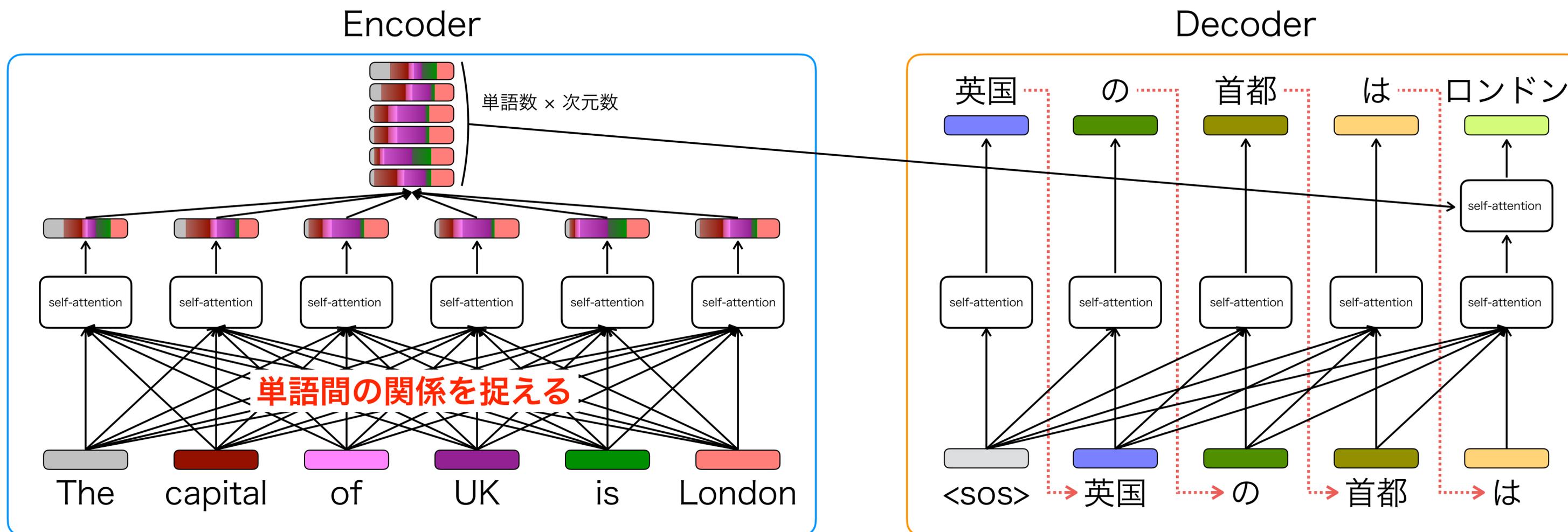
- 「穴埋め問題」を解くことで

- 大量の穴埋め問題を解く過程で、世界に関する一般的な知識や文法構造を獲得

- 自己注意(self-attention)により単語間の関係も表現可能なモデル

👍 Pros: 長距離依存を扱いやすい

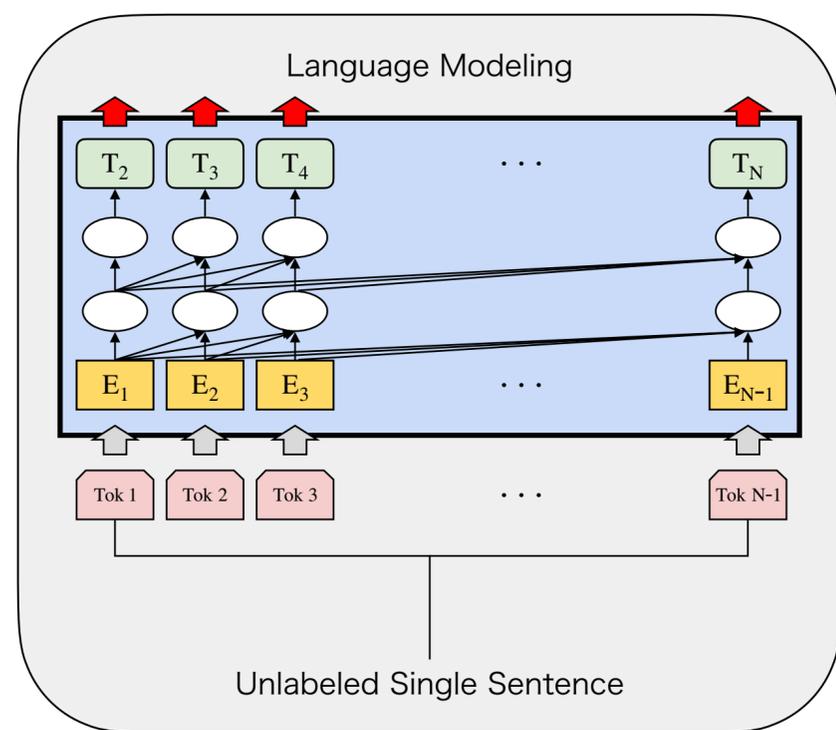
- 現在の大規模言語モデルを担うネットワーク構造



全単語をEncoderに同時に入力

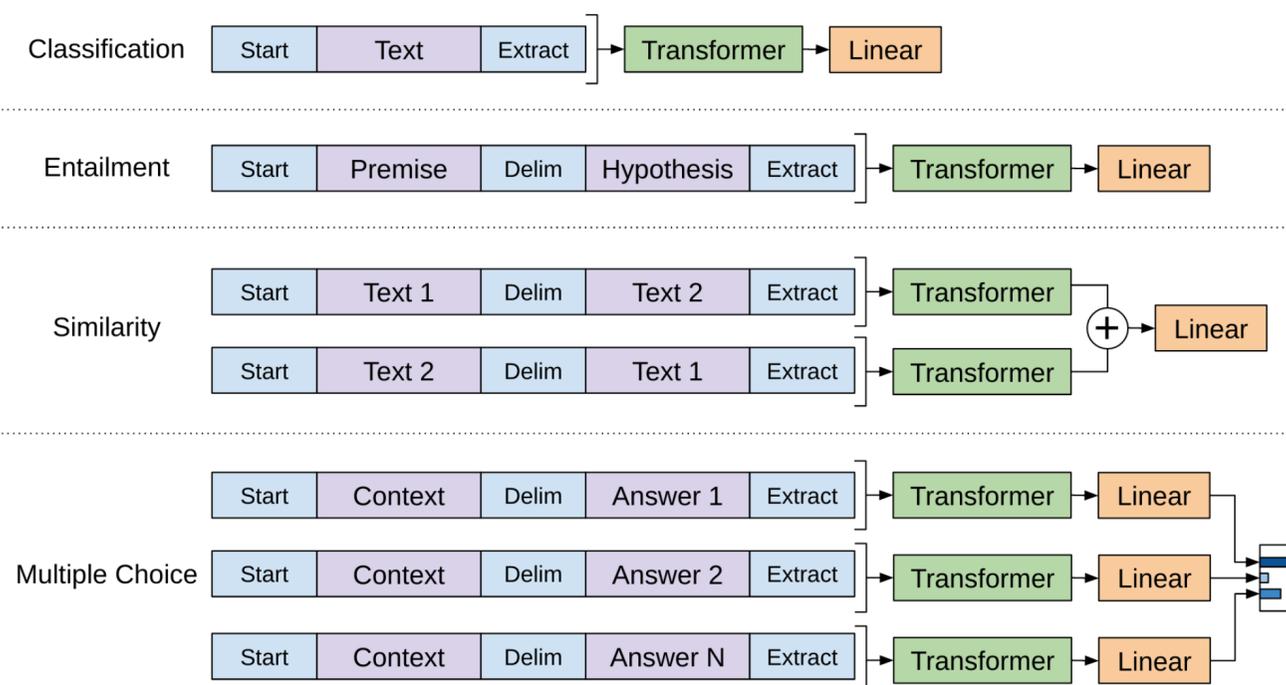
- Transformerを大量の言語パターンについて教師なしで事前学習
 - Language Modeling : 次の単語を予測
 - 自然な文脈 (単語の繋がり) について学習
- GPT-1, GPT-2, GPT-3などの複数のバージョンが存在

モデルとしてTransformerのデコーダを使用



事前学習

fine-tuning



一番最後の単語に対応する出力を入力として
タスク固有の出力をするLinear層を接続して学習

- 人間からのフィードバックに基づく強化学習
 - 有害なテキストを生成しないようにモデルを微調整
 - 会話データを用いたモデルの微調整により、会話ならではの砕けた表現を正しく理解し、適切な回答の生成が可能

Step 1: 教師あり学習

プロンプトと回答のペアによるデータセットを用いてGPT-3.5モデルをファインチューニング

Step 2: 報酬モデルの学習

プロンプトに対するモデルの複数の出力をパターンに対してアナレーターが順位付け
順位付けデータセットを用いて報酬モデルを学習

Step 3: 強化学習

Step1でファインチューニングしたGPT-3.5モデルとStep2の報酬モデルを用いて強化学習

報酬が最大となる方策を探索して最適な回答の生成

Step 1

Collect demonstration data and train a supervised policy.

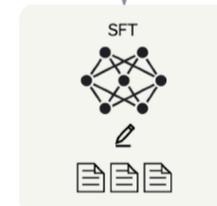
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



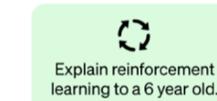
This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

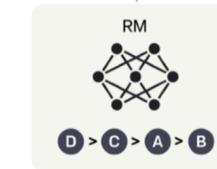
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.



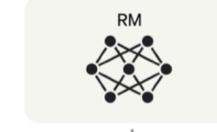
The PPO model is initialized from the supervised policy.



The policy generates an output.



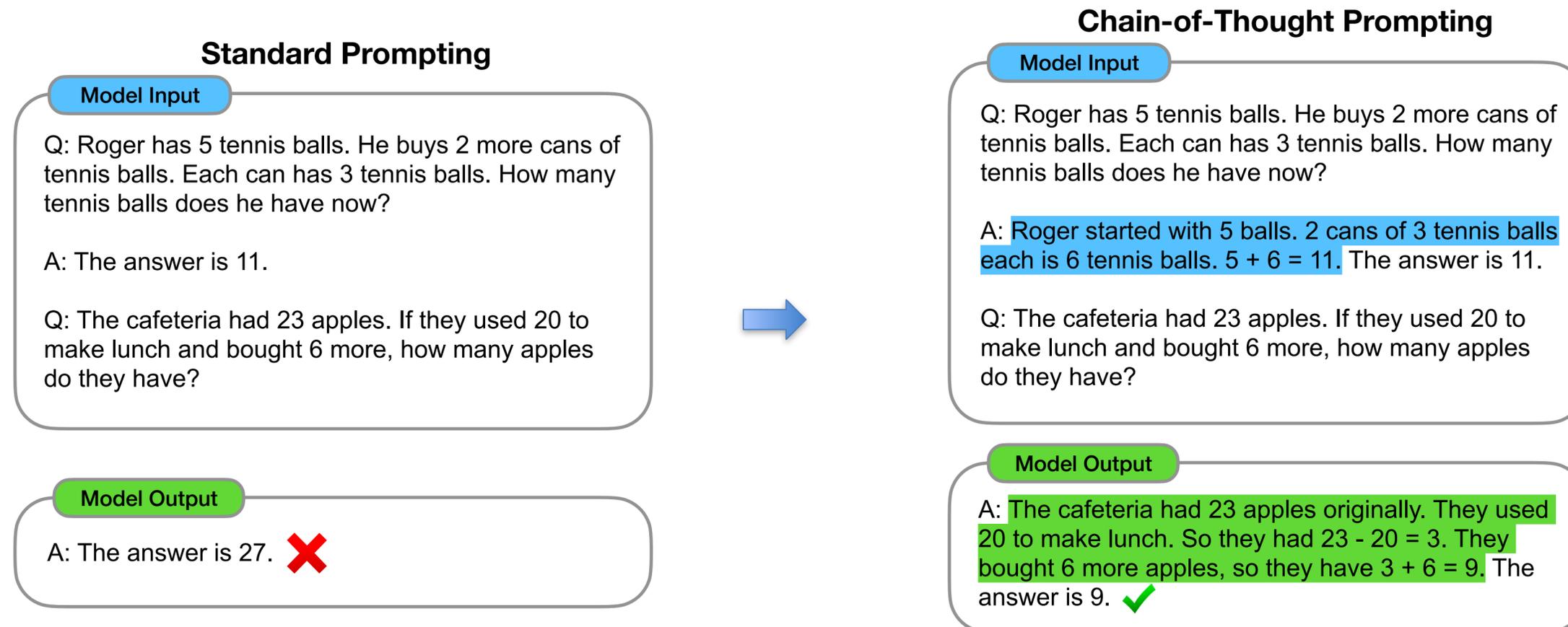
The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



- プロンプトの解答例に**考え方**も含めることで数学問題や常識推論などの性能を改善
 - Wei+ (2022) は考え方のテキストを手作業で作成
 - Kojima+ (2022) は単純に"Let's think step by step"の一文を追加
- 言語モデルが解答だけでなく**思考過程も出力**するように変化



- GPTのスコア, GPTと同スコアの受験者のパーセンタイルを報告 [OpenAI, 2022]

Exam	GPT-4	GPT-4 (no vision)	GPT-3.5
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)	298 / 400 (~90th)	213 / 400 (~10th)
LSAT	163 (~88th)	161 (~83rd)	149 (~40th)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)	710 / 800 (~93rd)	670 / 800 (~87th)
SAT Math	700 / 800 (~89th)	690 / 800 (~89th)	590 / 800 (~70th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)	157 / 170 (~62nd)	147 / 170 (~25th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)	165 / 170 (~96th)	154 / 170 (~63rd)
Graduate Record Examination (GRE) Writing	4 / 6 (~54th)	4 / 6 (~54th)	4 / 6 (~54th)
USABO Semifinal Exam 2020	87 / 150 (99th - 100th)	87 / 150 (99th - 100th)	43 / 150 (31st - 33rd)
USNCO Local Section Exam 2022	36 / 60	38 / 60	24 / 60
Medical Knowledge Self-Assessment Program	75 %	75 %	53 %
Codeforces Rating	392 (below 5th)	392 (below 5th)	260 (below 5th)
AP Art History	5 (86th - 100th)	5 (86th - 100th)	5 (86th - 100th)
AP Biology	5 (85th - 100th)	5 (85th - 100th)	4 (62nd - 85th)
AP Calculus BC	4 (43rd - 59th)	4 (43rd - 59th)	1 (0th - 7th)
AP Chemistry	4 (71st - 88th)	4 (71st - 88th)	2 (22nd - 46th)
AP English Language and Composition	2 (14th - 44th)	2 (14th - 44th)	2 (14th - 44th)
AP English Literature and Composition	2 (8th - 22nd)	2 (8th - 22nd)	2 (8th - 22nd)
AP Environmental Science	5 (91st - 100th)	5 (91st - 100th)	5 (91st - 100th)
AP Macroeconomics	5 (84th - 100th)	5 (84th - 100th)	2 (33rd - 48th)
AP Microeconomics	5 (82nd - 100th)	4 (60th - 82nd)	4 (60th - 82nd)
AP Physics 2	4 (66th - 84th)	4 (66th - 84th)	3 (30th - 66th)
AP Psychology	5 (83rd - 100th)	5 (83rd - 100th)	5 (83rd - 100th)
AP Statistics	5 (85th - 100th)	5 (85th - 100th)	3 (40th - 63rd)
AP US Government	5 (88th - 100th)	5 (88th - 100th)	4 (77th - 88th)
AP US History	5 (89th - 100th)	4 (74th - 89th)	4 (74th - 89th)
AP World History	4 (65th - 87th)	4 (65th - 87th)	4 (65th - 87th)
AMC 10 ³	30 / 150 (6th - 12th)	36 / 150 (10th - 19th)	36 / 150 (10th - 19th)
AMC 12 ³	60 / 150 (45th - 66th)	48 / 150 (19th - 40th)	30 / 150 (4th - 8th)
Introductory Sommelier (theory knowledge)	92 %	92 %	80 %
Certified Sommelier (theory knowledge)	86 %	86 %	58 %
Advanced Sommelier (theory knowledge)	77 %	77 %	46 %
Leetcode (easy)	31 / 41	31 / 41	12 / 41
Leetcode (medium)	21 / 80	21 / 80	8 / 80
Leetcode (hard)	3 / 45	3 / 45	0 / 45

Stanfordの大学院に入学できる

GRE (米国大学院の入学試験) や
 UBE (米国統一司法試験) などで高い成績を達成



Jim Fan
@DrJimFan

I don't give a damn about what is or isn't AGI. It doesn't matter.

Below is GPT-4's performance on many standardized exams: BAR, LSAT, GRE, AP, etc.

The truth is, GPT-4 can apply to Stanford as a student now. AI's reasoning ability is OFF THE CHARTS. Exponential growth is the scariest thing, isn't it!

<https://twitter.com/DrJimFan/status/1635694095460102145>

- GPT-4が苦手なタスク [Bubeck+, 2023]
 - 時事問題, 計算問題, 文字列

GPT-4 and ChatGPT: Symbolic manipulation

Answer the following questions.

Q: Who is the current president of the United States?

GPT-4:

Donald Trump

ChatGPT:

I'm sorry, but I am not able to browse the internet and my knowledge is limited to what was known up until 2021.

Q: What is $\sqrt{34324 * 2432}$?

GPT-4:

1287.6

ChatGPT:

The square root of $34324 * 2432$ is approximately 186.057539.

Q: What is the 13th letter of the word "supralapsarian"?

GPT-4:

The 13th letter of the word 'supralapsarian' is N.

ChatGPT:

The 13th letter of the word "supralapsarian" is "s".

ChatGPT(2025)

As of February 20, 2025, the current president of the United States is Donald J. Trump. He was inaugurated as the 47th president on January 20, 2025, marking his second, non-consecutive term in office.

The square root of $34324 * 2432$ is approximately 9136.52.

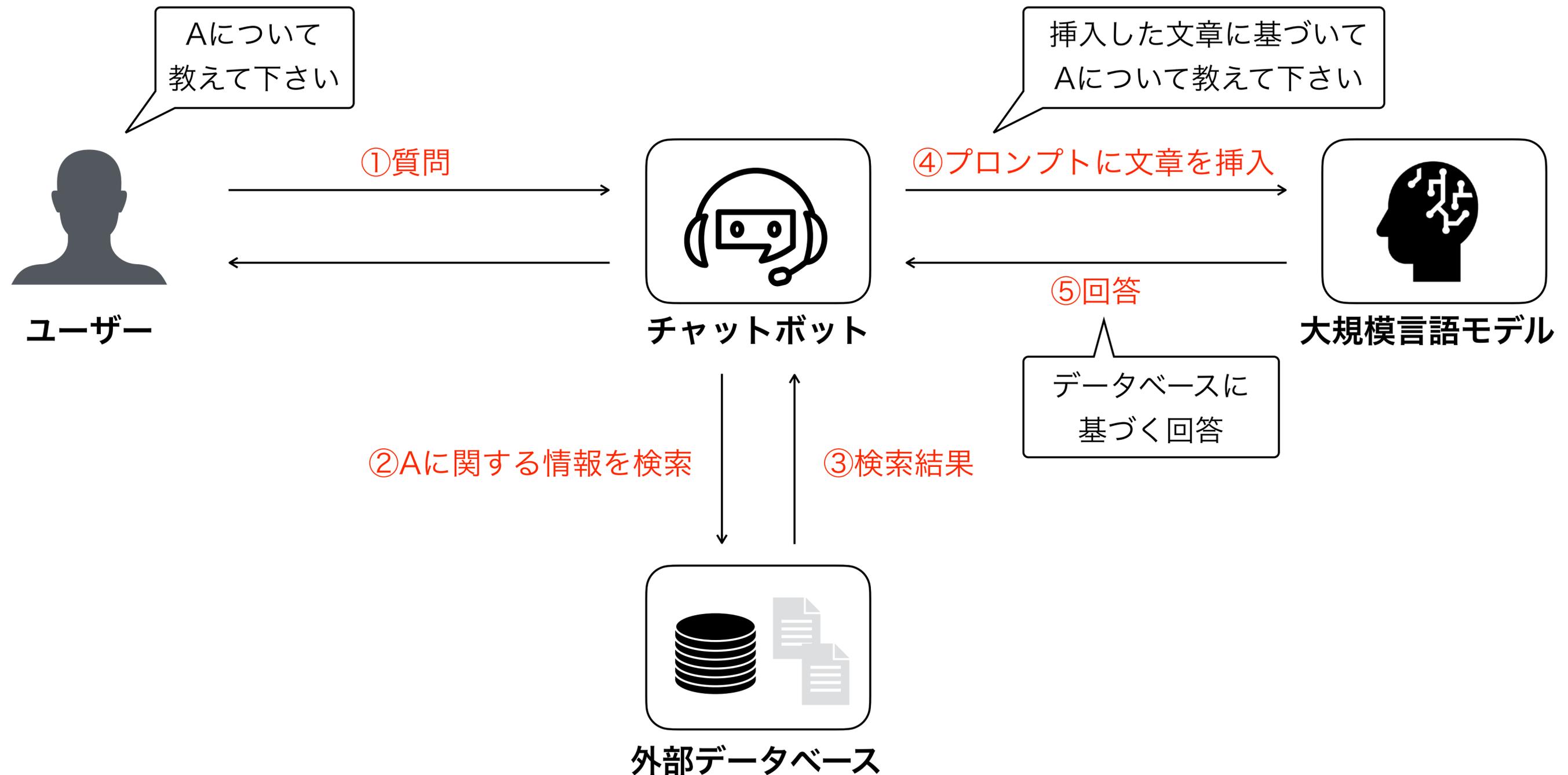
```
import math
# Calculate the square root of (34324 * 2432)
result = math.sqrt(34324 * 2432)
result
```

結果
9136.518374085394

The word "supralapsarian" has 14 letters. The 13th letter of the word is "a".
So, the 13th letter is "a".

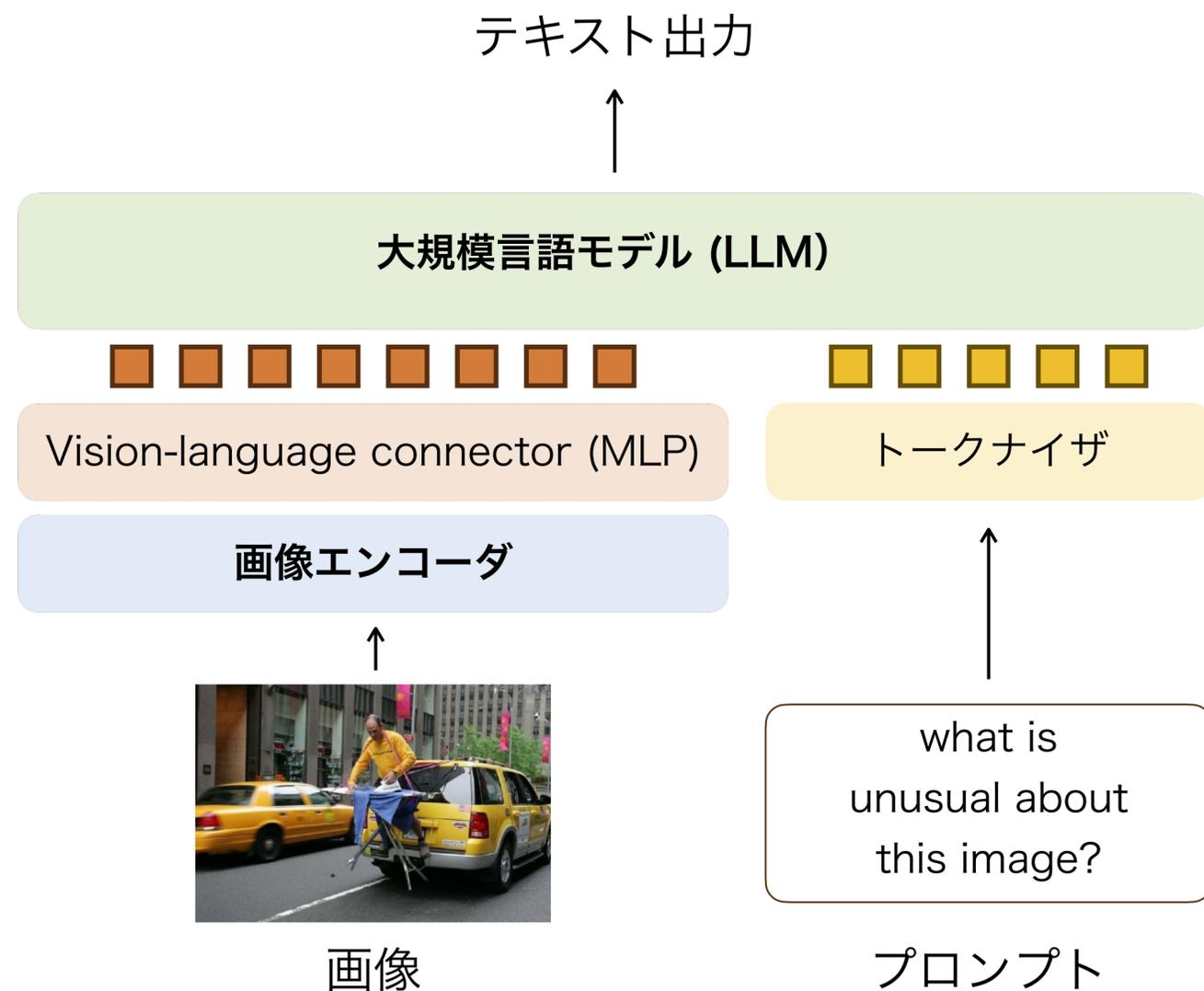
検索拡張生成 (RAG; Retrieval-Augmented Generation)

- 外部情報の検索を組み合わせることで、回答精度を向上させる技術



VLM (Vision-Language Model)

- ・ 画像エンコーダで得た視覚特徴をLLMに条件付けして画像内容の理解・推論・対話を自然言語で行えるようにしたマルチモーダルモデル
 - 画像特徴と言語特徴を橋渡しするvision-language connectorを導入



入力:



If there are factual errors in the questions, point it out; if not, proceed to answering the question.
What's happening in the desert?

質問に事実誤認がある場合は指摘してください。そうでない場合は、質問への回答に進んでください。
砂漠では何が起きているのでしょうか？

出力:

There are no deserts in the image. The image features a beach with palm trees, a city skyline, and a large body of water.

画像には砂漠は写っていません。
ヤシの木が生い茂るビーチ、街のスカイライン、そして広大な水域が写っています。

- 画像に対する質問応答



入力画像

User: What is unusual about this image?

この画像の何が異常なのでしょう？

LLaVA: The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

この写真の特異な点は、男性がミニバンかバンの荷台で衣類にアイロンをかけている点です。通常は、自宅など、より固定された安全な場所で、普通のアイロン台を使ってアイロンをかけるため、このような場所でアイロンをかけるのは一般的ではありません。写真に写っている光景は、車内で間に合わせのアイロン台を設置しているという点で特異であり、これは安全性と非常識性の両方を兼ね備えています。さらに、男性がこのような不安定な環境で、どのようにしてバランスと安定性を保って衣類にアイロンをかけているのかは不明です。

画像について一般的ではない事象について詳しく記述が可能

生成AIの**研究**への活用：藤吉AI先生（研究版）

- ・ 日々多くの論文が投稿されることから、短い時間で論文の概要を把握したい
 - アブストラクト：技術的な用語や論文独自の用語が多く、初見では内容が難しいことも

Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Albert Gu^{*1} and Tri Dao^{*2}

¹Machine Learning Department, Carnegie Mellon University

²Department of Computer Science, Princeton University
agu@cs.cmu.edu, tri@tridao.me

Abstract

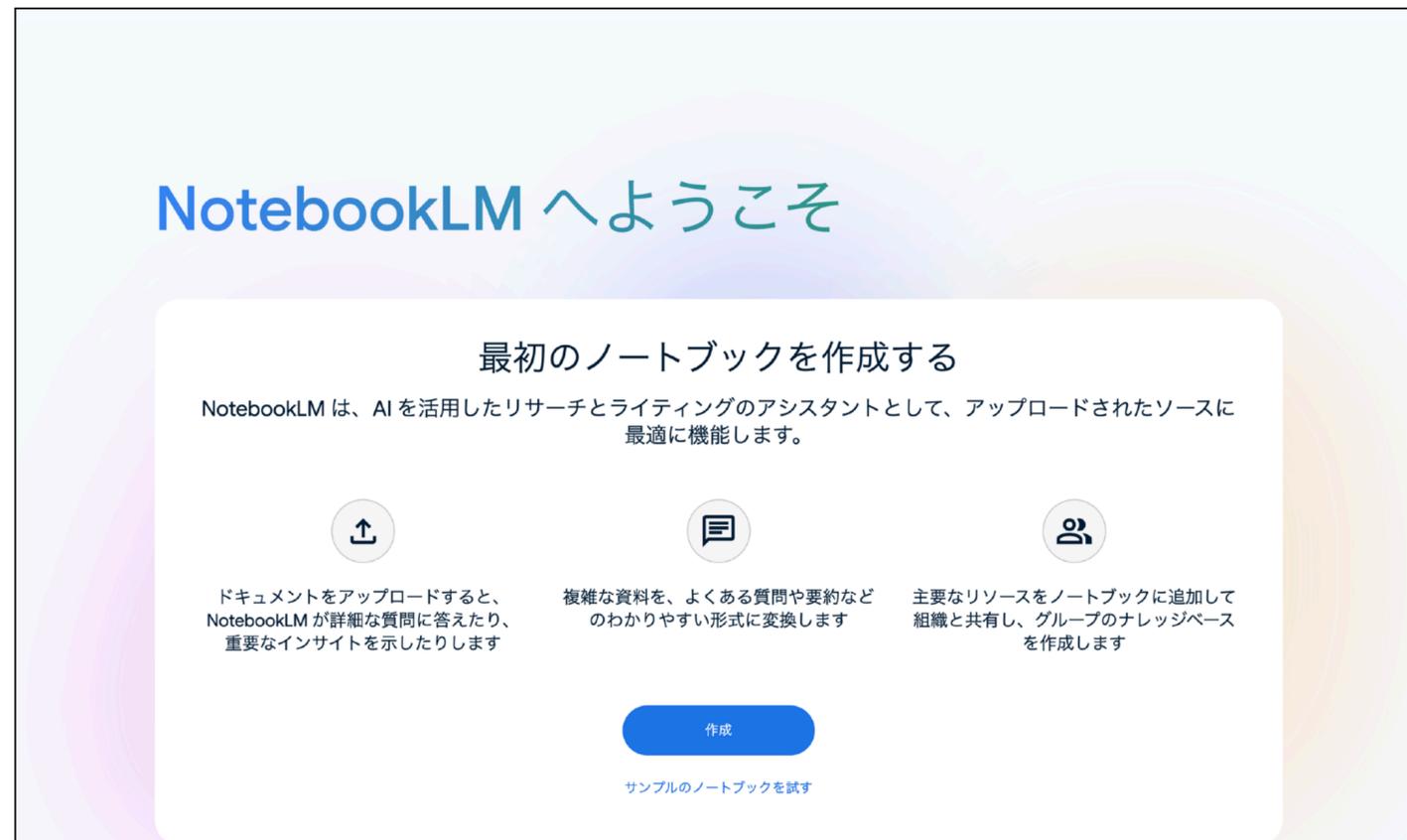
Foundation models, now powering most of the exciting applications in deep learning, are almost universally based on the Transformer architecture and its core attention module. Many subquadratic-time architectures such as linear attention, gated convolution and recurrent models, and structured state space models (SSMs) have been developed to address Transformers' computational inefficiency on long sequences, but they have not performed as well as attention on important modalities such as language. We identify that a key weakness of such models is their inability to perform content-based reasoning, and make several improvements. First, simply letting the SSM parameters be functions of the input addresses their weakness with discrete modalities, allowing the model to *selectively* propagate or forget information along the sequence length dimension depending on the current token. Second, even though this change prevents the use of efficient convolutions, we design a hardware-aware parallel algorithm in recurrent mode. We integrate these selective SSMs into a simplified end-to-end neural network architecture without attention or even MLP blocks (**Mamba**). Mamba enjoys fast inference (5× higher throughput than Transformers) and linear scaling in sequence length, and its performance improves on real data up to million-length sequences. As a general sequence model backbone, Mamba achieves state-of-the-art performance across several modalities such as language, audio, and genomics. On language modeling, our Mamba-3B model outperforms Transformers of the same size and matches Transformers twice its size, both in pretraining and downstream evaluation.

1 Introduction

Foundation models (FMs), or large models pretrained on massive data then adapted for downstream tasks, have emerged as an effective paradigm in modern machine learning. The backbone of these FMs are often *sequence models*, operating on arbitrary sequences of inputs from a wide variety of domains such as language, images, speech, audio, time series, and genomics (Brown et al. 2020; Dosovitskiy et al. 2020; Ismail Fawaz et al. 2019; Oord et al. 2016; Poli et al. 2023; Sutskever, Vinyals, and Quoc V Le 2014). While this concept is agnostic to a particular choice of model architecture, modern FMs are predominantly based on a single type of sequence model: the Transformer (Vaswani et al. 2017) and its core attention layer (Bahdanau, Cho, and Bengio 2015). The efficacy of self-attention is attributed to its ability to route information densely within a context window, allowing it to model complex data. However, this property brings fundamental drawbacks: an inability to model anything outside of a finite window, and quadratic scaling with respect to the window length. An enormous body of research has appeared on more efficient variants of attention to overcome these drawbacks (Tay, Dehghani, Bahri, et al. 2022), but often at the expense of the very properties that makes it effective. As of yet, none of these

現在、ディープラーニングにおける多くのエキサイティングなアプリケーションを支える基盤モデルは、ほぼ普遍的にTransformerアーキテクチャとその中核であるアテンションモジュールに基づいています。Transformerの長いシーケンスにおける計算効率の悪さに対処するために、線形アテンション、ゲート付き畳み込み、再帰モデル、構造化状態空間モデル (SSM) などの多くのサブ二乗時間アーキテクチャが開発されてきましたが、これらは言語などの重要なモダリティにおいてアテンションほどの性能を発揮していません。私たちは、これらのモデルの主要な弱点がコンテンツベースの推論を行う能力の欠如であることを特定し、いくつかの改良を加えました。第一に、SSMのパラメータを入力関数とするだけで、離散的なモダリティにおける弱点を補い、現在のトークンに応じてシーケンス長方向に情報を選択的に伝播または忘却することをモデルに可能にします。第二に、この変更により効率的な畳み込みの使用が妨げられるにもかかわらず、再帰モードでハードウェアに配慮した並列アルゴリズムを設計しました。これらの選択的SSMを、アテンションやMLPブロックさえも含まない簡素化されたエンドツーエンドのニューラルネットワークアーキテクチャ (Mamba) に統合しました。Mambaは高速な推論 (Transformerの5倍のスループット) とシーケンス長に対する線形スケーリングを享受し、実データでは100万長さのシーケンスまで性能が向上します。一般的なシーケンスモデルのバックボーンとして、Mambaは言語、音声、ゲノミクスなどの複数のモダリティにおいて最先端の性能を達成しています。言語モデルにおいて、私たちのMamba-3Bモデルは同サイズのTransformerを上回り、同等の事前学習および下流評価において、サイズが2倍のTransformerに匹敵します。

- Googleが開発したLLM+RAGによる研究・メモ作成用のチャットサービス
 - アップロードしたコンテンツ（PDF, .txt, Markdown, 音声）に基づいてLLM（Gemini）が回答
 - 作成したノートブックは簡単に他者へ共有が可能



- 過去に行った研究で引用数が多い論文や執筆記事等をアップロード

NotebookLM

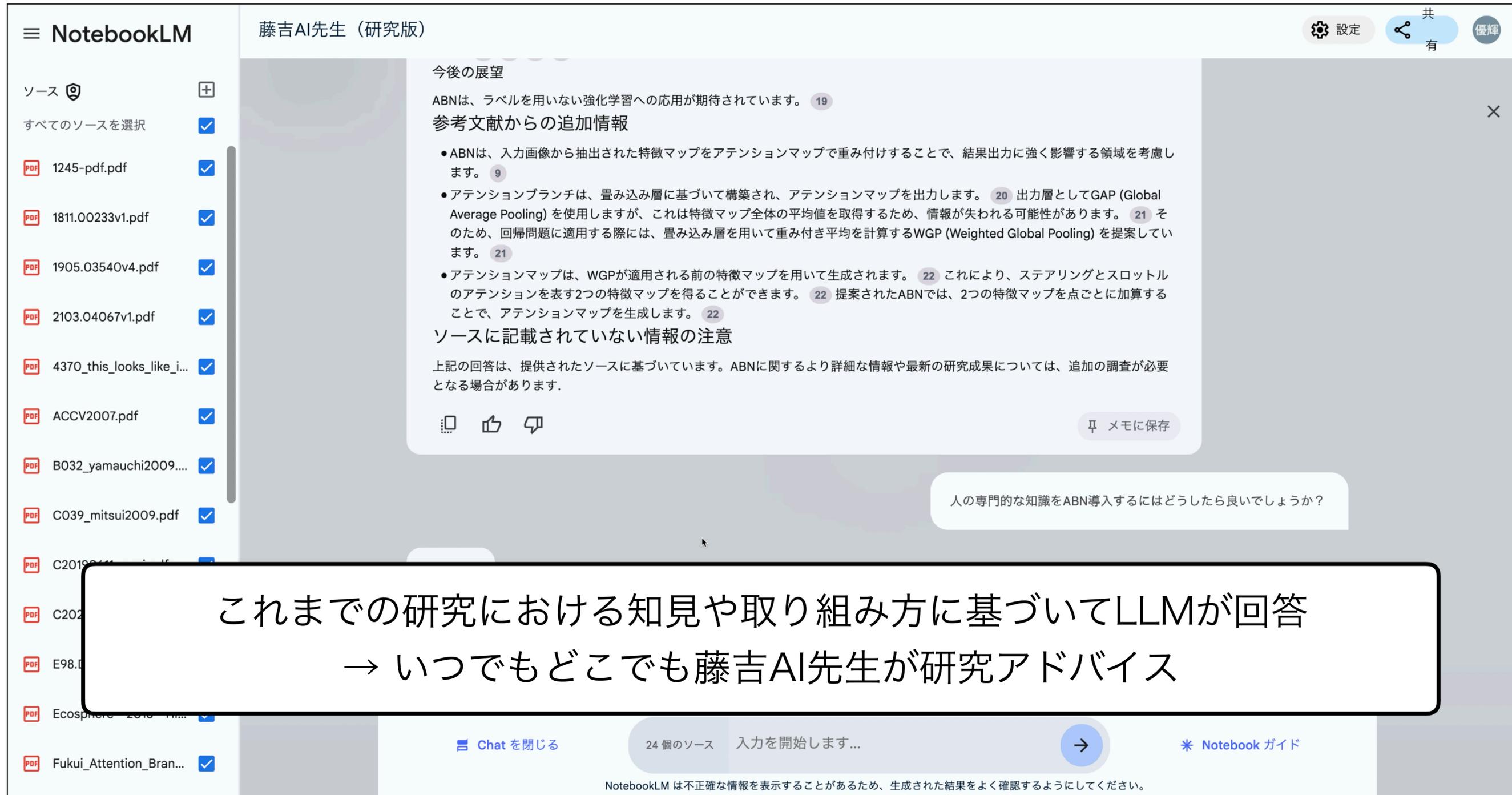
Hironobu Fujiyoshi
Professor, Chubu University
確認したメール アドレス: isc.chubu.ac.jp - ホームページ
Computer vision

タイトル	引用先	年
Moving target classification and tracking from real-time video AJ Lipton, H Fujiyoshi, RS Patil Proceedings fourth IEEE workshop on applications of computer vision, WACV'98 ...	2097	1998
A system for video surveillance and monitoring RT Collins, AJ Lipton, T Kanade, H Fujiyoshi, D Duggins, Y Tsin, ... VSAM final report 2000 (1-68), 1	1812	2000
Algorithms for cooperative multisensor surveillance RT Collins, AJ Lipton, H Fujiyoshi, T Kanade Proceedings of the IEEE 89 (10), 1456-1477	870	2001
Real-time human motion analysis by image skeletonization H Fujiyoshi, AJ Lipton, T Kanade IEICE TRANSACTIONS on Information and Systems 87 (1), 113-120	737	2004
Attention branch network: Learning of attention mechanism for visual explanation H Fukui, T Hirakawa, Y Yamashita, H Fujiyoshi Proceedings of the IEEE/CVF conference on computer vision and pattern ...	615	2019
A system for video surveillance and monitoring: VSAM final report R Collins, A Lipton, T Kanade, H Fujiyoshi, D Duggins, Y Tsin, D Tolliver, ... Technical report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University	517	2000
Deep learning-based image recognition for autonomous driving H Fujiyoshi, T Hirakawa, T Yamashita IATSS research 43 (4), 244-252	491	2019
Real-time human detection using relational depth similarity features S Ikemura, H Fujiyoshi Computer Vision-ACCV 2010: 10th Asian Conference on Computer Vision ...	167	2011
Boosted random forest Y Mishina, R Murata, Y Yamauchi, T Yamashita, H Fujiyoshi IEICE TRANSACTIONS on Information and Systems 98 (9), 1630-1636	128	2015
Embedding human knowledge into deep neural network via attention map M Mitsuhara, H Fukui, Y Sakashita, T Ogata, T Hirakawa, T Yamashita, ... arXiv preprint arXiv:1905.03540	91	2019
Pedestrian detection based on deep convolutional neural network with ensemble inference network H Fukui, T Yamashita, Y Yamauchi, H Fujiyoshi, H Murase 2015 IEEE Intelligent Vehicles Symposium (IV), 223-228	71	2015



<https://scholar.google.com/citations?user=CIHKZpEAAAAJ>

https://notebooklm.google.com/notebook/5cb9a600-5acf-493d-a13d-cae157dd6434



藤吉AI先生 (研究版)

設定 共有 優輝

今後の展望

ABNは、ラベルを用いない強化学習への応用が期待されています。 19

参考文献からの追加情報

- ABNは、入力画像から抽出された特徴マップをアテンションマップで重み付けすることで、結果出力に強く影響する領域を考慮します。 9
- アテンションブランチは、畳み込み層に基づいて構築され、アテンションマップを出力します。 20 出力層としてGAP (Global Average Pooling) を使用しますが、これは特徴マップ全体の平均値を取得するため、情報が失われる可能性があります。 21 そのため、回帰問題に適用する際には、畳み込み層を用いて重み付き平均を計算するWGP (Weighted Global Pooling) を提案しています。 21
- アテンションマップは、WGPが適用される前の特徴マップを用いて生成されます。 22 これにより、ステアリングとスロットルのアテンションを表す2つの特徴マップを得ることができます。 22 提案されたABNでは、2つの特徴マップを点ごとに加算することで、アテンションマップを生成します。 22

ソースに記載されていない情報の注意

上記の回答は、提供されたソースに基づいています。ABNに関するより詳細な情報や最新の研究成果については、追加の調査が必要となる場合があります。

メモに保存

人の専門的な知識をABN導入するにはどうしたら良いでしょうか？

Chat を閉じる 24 個のソース 入力を開始します... Notebook ガイド

NotebookLM は不正確な情報を表示することがあるため、生成された結果をよく確認するようにしてください。

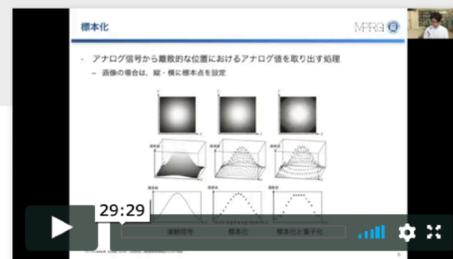
これまでの研究における知見や取り組み方に基づいてLLMが回答
→ いつでもどこでも藤吉AI先生が研究アドバイス

生成AIの**教育**への活用：藤吉AI先生（教育版）

ロボットビジョン

01：画像の標本化・量子化、濃淡画像の処理 02：空間フィルタリング 03：画像の幾何変換、イメージモザイク 04：ハフ変換、フーリエ変換 05：二値化、二値画像処理 06：テンプレートマッチング 07：特徴点検出・記述（SIFT）
08：コンピュータビジョン 09：アフィン不変なキーポイントマッチング 資料：http://mprg.cs.chubu.ac.jp/~hf/Tutorial_IP.pdf

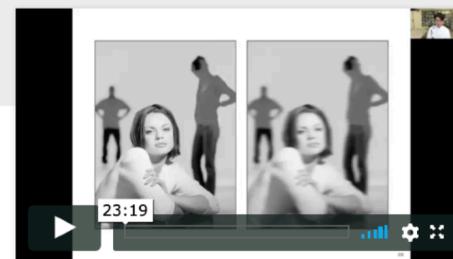
1 - 12 of 16 共有



01-01：画像の標本化・量子化

Hironobu Fujiyoshi

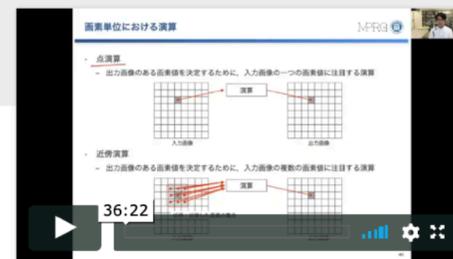
画像の標本化・量子化 デジタル画像の表現 資料：
http://mprg.cs.chubu.ac.jp/~hf/Tutorial_IP.pdf



01-02：濃淡画像の処理

Hironobu Fujiyoshi

濃淡画像処理 画質とは 濃淡ヒストグラム コントラスト 線形変換、ガンマ補正 資料：...



02-01：空間フィルタリング...

Hironobu Fujiyoshi

画像処理の基本モデル 近傍演算 平滑化 [1] 移動平均フィルタ [2] 加重平均フィルタ（ガウシアンフ...



02-02：空間フィルタリング...

Hironobu Fujiyoshi

画像の微分とは エッジ抽出 [1] 微分フィルタ [2] プリューウィットフィルタ [3] ソーベルフィルタ [4] ...



03-01：画像の幾何変換

Hironobu Fujiyoshi

画像の幾何変換 [1] 平行移動 [2] 拡大縮小 [3] 回転 [4] スキュー 変換の組み合わせ アフィン変換 資料：...



03-02：画像の幾何変換（イ...

Hironobu Fujiyoshi

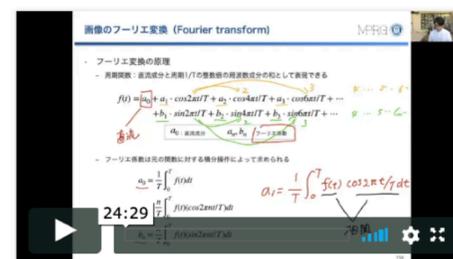
画像の再配列 画像データの内挿 イメージモザイク 資料：...



04-01：ハフ変換

Hironobu Fujiyoshi

ハフ変換による直線の検出 円の検出 資料：
http://mprg.cs.chubu.ac.jp/~hf/Tutorial_IP.pdf



04-02：フーリエ変換

Hironobu Fujiyoshi

フーリエ変換の原理 逆フーリエ変換 資料：
http://mprg.cs.chubu.ac.jp/~hf/Tutorial_IP.pdf



05-01：二値化

Hironobu Fujiyoshi

グレースケール変換 二値化 判別分析法 資料：
http://mprg.cs.chubu.ac.jp/~hf/Tutorial_IP.pdf



05-02：二値画像処理

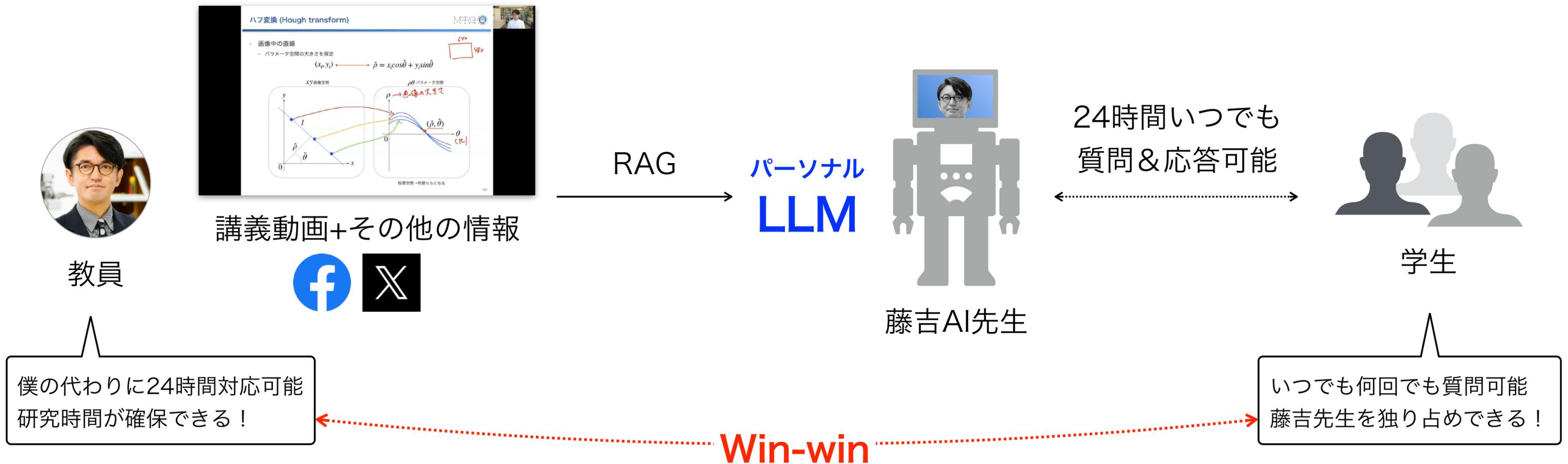
Hironobu Fujiyoshi

連結性 輪郭線追跡 収縮・膨張 ラベリング 形状特徴パラメータ 資料：...

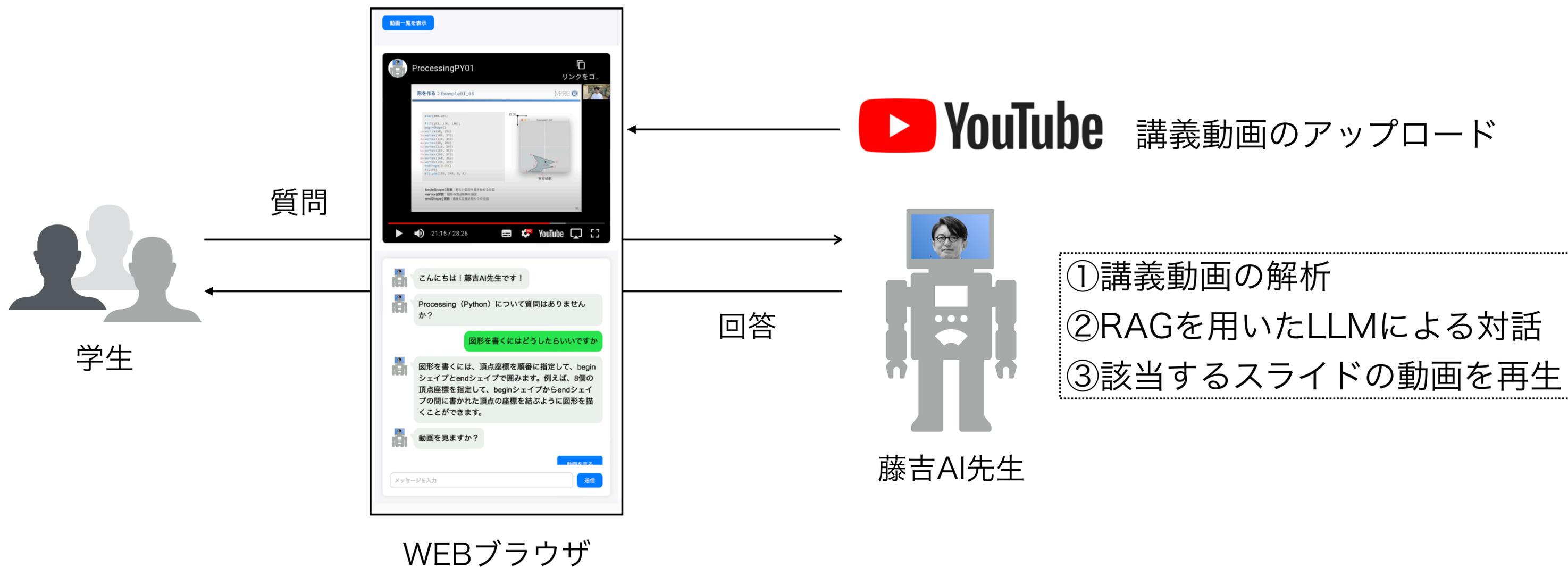
講義「ロボットビジョン」の動画（パスワード：MPRG）

<https://vimeo.com/showcase/7170791>

- ・ 教員が作成した講義動画コンテンツを用いてLLMを学習（復習）に活用
→ 講義の内容に沿った対話が可能

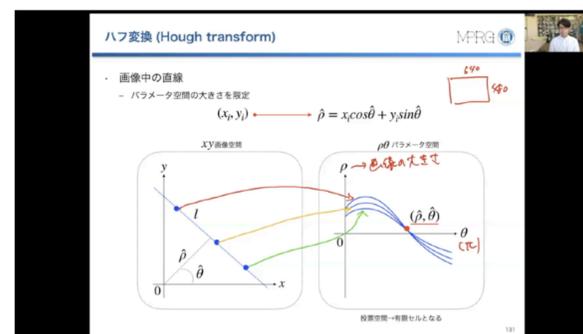


- ・ 教員が作成したコンテンツを用いてパーソナライズしたLLMを教育に活用
→ 講義の内容に沿った対話が可能



1. 講義動画からスライドの切り替えタイミングを自動検出
2. スライド毎の発話を文章化

講義動画



スライド切り替えタイミング：

↓ 0分0秒

↓ 2分30秒

↓ 5分21秒

↓ 8分47秒

↓ 11分32秒

↓ 26分13秒



スライド1の
発話文章



スライド2の
発話文章



スライド3の
発話文章



スライド4の
発話文章

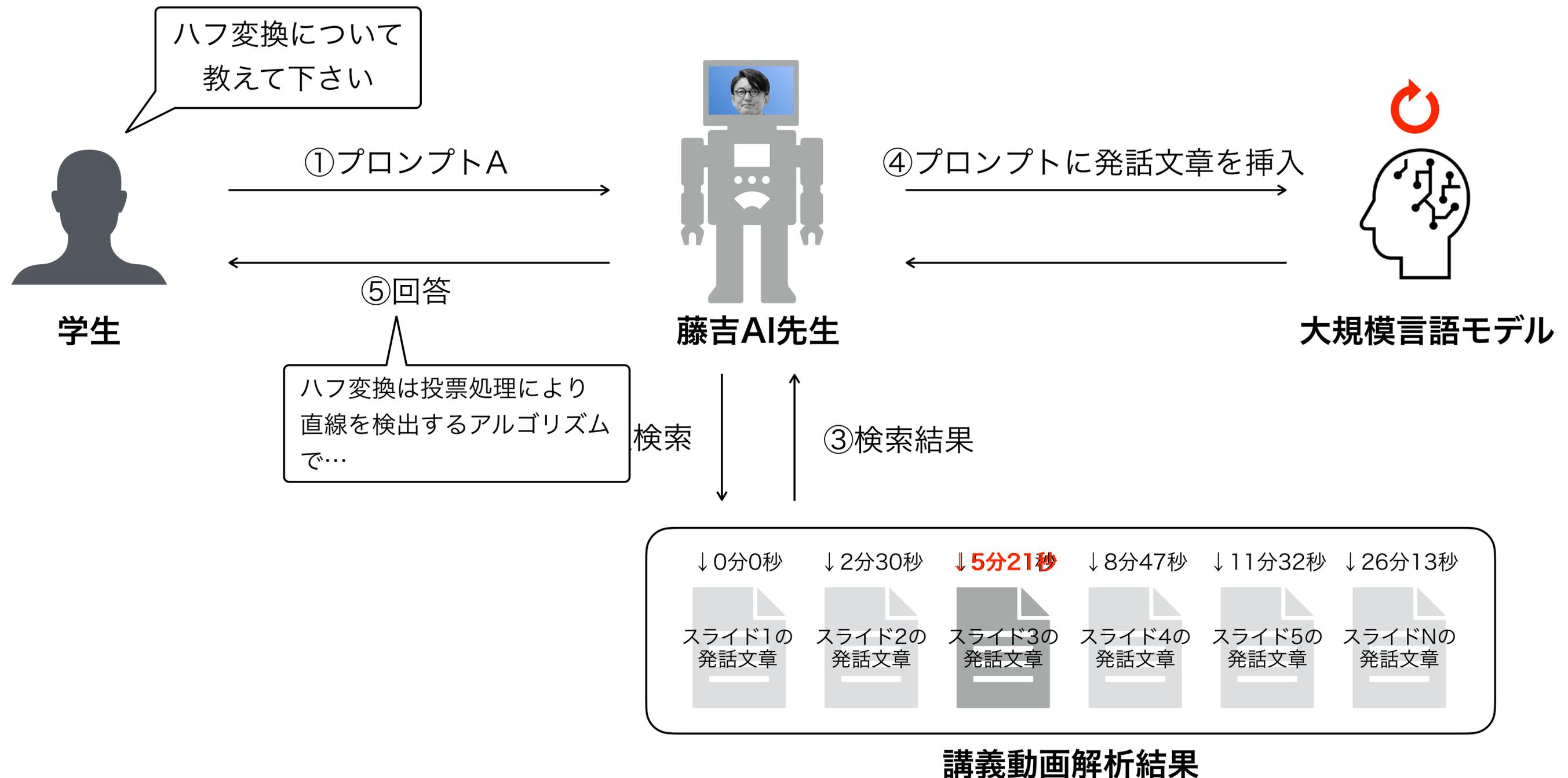


スライド5の
発話文章



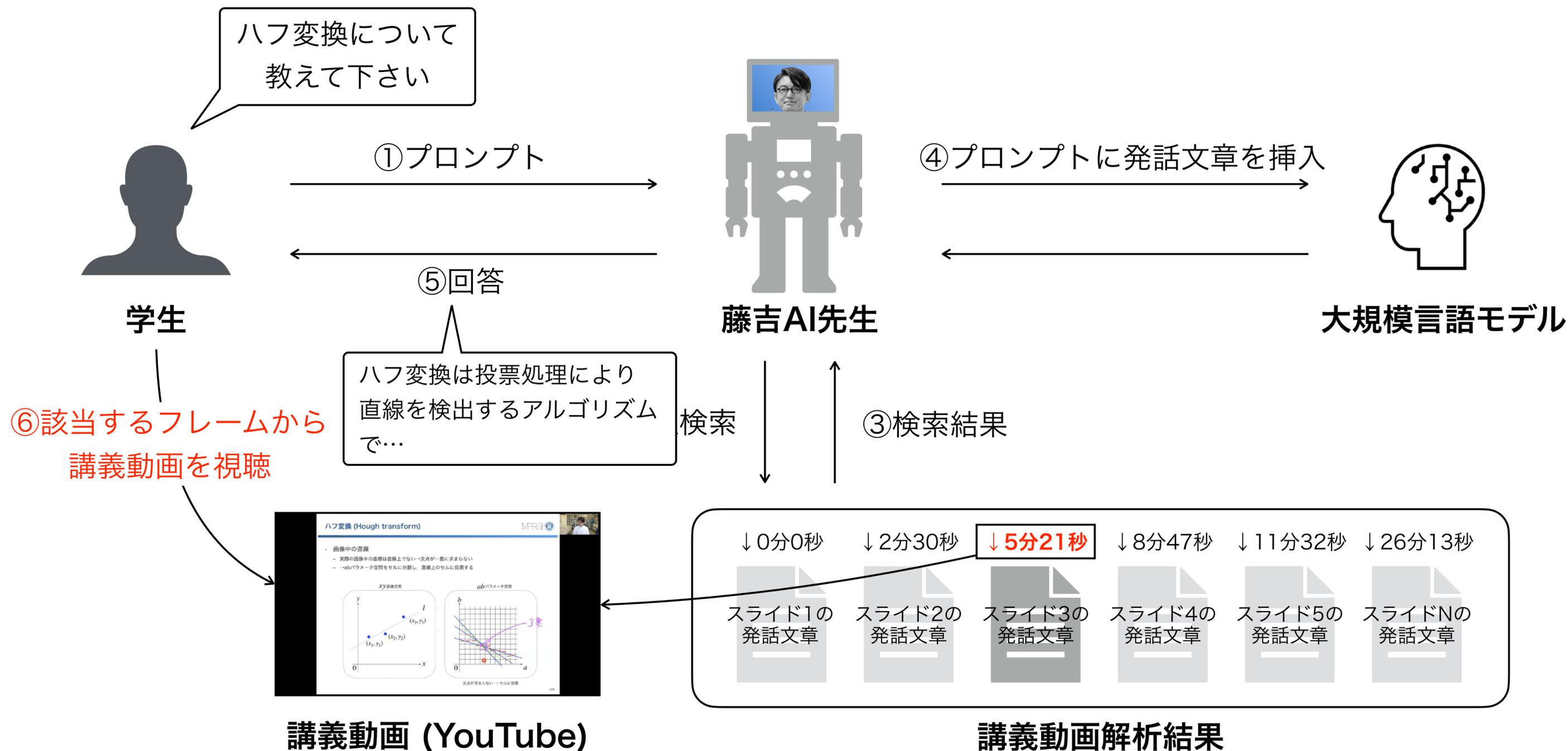
スライドNの
発話文章

- 質問に対する講義動画の該当する発話文章を検索してLLMのプロンプトに追加



藤吉AI先生：③該当するスライドの動画を再生

- ・ 質問と回答に該当するスライドの動画フレームから再生



質問	Chat-GPT	藤吉AI先生 (プログラミングの授業版)
<p>図形を書くには どうしたらいいですか？</p>	<p> 図形を描く方法はいくつかありますが、目的や使用するツールによって異なります。以下に、一般的な方法をいくつか紹介します。</p> <p>1. 手描き 紙とペンを使って手描きで図形を描く方法です。これが最もシンプルで直感的な方法です。</p> <p>2. コンピューターソフトウェアを使用</p> <p>a. グラフィックデザインソフトウェア</p> <ul style="list-style-type: none">• Adobe Illustrator: プロフェッショナルなベクターグラフィックデザインソフト。複雑な図形やデザインを作成するのに適しています。• CorelDRAW: Adobe Illustratorに似たソフトウェアで、ベクターグラフィックの作成に適しています。 <p>↓</p> <p>h 図形描画専用ソフト</p> <p> ChatGPT にメッセージを送信する </p>	<p> 図形を書くには、頂点座標を順番に指定して、beginシェイプとendシェイプで囲みます。例えば、8個の頂点座標を指定して、beginシェイプからendシェイプの間に書かれた頂点の座標を結ぶように図形を描くことができます。</p> <p> 動画を見ますか？</p> <p>動画を見る</p>

↓

 一般的な回答

↓

 講義内容に沿った回答

動画一覧を表示



<http://ikz1-123-68031.vs.sakura.ne.jp/ja/share/s871RTFeFoXTL19IZK0Q2BX2sso3iGApH5AbmP7CIOM>



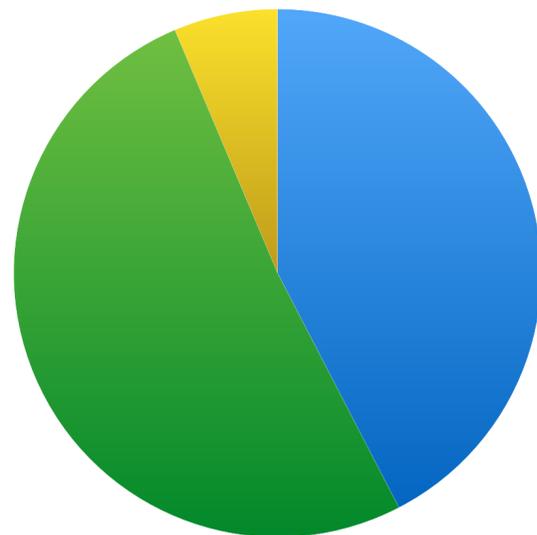
こんにちは！藤吉AI先生です！



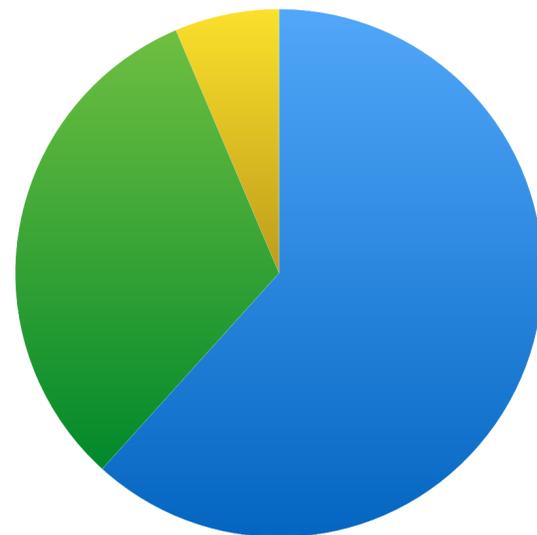
ロボットビジョン（画像処理）について質問はありませんか？

- ・ 講義受講者にアンケートを実施
 - 学部生：43人 「ロボット工学入門 (1年生)」 「ロボットビジョン (3年生)」
 - 社会人：4人 「CU Synergy Program AI講座」

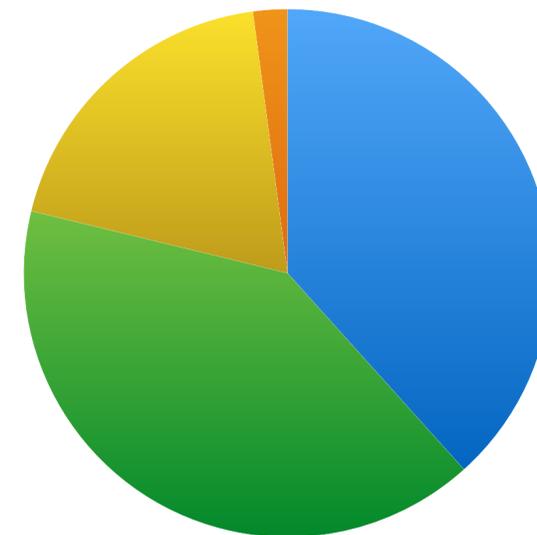
● 強くそう思う
● そう思う
● 普通
● 思わない



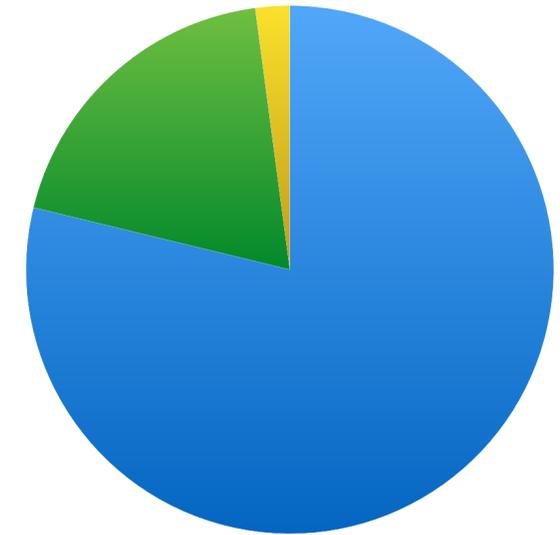
復習に役に立ったか？



動画への接続は適切だったか？



回答は適切だったか？



他の講義でも実現して欲しい？

動画への接続性は高く，講義の復習に役に立つことが分かった

- ・ 生成AIの利用

- 定型業務の効率化：e.g., 議事録の要約
- クリエイティブな提案の補助：e.g., アイディアの壁打ち

- ・ 生成AIの課題

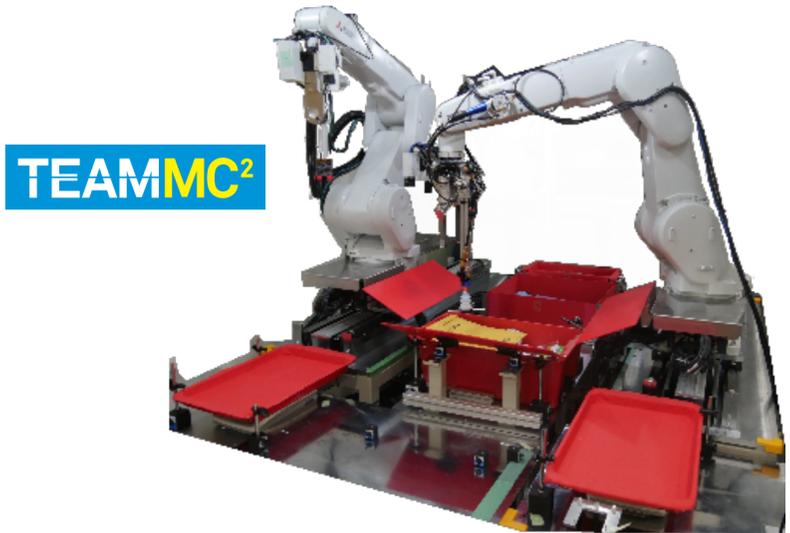
- モデルに依存する出力精度
- ハルシネーション（AIがもっともらしい嘘をつく）のリスク
- 敵対的プロンプトへの対策が不十分
- 学習データの扱い（著作権）

→ **AIの仕組みを知り、その生成物の内容を盲信せず、必ず根拠や裏付けを自ら確認する力**

フィジカルAI

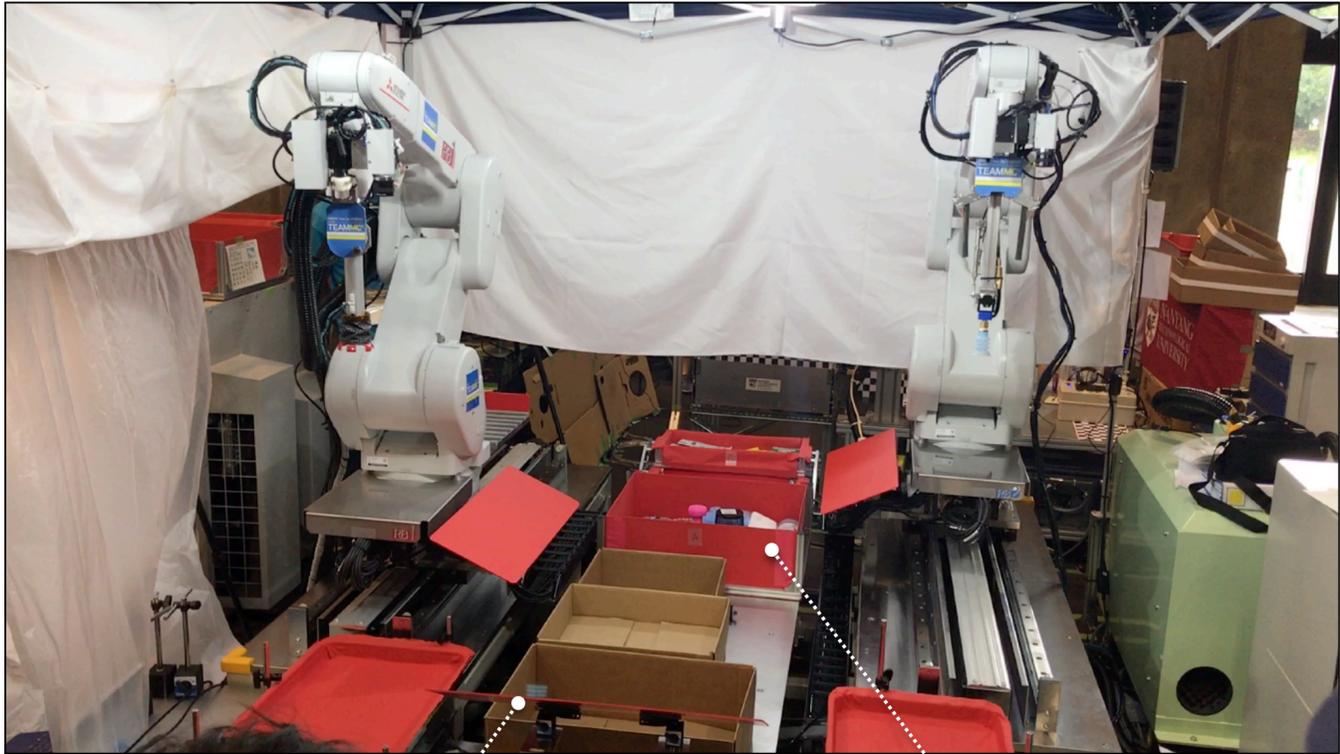
Amazon Robotics Challenge

- Amazonが2015～2017年に開催した国際ロボティクス競技
 - 「倉庫での商品ピッキング・仕分けを自律ロボットでどこまで実現できるか」を競う大会



Stow Task: **3rd place**

Amazon Robotics Challenge Pick Task



Box(S, M, L)

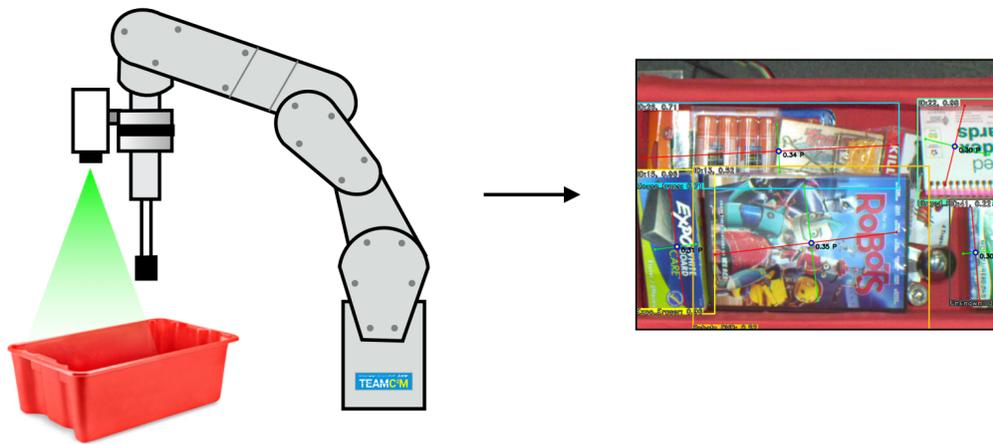
Shelf

Order list

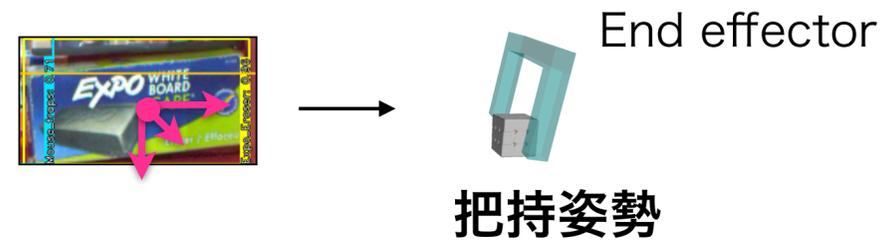
ピック対象	プレース
	→ S box
item 19	
	→ L box
item 30	
	→ M box
item 11	
	⋮

- モジュール構成：深層学習は「①物体認識」のみに利用

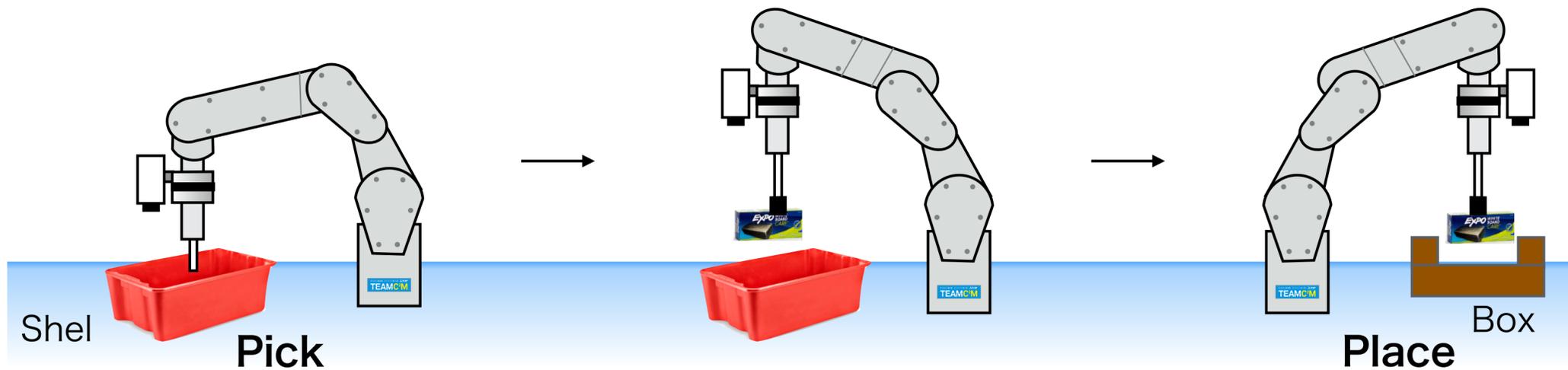
① 物体認識 (Object recognition)



② 把持位置検出 (Grasp detection)



③ 経路計画 (Motion planning)



④ 制御 (Robot control)



・従来のロボット制御

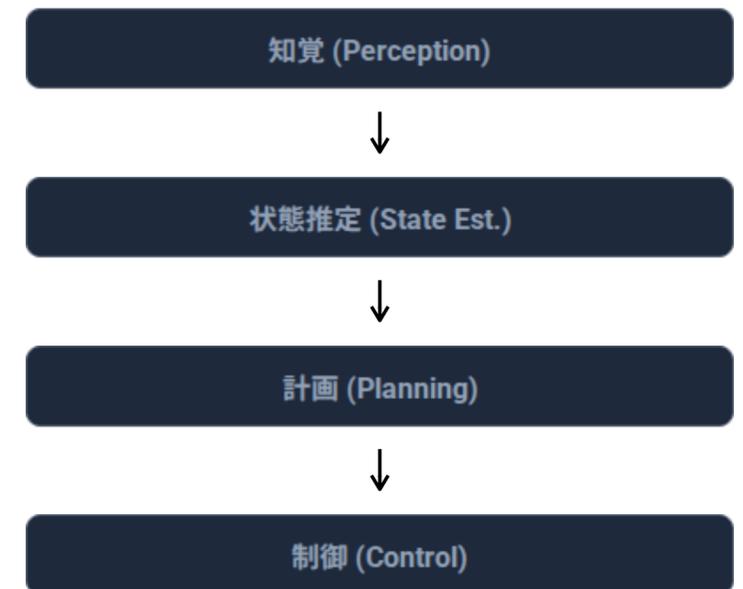
- 分業制 (Modular) : 各工程が独立しており、前段のエラーが後段に波及 (伝言ゲーム)
- 閉じた世界 : 定義された物体・環境以外には対応できない
- ルールベース : 「AならばB」という明示的なプログラムが必要

・VLA (Vision-Language-Action) モデル

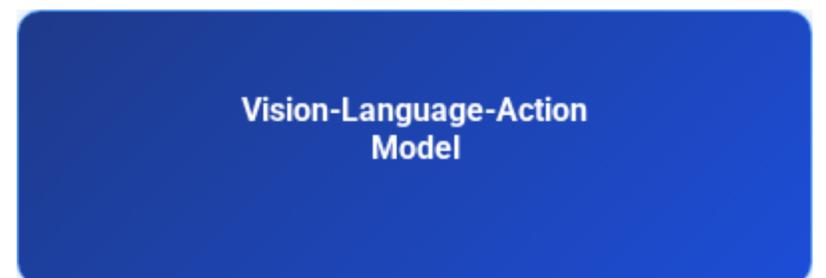
- 一気通貫 (**End-to-End**) : 画像と指示を入れれば、直接「動作」を出力
- **開かれた世界** : 未知の物体も「意味」で理解して対応可能
- **データ駆動** : ルールを書くのではなく、大量のデータから振る舞いを学ぶ

→ フィジカルAI : 視覚 + 言語 + 行動 (VLA) で実世界タスクを遂行

従来の制御スタック



VLAの制御



- 学習データ

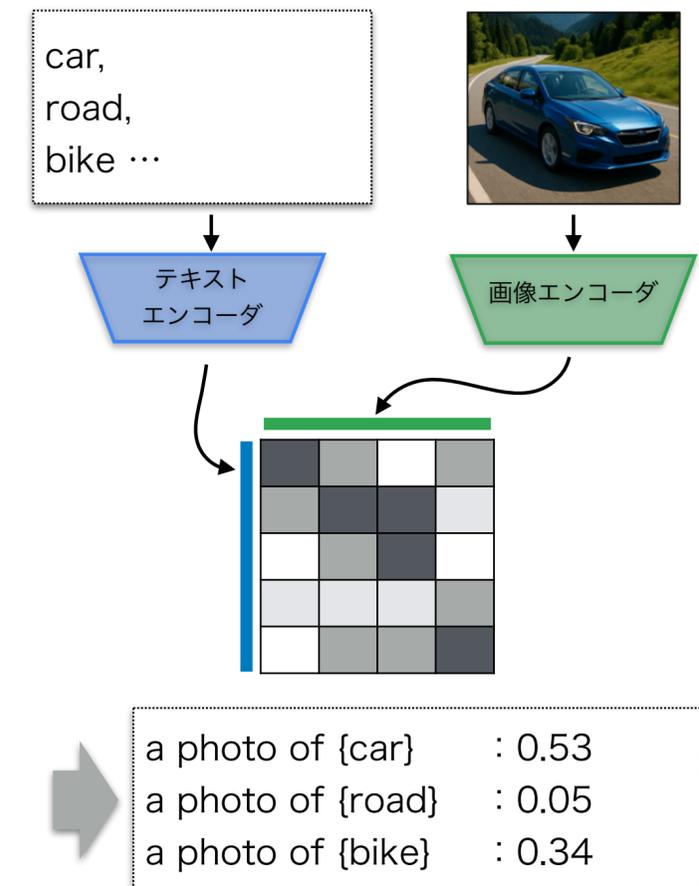
- インターネット上の数十億規模の「画像とテキストのペア」

- 学習プロセス

- 画像と単語（言語）とのアライメント学習
- 共通の埋め込み空間を獲得：画像の特徴と言語の特徴を同じ特徴にマッピング
- ex. 「carの画像」の特徴ベクトルと「car」という単語の特徴ベクトルが空間上の同じ場所（近い距離）に配置

→ 画像を見ただけで、言葉（知識）へのアクセスが可能

画像とテキストのペアで学習



- ・ 大規模言語モデル (LLM) の利用
 - WEB上の膨大なテキストから得た「世界知識」を内包している
 - 画像エンコーダとLLMを結びつけ、高度なVQA推論を実現

・ Semantic Understanding (意味的理解)

- 属性知識 :

水のボトル、空で潰れたコーラ、空いていないコーラ

- 関係性理解 :

水のボトル → 「キャップがしっかり閉まっており、中身も入っているため、まだ飲むことができます。」

空で潰れたコーラ → 「缶が大きく潰れており、飲み口も開いています。明らかに飲み終わった後のゴミです。」

空いていないコーラ → 「プルタブが開いておらず、「未開封」の状態です。」

→ 画像を見ただけで、言葉 (知識) へのアクセスが可能

VLM : LLMの活用

Q. Please describe what you see in the image.



画像エンコーダ

大規模現言語モデル (LLM)

A. Water Bottle: Closed / Full
Standing Coke: Unopened / Full
Lying Coke: Opened / Empty / Crushed

- VLA (**V**ision-**L**anguage-**A**ction) の定義
 - VLMにロボット制御能力（アクションの生成）を統合したモデル

- 三つのコアモジュール

- 画像エンコーダ：（知覚）

ロボットのカメラ映像をLLMが理解できる「意味のベクトル」に圧縮・翻訳

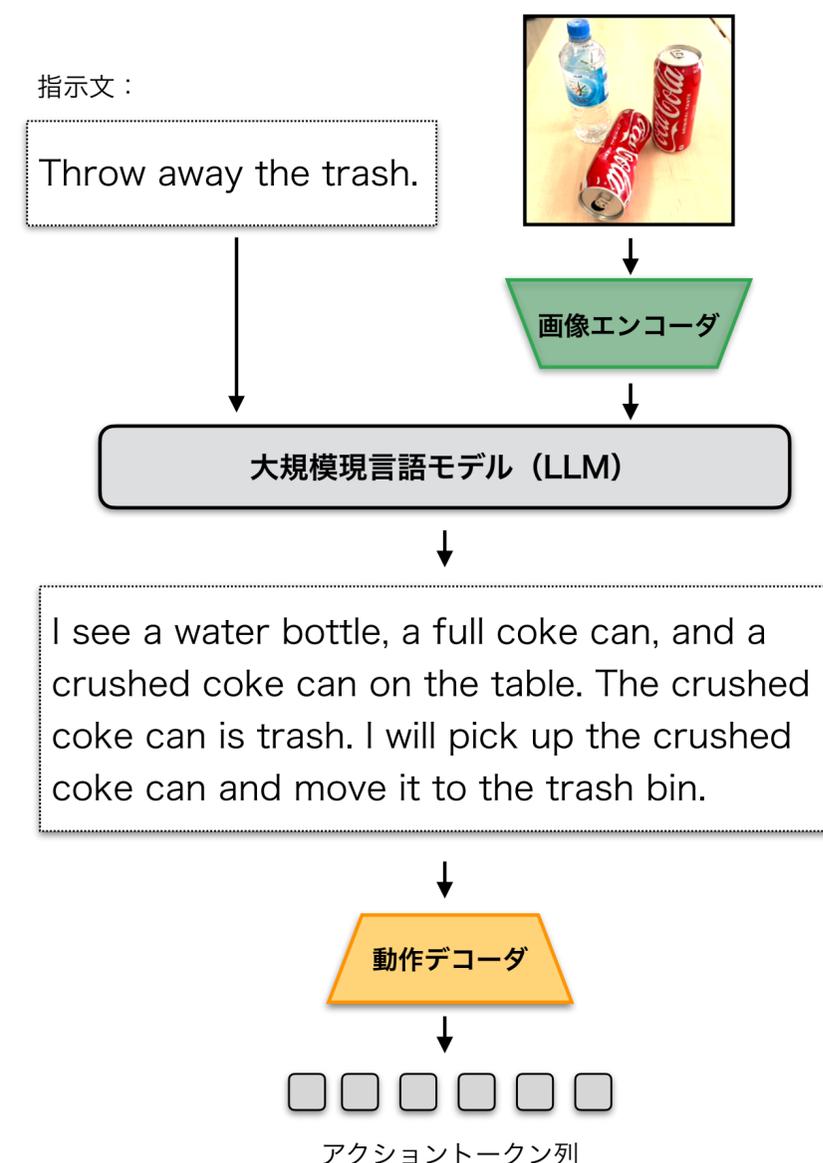
- LLM：（推論）

視覚情報と言語指示を統合し、次に行うべきことを推論・決定

- 動作デコーダ：（制御）

抽象的な「トークン」を、物理的な「ロボット制御信号」に変換

VLA



- GoogleのRT-2 (Robotic Transformer 2)
 - VLAの出力は推論 (Reasoning) と行動 (Action) の2パートで構成



Input Instruction: "Throw away the trash."

Model Output: I see a water bottle, a full coke can, and a crushed coke can on the table. The crushed coke can is trash. I will pick up the crushed coke can and move it to the trash bin.

<action_token_104> <action_token_045> <action_token_012> <action_token_255> <action_token_128> <action_token_001> <action_token_104> ...

1. 推論パート (Reasoning / Chain-of-Thought)

- モデルはまず自然言語で状況を整理

"I see a water bottle, a full coke can, and a crushed coke can... The crushed coke can is trash."

- ▶ 認識：画像内の3つの物体（水、新品コーラ、潰れたコーラ）を正しく認識
- ▶ 判断：「ゴミ＝潰れた缶」という常識的な結びつき (Semantic Understanding) を行う
 - VLAの強みであり、従来のロボット制御では難しい



Input Instruction: "Throw away the trash."

Model Output: I see a water bottle, a full coke can, and a crushed coke can on the table. The crushed coke can is trash. I will pick up the crushed coke can and move it to the trash bin.

<action_token_104> <action_token_045> <action_token_012> <action_token_255> <action_token_128> <action_token_001> <action_token_104> ...

2. 行動パート (Action Tokens)

- ロボットの手先位置やグリッパーの状態を表す数値を0~255などに離散化したもの

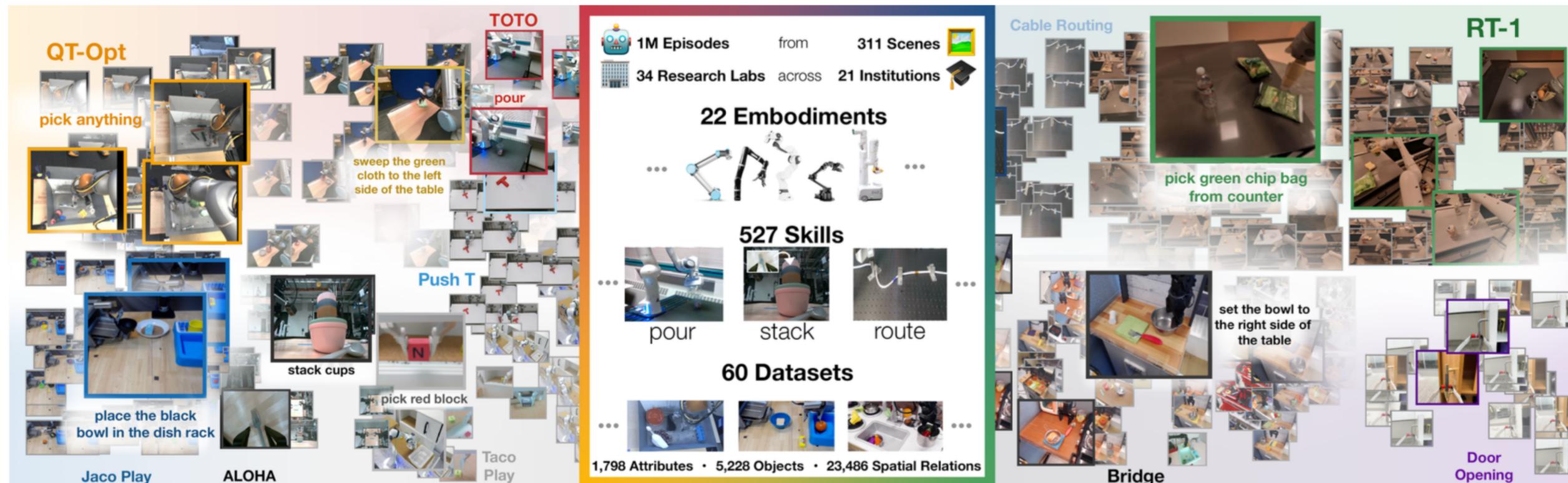
- ▶ 座標データ (x, y, z) : 「手前の潰れた缶」の位置へアームを移動させる数値
- ▶ 回転データ (roll, pitch, yaw) : 缶を上から掴むための手首の角度
- ▶ グリッパー状態 : 「開く(0)」もしくは「閉じる(1)」

- 実際にロボットがどう動くか (デコード結果 : トークン列)

1. Token Step 1-10 : アームを座標 (x=300, y=100, z=200) へ移動 (缶の真上)
2. Token Step 11-15 : グリッパーを全開にする
3. Token Step 16-20 : 座標 (x=300, y=100, z=0) へ下降 (缶の位置)
4. Token Step 21 : グリッパーを閉じる (ここで「新品」ではなく「潰れた缶」を選んでいるのが重要)
5. Token Step 22-30 : Z軸を高く上げる (安全確保)
6. Token Step 31-50 : 画面外のゴミ箱座標へ移動

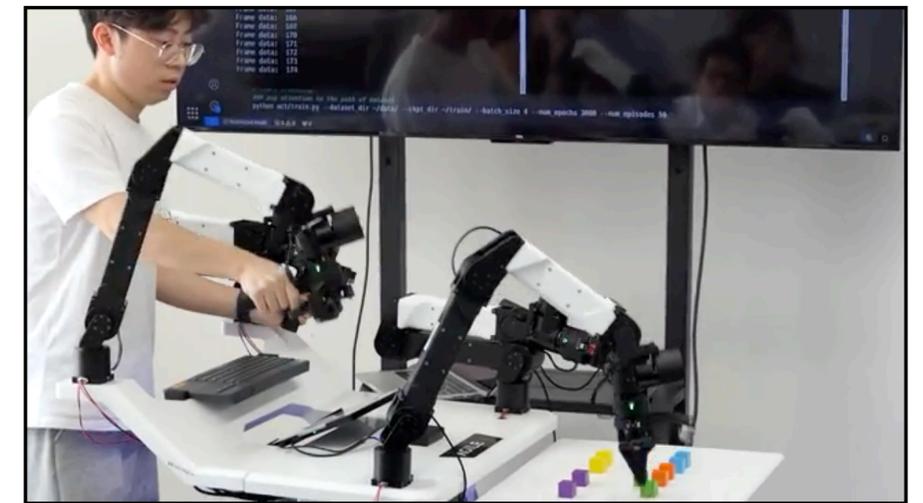
→ VLAは「どれがゴミか？」という認識と「どう腕を動かすか？」という制御を一つのニューラルネットワーク内で完結して出力

- 多様なロボットデータの統合：Open X-Embodiment (OXE)
 - 異機種間学習：22種類の異なるロボット（Franka, UR5, WidowXなど）のデータ（100万以上のエピソード）を混ぜて学習
 - 汎用的な物理法則：多くのロボットデータから「物を掴む」という行為の普遍的な概念を獲得
 - Co-training：知識（Webデータ）と技能（OXEデータ）を混ぜることで、破滅的忘却を防ぎながら身体性を獲得



<https://robotics-transformer-x.github.io/>

- ・ ターゲットロボットへの適応
 - 効率的な転移学習：ゼロから学ぶのではなく、すでに「掴む概念」を獲得したVLAモデルに「このアームでの掴み方」を微調整
 - アクション空間の調整：モデルが出力するトークンをロボット固有の関節角度やモーター指令に変換
- ・ ファインチューニング用データ
 - Teleoperation (遠隔操作)：人間がVRコントローラ等でロボットを操作し、正解の動き (デモ) を記録
 - データセットの構成
 - ・ 入力：カメラ画像 + 言語指示 (例：「赤いブロックを積んで」)
 - ・ 出力：一連のアクション (関節角度の軌道)
 - 多様性の確保：50~100件のデータでも、照明・背景・物体の配置を変えて収録することで汎用性が向上



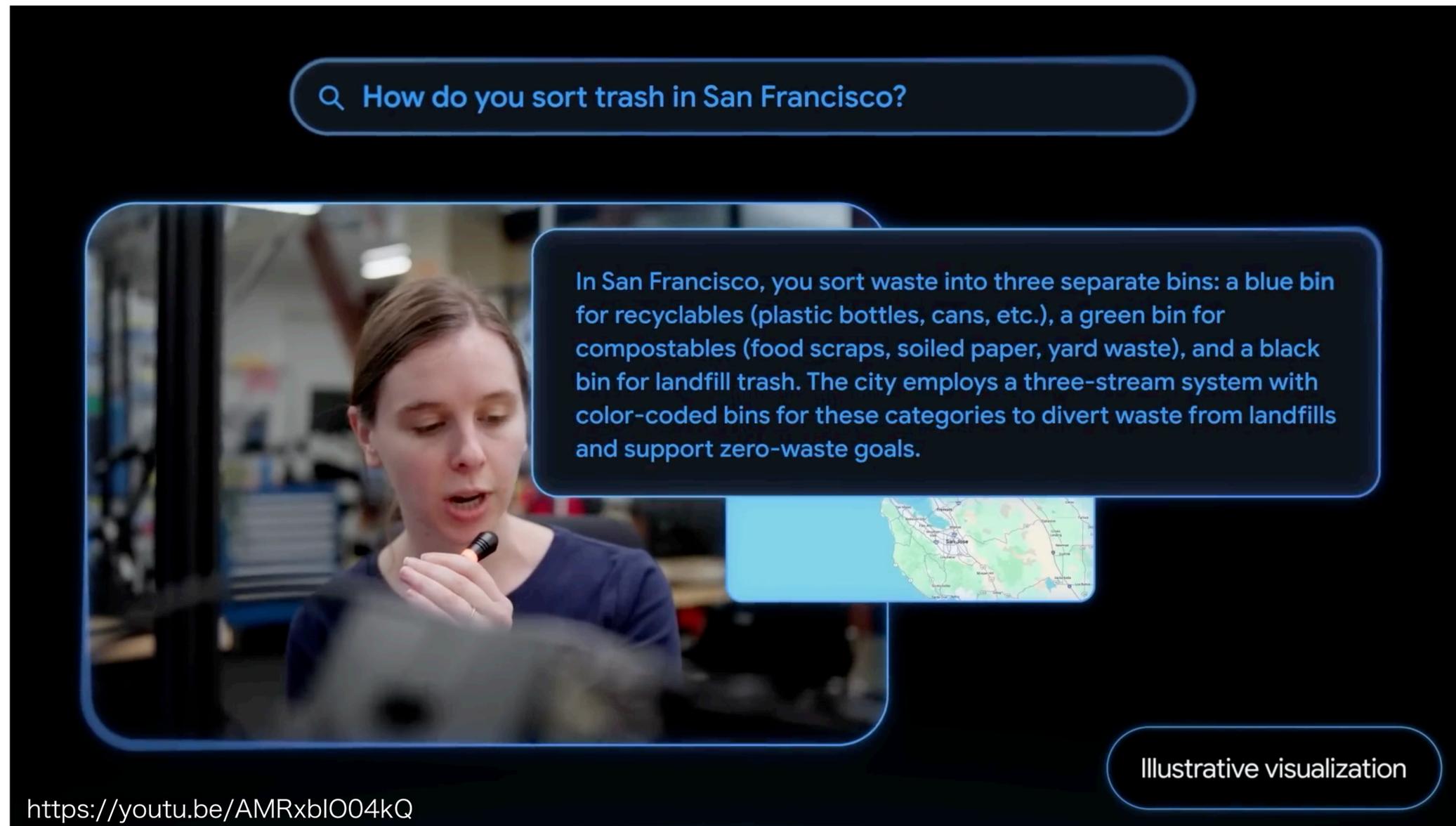
<https://www.youtube.com/watch?v=OXDy5YfnD78>

→ 人間がお手本を見せるTeleoperationデータを少量用意するだけで適応

- ・ 見たことのない物体やタスクの組み合わせにも対応可能
 - 把持タスクとして学習していないラグビーボールを把持する例
 - 言語指示：Can you pack the ball in the white bag ?



- ・ 外部知識を利用した適切な動作の実現
 - 地域のルール（外部知識）と捨てるべき対象は何か（VLMの世界知識）に従ってゴミを分別
 - 言語指示：How do you sort trash in San Francisco?



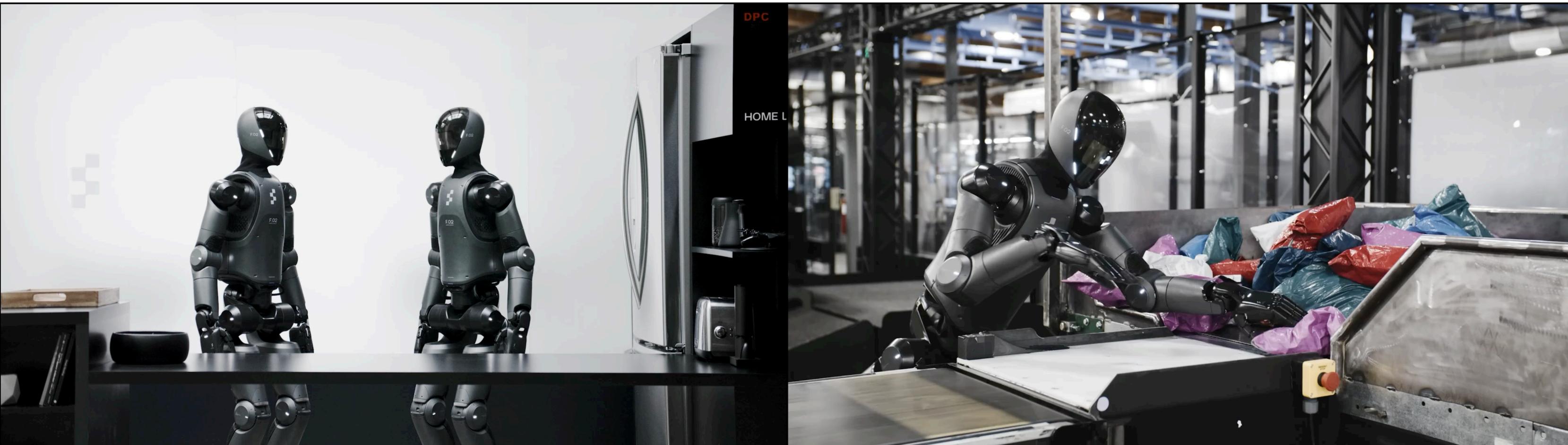
Q How do you sort trash in San Francisco?

In San Francisco, you sort waste into three separate bins: a blue bin for recyclables (plastic bottles, cans, etc.), a green bin for compostables (food scraps, soiled paper, yard waste), and a black bin for landfill trash. The city employs a three-stream system with color-coded bins for these categories to divert waste from landfills and support zero-waste goals.

Illustrative visualization

<https://youtu.be/AMRxbIO04kQ>

- ・ “見て理解し、言葉で指示を受け、状況に応じて人間のように行動する”
 - HELIX (Figure)



<https://www.youtube.com/watch?v=9pHk3A0H9hM>

<https://www.youtube.com/watch?v=lkc2y0yb89U>

1. マルチモーダル統合のさらなる進化

- テキスト・音声・映像・センサ情報等の視覚以外の情報と統合された「マルチモーダル認識」が主流
- VLM (Vision-Languageモデル) を起点とし、状況理解・意図理解・因果推論へと進化
(例：画像だけでなく、その背後の「なぜそう見えるか」までを推論するAI)

2. 世界モデルとの融合 (World Model + Vision)

- Vision Transformerと強化学習・世界モデル (World Models) の統合
- AIが「見たもの」をもとに、未来の状態を予測・シミュレートする能力
(例：DINO-WMなどが先駆的事例)

3. 空間・時間スケールの拡張

- 静止画 → 動画 → 長期的時系列へと拡張 (例：数十秒・数分先を予測する視覚AI)
- 3Dビジョン, NeRF, Gaussian Splattingといった技術の融合による空間認識

4. VLA(E2E)モデルにおける安全性の担保

- 物理的な安全性に関する強力なセマンティック理解

・ 設立目的

- ディープラーニングを中心とする技術による日本の産業競争力の向上
- 理事長：松尾豊（東京大学）



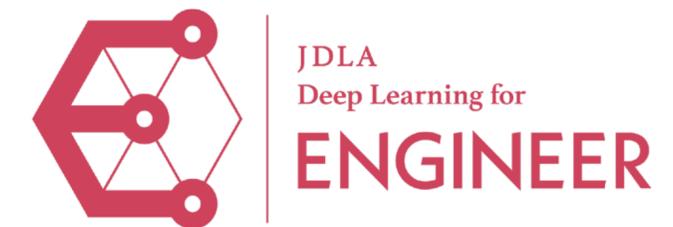
・ G検定

- AI・ディープラーニングの活用リテラシー習得のための検定試験
- 年間6回：1月・3月・5月・7月・9月・11月の開催



・ E資格

- ディープラーニングの理論を理解し，適切な手法を選択して実装する能力や知識を有しているかを認定
- 受験資格：JDLA認定プログラムを試験日の過去2年以内に修了



- ・ AIの基礎を体系的に理解し実践的技術を学ぶ
 - 2コマ（16時～19時の3時間）× 全15回
 - 講義+プログラム演習, プロジェクト課題



CU Synergy-Program：AI基礎と応用



		内容
第1回	4月5日(火)	AIと機械学習(1)
第2回	4月12日(火)	AIと機械学習(2)
第3回	4月19日(火)	AIと機械学習(3)
第4回	4月26日(火)	ニューラルネットワーク基礎
第5回	5月10日(火)	畳み込みニューラルネットワークの基本
第6回	5月17日(火)	畳み込みニューラルネットワークの発展
第7回	5月24日(火)	畳み込みニューラルネットワークの応用
第8回	5月31日(火)	時系列データに適したディープラーニング

		内容
第9回	6月7日(火)	深層生成モデル(1)
第10回	6月14日(火)	深層生成モデル(2)
第11回	6月21日(火)	深層強化学習(1)
第12回	6月28日(火)	深層強化学習(2)
第13回	7月5日(火)	プロジェクト課題(1)
第14回	7月12日(火)	プロジェクト課題(2)
第15回	7月19日(火)	AIの今後



- ・ 深層学習を利活用するための技術と最新動向を学ぶ
 - 深層学習技術の事業化・ビジネス活用のワークショップ（外部講師）
 - 講義+プログラム演習, アドバンス・プロジェクト課題（全8回）



CU Synergy-Program：AI基礎と応用・アドバンスコース



		内容	担当
第1回	10月4日(火)	知識蒸留+知識転移グラフ	藤吉
第2回	11月1日(火)	半教師あり学習、自己教師あり学習	山下, 平川
第3回	11月15日(火)	グラフニューラル ネットワーク	山下
第4回	11月29日(火)	Vision Transformer	藤吉
第5回	12月6日(火)	深層学習の事業化・ビジネス活用	外部講師
第6回	12月20日(火)	深層強化学習	山下, 平川
第7回	1月10日(火)	アドバンス・プロジェクト 課題1	藤吉, 山下, 平川
第8回	1月24日(火)	アドバンス・プロジェクト 課題2	藤吉, 山下, 平川

A screenshot of a YouTube video player. The video shows a man in a white shirt standing in front of the entrance to Bldg. 16, which is the Machine Perception and Robotics Group (MPRG) building. A red car is parked to the right. The video title is 'MPRG Virtual 360° Lab Tour' and it is marked as '限定公開' (Limited Public). The video has 6 views and was uploaded on 2020/06/25. The channel name is 'MPRG' and there is a 'チャンネル登録' (Subscribe) button.

<https://www.youtube.com/watch?v=72GkuKF5T2E>