Multi-modal LLM

video recognition & vision and Language group

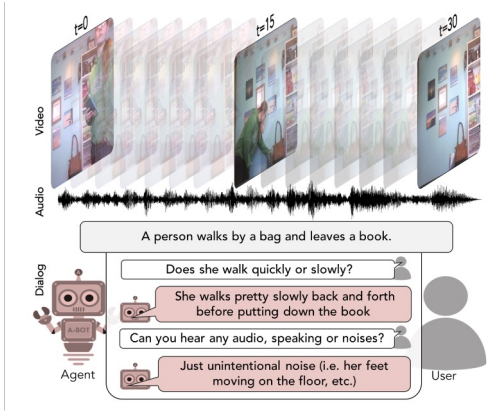
サーベイ資料の構成

- **目的：**AI分野のFuture Topicsとなるものを導く
- **サーベイ内容 1：GPT関連**
  LLM系の研究がどうなっているのか、どうやって自分の研究に使うか
- **サーベイ内容 2：Multimodal LLM関連**
  Multimodal LLMのテーマ設定、手法、データセット
- **サーベイ内容 3：LLM+Robotics in ICRA 2023**
  ロボティクスのトップ会議ICRA 2023でLLM + Roboticsの速報
- **サーベイ内容 4：国・組織の声明**
  欧州のAI規制案に関して述べる

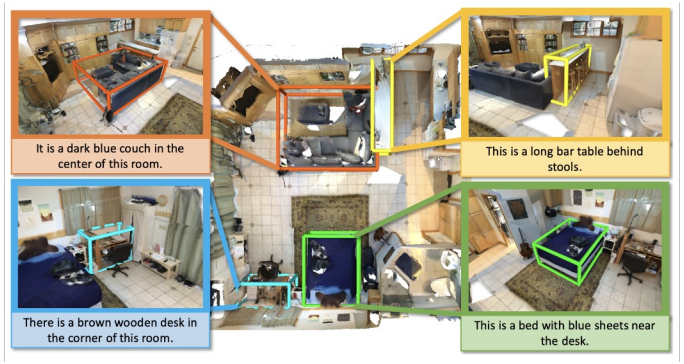
Multi-modal タスク（ Vision ＋Language ）

Computer Visionでのmulti-modal + languageの検討：

Visual Question Answering [Antol+, ICCV'15]
Image ＋ Language



Audio Visual Scene-Aware Dialog
[Alamri+, CVPR'19]
Video + Audio + Language

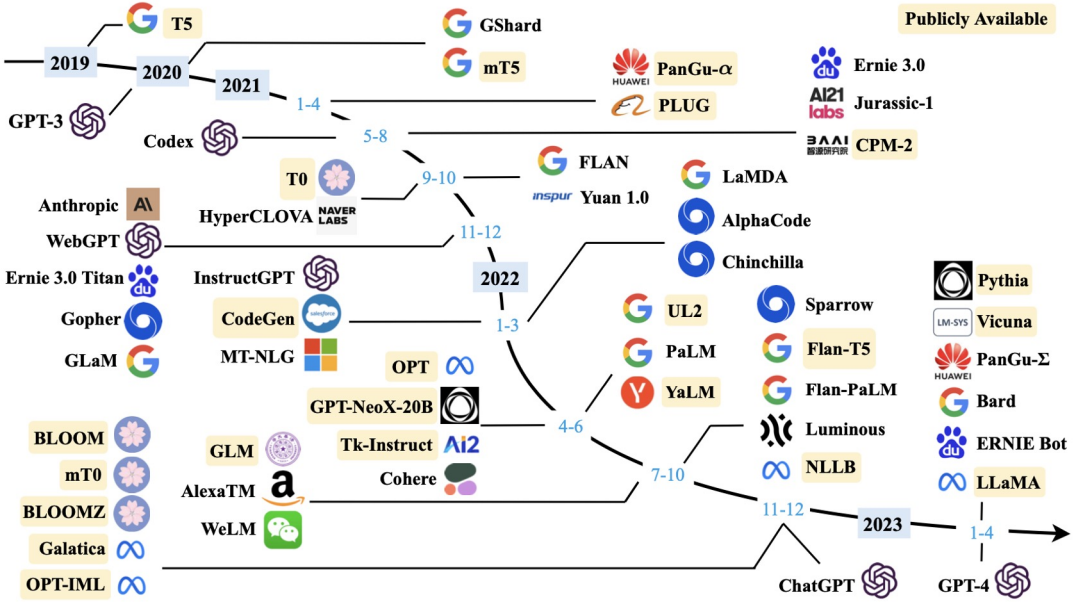


ScanRefer [Chen+, ECCV'20]
3D + Language

視覚・音声などのmulti-modalの詳細情報を言語を介して理解する。形式としては、question answering, dialog, summarizationなどがある。

LLMの発展

GPT系を代表としたLLMが迅速的に発展

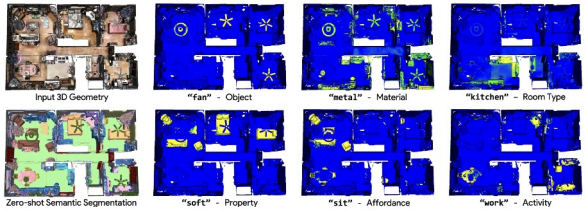


LLMのタイムライン（ Zhao, Wayne Xin, et al. "A survey of large language models." *arXiv preprint arXiv:2303.18223* (2023). ）

LLMがMulti-modal タスクへの導入例 (1/2)

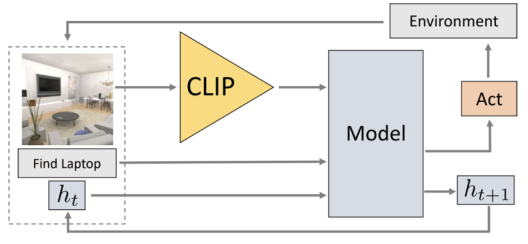
- CLIPがMulti-modalタスクの主流特徴抽出器に：

(CLIPのサーベイ参照資料：https://xn--techblog-y63gww.exawizards.com/entry/2023/05/10/055218)



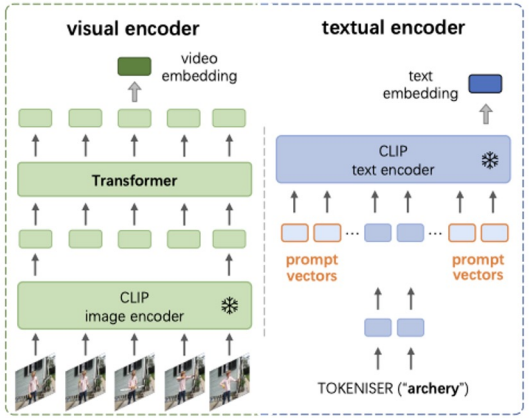
OpenScene 3D Scene Understanding with Open Vocabularies

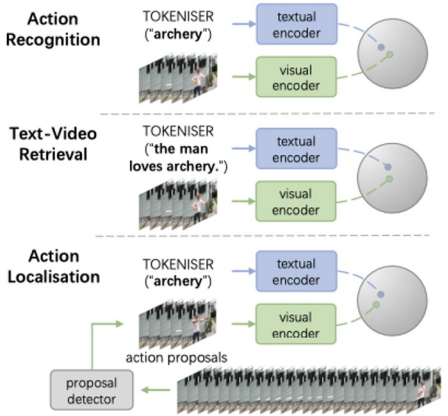
CLIPを3D点群＋言語タスクへ導入し、
Zero-shotで点群の認識



Simple but Effective: CLIP Embeddings for Embodied AI

CLIPをEmbodied AIタスクへ導入し、
シンプルなモデルでSOTA達成





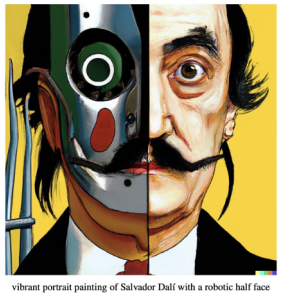
Prompting Visual-Language Models for Efficient Video Understanding

CLIPをVideo＋Languageタスクへ導入し、
優れた性能でZero-shotも対応

LLMがMulti-modal タスクへの導入例（2/2）

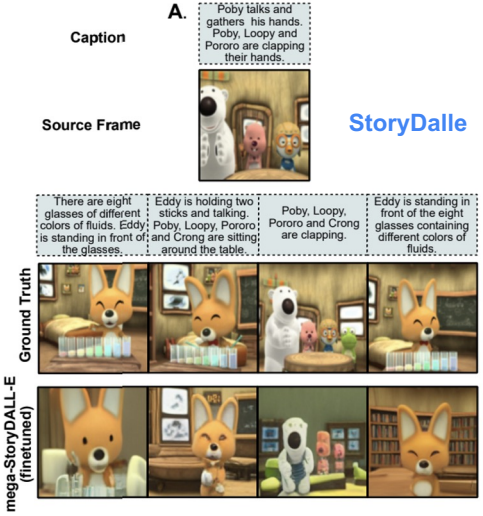
- DALL・E1,2, Diffusion modelによりText-to-Imageがホット：

DALLE 2



GLIGEN OpenSet Grounded Text-to-image Generation

StoryDalle



DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

LLMの導入の影響 (1/4)

LLMの導入によりモデル評価問題、信頼性問題の色々（1/2）：

- これまで基本的なタスクだったClassificationが置き換えられる?
  - テキスト出力をZero-shotのような形式で評価するのが基本となる?
    - ベンチマークとしてはClassificationの方が扱いやすそうなので残る可能性も?
  - テキスト出力の場合の評価方法
    - 最近のLLMの生成文の評価、Referenceがあると、BLEUとかもLLMトレンド前のように使われている
    - 生成した文章で、LLMでもう一回質問応答で評価するなどもある。
    - ChatGPTで評価するとか、人の評価との相関が高いという説がある。
    - 新しい指標が出ているわけではない。
    - 3次元でLLMを利用する際に一度3次元からDenseCaptioningをして、そこからの出力をLLMに入力して、Zero−shotを実現、
      - Captioningの精度の方がもはや重要

- アノテーションの質どうやって保つ、LLMに頼りすぎると複雑になってくる。
  - Active Learningとか、
  - 優秀なAnnotatorをどうやって訓練とか
  - データを蒸留し、優れたデータを残していく。

LLMの導入の影響 (2/4)

LLMの導入によりモデル評価問題、信頼性問題の色々（2/2）：

- 評価時に用いるデータが既にLLMの学習データに含まれているという懸念
  - LLMのマルチモーダルタスクへの導入が進むことで言語分野だけでなく画像などその他の分野でも評価の信頼性という点で問題が出る可能性あり
    - 学習データが明確にされているLLMが重宝される?
    - そもそもLLM導入以前から評価の適切さについて疑問視されているところもあり...
  - 適切な評価方法の確立は今後の研究において重要な点の一つ
    - もしくはアプリケーション寄りの問題設定で評価する研究が増える?

- 使う時にLLMなのか、人間なのかが判断できない？

- ChatGPTは言語情報の理解が望ましい→
  - 分析
  - 言語理解が正しい方向で理解されているか

LLMの導入の影響 (3/4)

社会への影響 (1/2) :

- ChatGPTなどを代表としたChat AIが普及してきた。
  - 今後もChat AIが伸びていく。

- ChatGPTの導入により、情報格差がどうなる？
  - 縮まったか、開いたか、意見がバラバラ
  - ChatGPTが上手く使えない方が圧倒的に多いかも
  - 信頼性を検討せずに使っている人が多い
    - ChatGPTを信用しすぎて、様々なソースから情報を確認しない人がいる
    - 現状のChatGPTが堂々と間違えた情報を出す場合がある

LLMの導入の影響 (4/4)

社会への影響 (2/2) :

- 生成画像やテキストの利用について：
  - EUの声明で：
    - 使用には危険視されず
    - データの透明性が重要視
    - ChatBotから生成したものですよ”とタグつける主張
  - 日本では：
    - まだgray zoneな場合があり、手修正加えたら人間だと
  - ICCVなどの国際会議査読：
    - ChatGPTなどを使わない
  - イラストだと色々問題あるが、テキスト単体では製造物として楽しめない

- 危惧する面もある：
  - 詐欺に悪用。

Post-LLM時代でMulti-modalの課題設定(1/3)

- AGIを目指すためには：
  - 色々なモダリティ、あらゆる情報を入れていくべき（すでに色々検討されてきている）。
  - マルチモーダルの問題点：
    - あらゆるSensorsとTextのAlignmentが課題になる。
    - インタネットでデータがある（画像など）が、他のモダリティがあまりリソースがない
    - どうやってあらゆるモダリティのEncodingも問題があるかも？
  - 新しいモダリティの導入で活躍させるためにの技術系の研究（どうやって異なるセンサーデータ上手く扱う）
    - 新しいモダリティLLMでZero一shot認識
    - モダリティの中間表現がどのように学習する？
    - モダリティの中間表現の分析など（可視化など）が新たなテーマになる
    - 時系列情報ではないモダリティもあり、どうやって対応するのが難しい面がある
    - 大規模化できないセンターデータもある？
      - 小規模データで学習できる？
        - LIMA 質が高いPromptsを用意することで、
    - 新しいモダリティの導入で”新規な能力”がある？これまでに全然できなかったことなど
  - AutoGPTなどのように、外部（例えば、Searching Engineを入れたり）
  - 人間の行動情報を入れていく。
    - これまでには、Object Trackingでダイレクト
    - 行動履歴のようなデータ
    - メタバース的な観点、
    - 一回バーチャルでやるのが先かも。

Post-LLM時代でMulti-modalの課題設定(2/3)

- 大規模モデルでデータセットを集める。
  - ビデオ詳細認識のデータセット作成
  - BLIP2
  - QAmodel、ChatGPT
  - 説明文から編集文をつくる（InstructPix2Pix）
  - 簡単なテキストでデータでLLMで拡張して学習。

- CLIPの特徴（自然言語に紐づいた特徴）を利用して、他の研究に利用する。
  - 画像生成
  - Embodied AI
  - Open vocabularyな（語彙に制約されない）画像認識や物体検出

- LLMがreasoning能力、reasoning能力を使って高レベルの文章をより詳細的に分解してRobotics応用。
  - InstructionをChatGPTでより詳細的なStep一by一Step

- 人間の感性理解・Social Intelligenceなども課題の視野に。
- インタラクション的な方向性へシフト。
- ダイナミックで情報を収集していく。
- Promptの使用が気になり、Promptの設計が大事。

Post-LLM時代でMulti-modalの課題設定(3/3)

- マルチAgent、LLMを搭載したAgentとか
  - 現実に近いような世界がLLMだとシミュレーションできたら、

- 個人のChatBotを作る
  - 自分っぽいものを作る
    - 今はみんながSLACKや、メールデータで学習させるのが多い
  - 目的：
    - お医者さんへ連絡をしてくれる
    - 忙しい時に会議に出てもらう
    - 自分の代わりに働いてくれる
    - 有名な学者、、
  - 一週間分のマルチモーダルデータ（対話など）

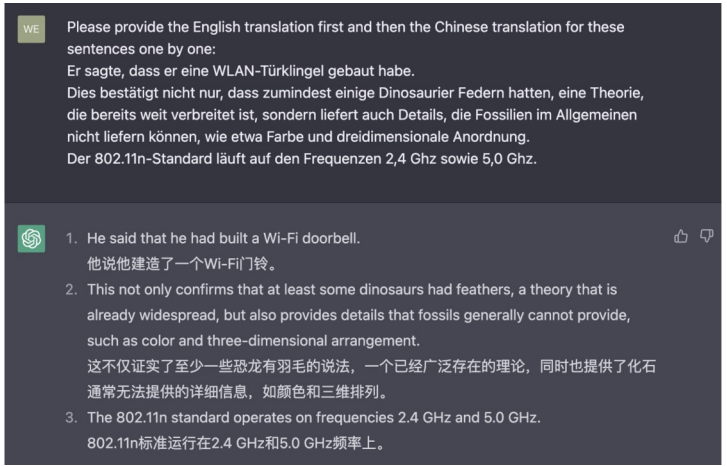
- 画像が言語により豊かの情報を持つ、テキスト情報では埋められない
  - data 構造で差を縮む：
    - Scene Graph：3次元的な情報
    - AMRepresentation：抽象的な構造表現
    - LLMではテキスト（構造化と言えるでしょう）とコードとか
      - 自動的にされているとは：データの裏にある構造が学習できる
      - 階層性：
    - Reasoningなどの複雑なタスクにおいて、言語の構造が良くない場合がある、、
    - Arrival（Science Fiction）：

下記からは論文まとめ

*GPT*関連

タイトル：Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine

- **概要**：ChatGPT (GPT-4) に翻訳を依頼した場合の精度を既存の商業システムと比較
- **新規性**：言語圏によってChatGPTの翻訳精度が大きく異なることを示した。またPivot (中間言語を挟んで翻訳する）を使用することでChatGPTの精度が向上することを示した。
- ChatGPT を翻訳器として使用することについては他にも論文が存在し、文法間違いや温度の違いによる翻訳精度の違いが調査されている
  - Unleashing the Power of ChatGPT for Translation: An Empirical Study
  - Towards Making the Most of ChatGPT for Machine Translation



まとめBY: Takuto Nariura

Scaling Transformer to 1M tokens and beyond with RMT

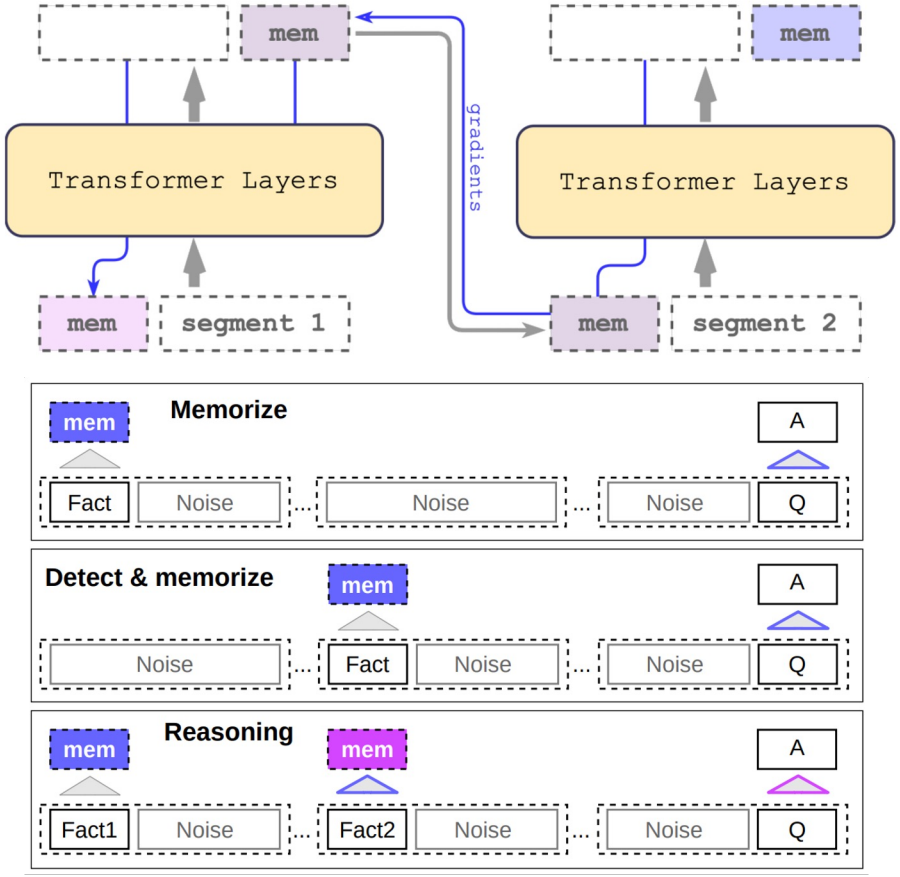
会議：arXiv preprint
著者：Aydar Bulatov, Yuri Kuratov, Mikhail S. Burtsev

概要：Transformer (BERT: 系列長512)に再帰的なメモリを導入することで有効系列長を200万まで増やした

実験：

- bAblで人工タスク (右図参照→)
- QuALITYでlong QA

知見：5セグメント(512x再帰5ループ)以上のタスクで訓練すると、それ以上の系列長にも汎化し始める



まとめBY: Seitaro Shinagawa

LIMA: Less Is More for Alignment

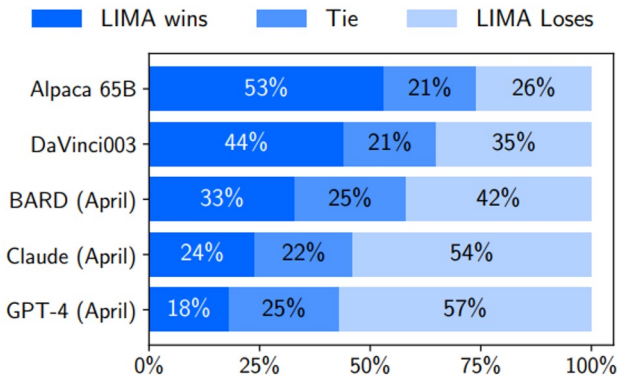
会議：arXiv preprint

オープンなLLMであるLLaMa-65Bを慎重に選んだ1,000サンプルのプロンプトで微調整するとGPT-4と人手評価で遜色ない結果を得られた

- 750サンプルはQAサイトの上位の質問回答から、250サンプルは著者が全体の質や多様性を考えて追加
- LIMAの出力がGPTより好まれた回答は43%

Source	#Examples	Avg Input Len.	Avg Output Len.
Training			
Stack Exchange (STEM)	200	117	523
Stack Exchange (Other)	200	119	530
wikiHow	200	12	1,811
Pushshift r/WritingPrompts	150	34	274
Natural Instructions	50	236	92
Paper Authors (Group A)	200	40	334
Dev			
Paper Authors (Group A)	50	36	N/A
Test			
Pushshift r/AskReddit	70	30	N/A
Paper Authors (Group B)	230	31	N/A

プロンプトのソース元



LIMAと他のLLMとの比較

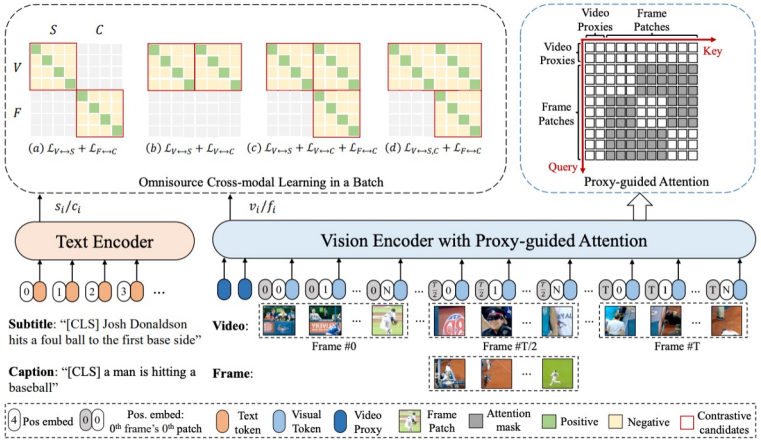
まとめBY: Seitaro Shinagawa

タイトル：CLIP-VIP: Adaptinig Pre-trained Image-Text Model to Video-Language Representation Alignment

会議 :ICLR 2023

著者：

- 概要：画像-テキストの事前学習モデルであるCLIPを動画-テキストタスクに適応する（post pre-training）ことを試みた論文。適応するのが難しい原因を分析し、またその分析に基づいた手法を提案
- 新規性：初めて、画像-テキストの事前学習モデルをビデオに適用する際の分析を怒った
- データ数が少ないと、CLIPの事前学習で学んだことを忘れてしまう＆over fittingする。また、画像とビデオのテキストに関するドメインギャップが大きいと適応が難しい。
- 1. 多様なビデオデータセットを用いる 2. 補助キャプションを用意、3. ビデオ全体、フレーム、キャプション、補助キャプションの4つを使ったロスの提案



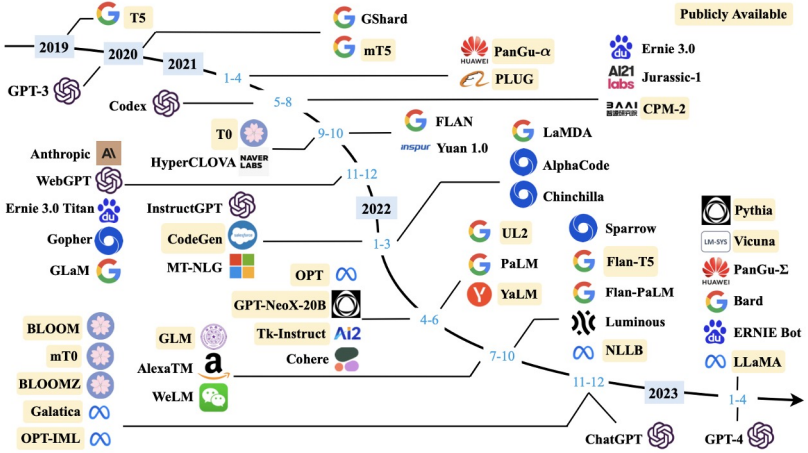
まとめBY:

A Survey of Large Language Models (arXiv 2023)

概要：最近のLLMs（パラメータ数が数百億以上の事前学習言語モデルを意味する）を網羅的にサーベイした論文。LLMsの背景、LLMに関する重要な技術、LLMの使用、LLMsの評価、について詳細的に述べた。また、LLMsを活用する際に有用な情報として、主流なLLMsモデル、学習する際に使われたデータセットなどのリソースについても述べた。

LLMのEmergent ability：BERTなどの小型LMに出現しない、LLMなどのパラメータ数や学習データが膨大になる際に出現する下記らの能力：

- **In-context learning:** 指示やタスクのデモなどから学習せずにテストインスタンスを補完できる。
- **Instruction following:** instruction tuningした後に、instruction guidedでunseenタスクに汎化可能。
- **Step-by-step reasoning:** chain-of-thought推理を利用し中間の推理ステップが提示された場合に問題を解決できる。



LLMsのタイムライン

まとめBY: Qiu Yue

A Survey of Large Language Models (arXiv 2023)

LLMの鍵となる技術：

- **Scaling：**パラメータ数、学習データサイズ、学習の量、学習効率、データの質がモデル性能に重要。
- **Training：**distributed trainingや最適化のtrickなどが重要。
- **Ability eliciting:** モデルの推理能力を誘導するために、task instructionsやchain-of-thought promptingなどが重要。
- **Alignment tuning：**LLMsが学習データから有害、biased的な情報を学習する可能性があるため、人間が介してモデルの評価や学習を誘導する**R**einforcement **L**earning with **H**uman **F**eedbackなどが重要。
- **Tools manipulation:** 数値計算、サーチエンジンで新しい知識を学習など外部ツールでLLMsを強める。

Corpora	Size	Source	Latest Update Time
BookCorpus [122]	5GB	Books	Dec-2015
Gutenberg [123]	-	Books	Dec-2021
C4 [72]	800GB	CommonCrawl	Apr-2019
CC-Stories-R [124]	31GB	CommonCrawl	Sep-2019
CC-NEWS [27]	78GB	CommonCrawl	Feb-2019
REALNEWs [125]	120GB	CommonCrawl	Apr-2019
OpenWebText [126]	38GB	Reddit links	Mar-2023
Pushift.io [127]	-	Reddit links	Mar-2023
Wikipedia [128]	-	Wikipedia	Mar-2023
BigQuery [129]	-	Codes	Mar-2023
the Pile [130]	800GB	Other	Dec-2020
ROOTS [131]	1.6TB	Other	Jun-2022

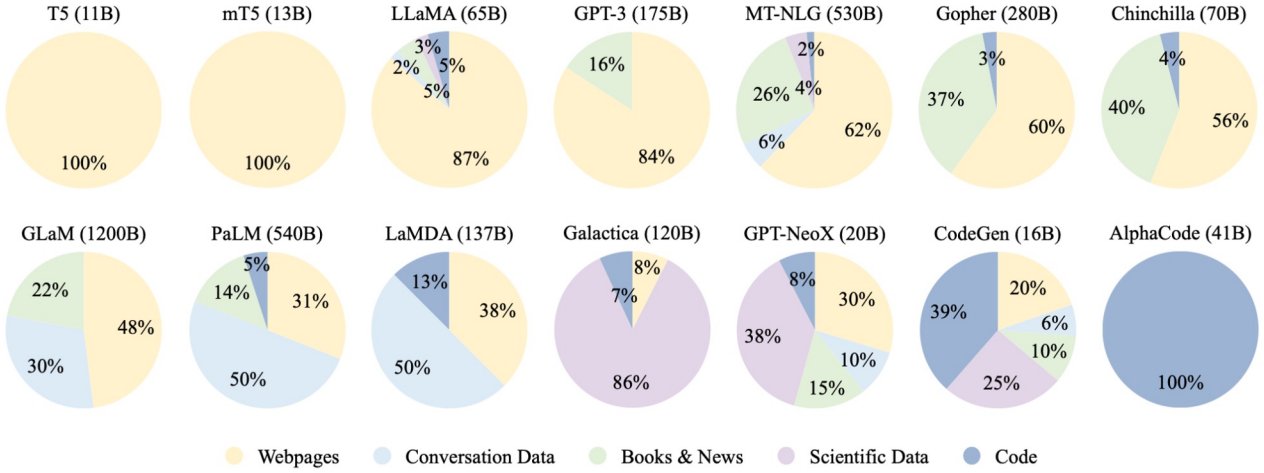
LLMsの学習に広く使われている Corpora

まとめBY: Qiu Yue

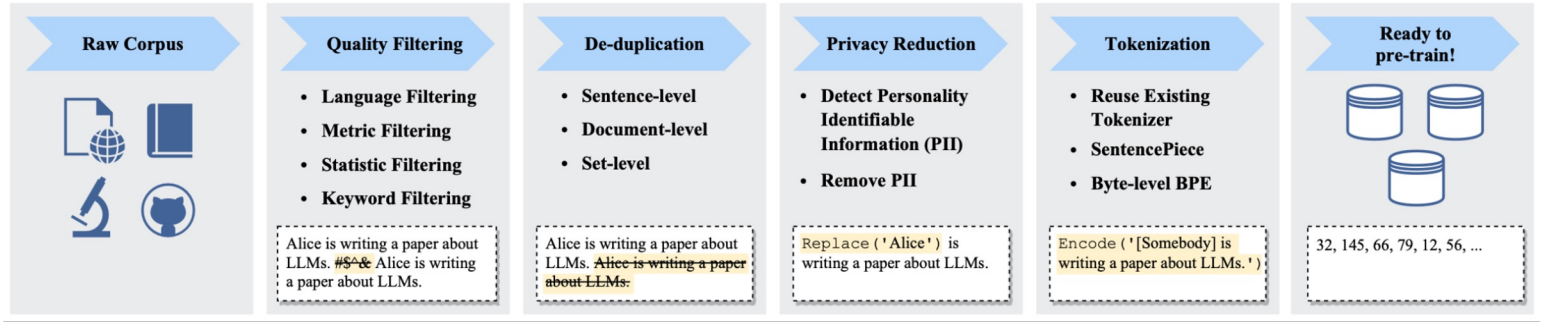
A Survey of Large Language Models (arXiv 2023)

Pretrainデータに関して：

データ分布：LLMsの学習に大量な言語データが必要となる。右図で現在のLLMs pre-training時に使ったデータの分布を示す。



データ処理プロセス：学習データの質を高めながら、学習可能にデータをトークン化するなど。



まとめBY: Qiu Yue

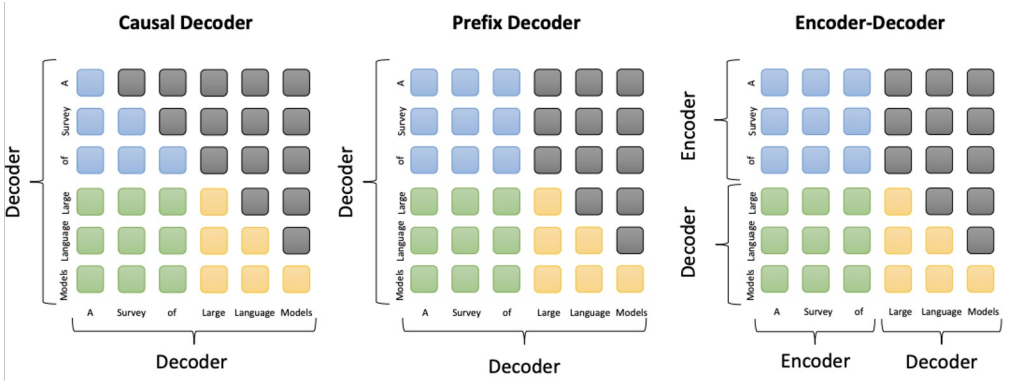
A Survey of Large Language Models (arXiv 2023)

LLMsのモデル構成：

LLMsの model configuration：

Model	Category	Size	Normalization	PE	Activation	Bias	#L	#H	d_{model}	MCL
GPT3 [55]	Causal decoder	175B	Pre Layer Norm	Learned	GeLU	✓	96	96	12288	2048
PanGU- α [74]	Causal decoder	207B	Pre Layer Norm	Learned	GeLU	✓	64	128	16384	1024
OPT [80]	Causal decoder	175B	Pre Layer Norm	Learned	ReLU	✓	96	96	12288	2048
PaLM [56]	Causal decoder	540B	Pre Layer Norm	RoPE	SwiGLU	×	118	48	18432	2048
BLOOM [68]	Causal decoder	176B	Pre Layer Norm	ALiBi	GeLU	✓	70	112	14336	2048
MT-NLG [97]	Causal decoder	530B	-	-	-	-	105	128	20480	2048
Gopher [59]	Causal decoder	280B	Pre RMS Norm	Relative	-	-	80	128	16384	2048
Chinchilla [34]	Causal decoder	70B	Pre RMS Norm	Relative	-	-	80	64	8192	-
Galactica [35]	Causal decoder	120B	Pre Layer Norm	Learned	GeLU	×	96	80	10240	2048
LaMDA [96]	Causal decoder	137B	-	Relative	GeGLU	-	64	128	8192	-
Jurassic-1 [90]	Causal decoder	178B	Pre Layer Norm	Learned	GeLU	✓	76	96	13824	2048
LLaMA [57]	Causal decoder	65B	Pre RMS Norm	RoPE	SwiGLU	✓	80	64	8192	2048
GLM-130B [82]	Prefix decoder	130B	Post Deep Norm	RoPE	GeGLU	✓	70	96	12288	2048
T5 [72]	Encoder-decoder	11B	Pre RMS Norm	Relative	ReLU	×	24	128	1024	512

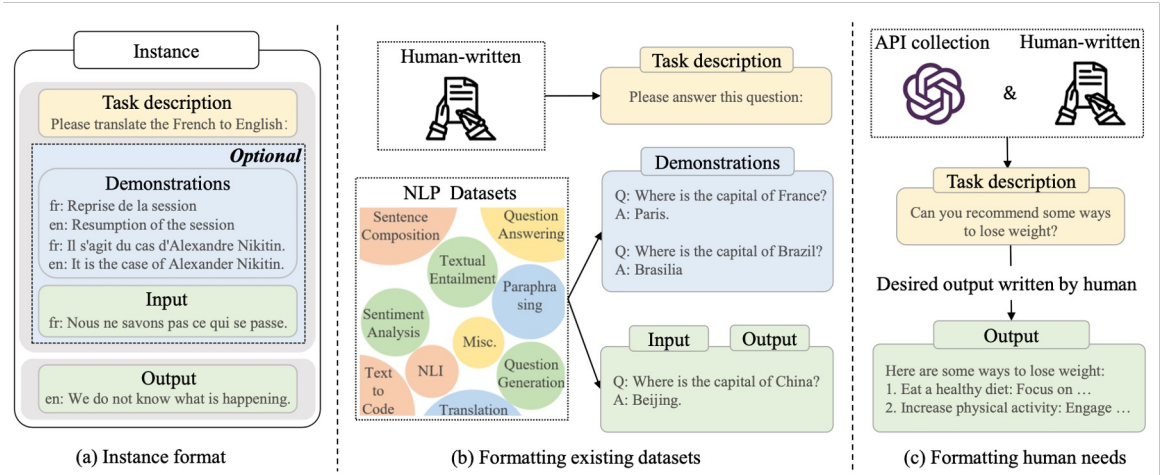
3種類のモデル構造：



まとめBY: Qiu Yue

A Survey of Large Language Models (arXiv 2023)

LLMsの重要技術１Instruction tuning：データセット全体ではなく、instructionのフォーマットで小さいfine-tuningデータ集を構築してfine-tuningする方がLLMsの性能を破壊せずに、性能の向上や未知タスクへの汎化性能を高められる。



まとめBY: Qiu Yue

LLMsの重要技術2 Alignment tuning：人間のフィードバックによりリワードモデルを学習。その上、人間にアラインしたRL fine-tuningを行う。

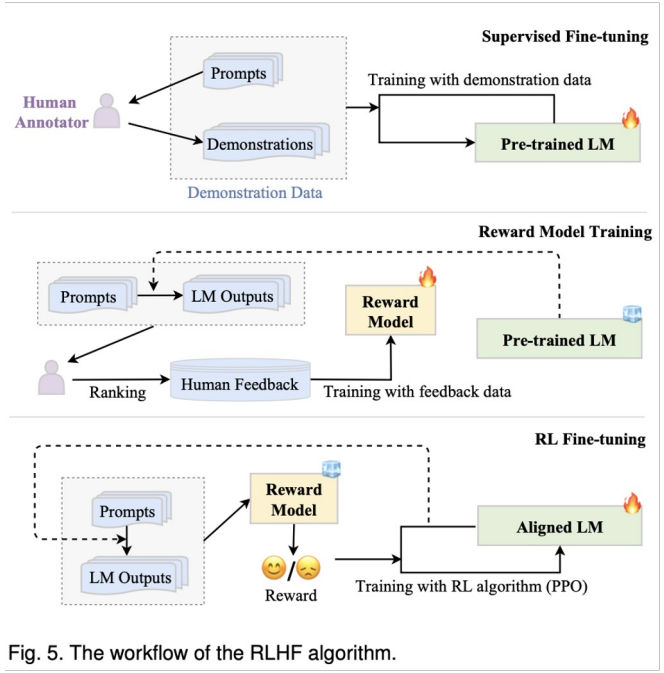
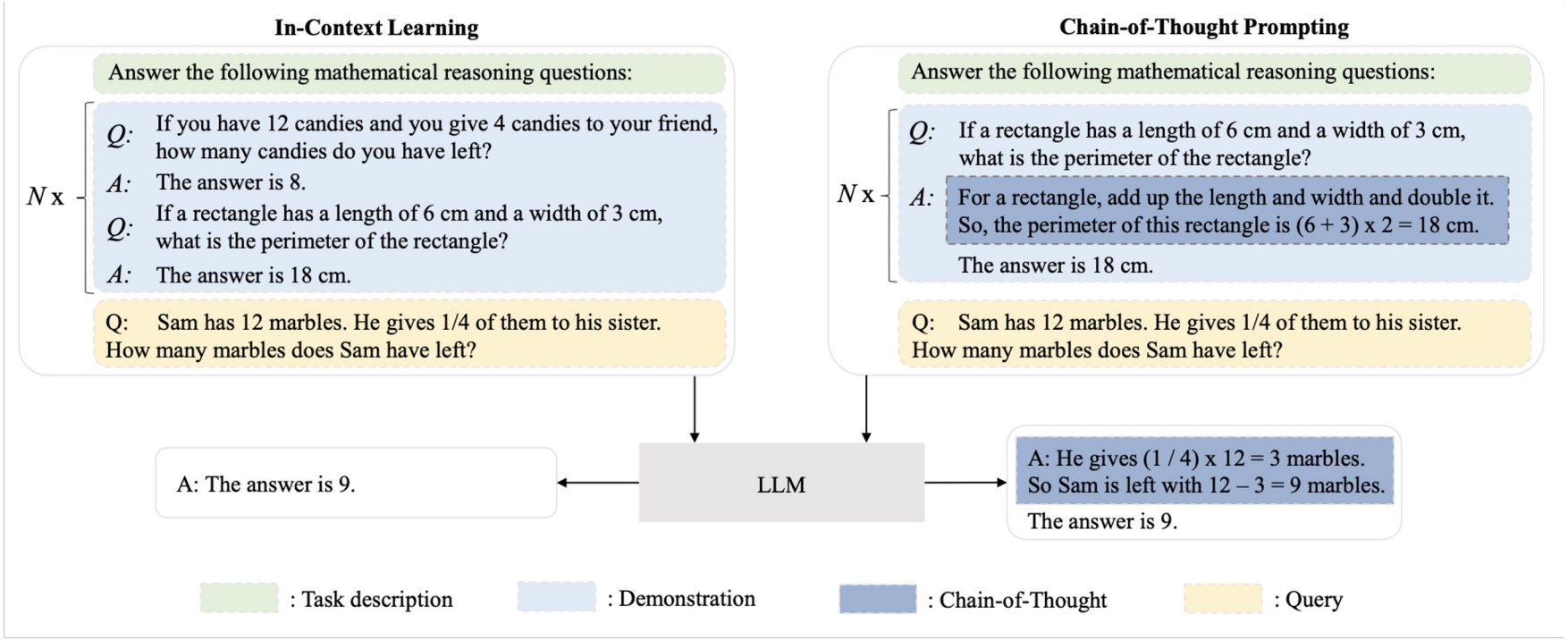


Fig. 5. The workflow of the RLHF algorithm.

A Survey of Large Language Models (arXiv 2023)

LLMsの使用：In-context learningやchain-of-thought などを利用したpromptingで学習ずみのLLMsを使用。



まとめBY: Qiu Yue

LLMsのfuture directions :

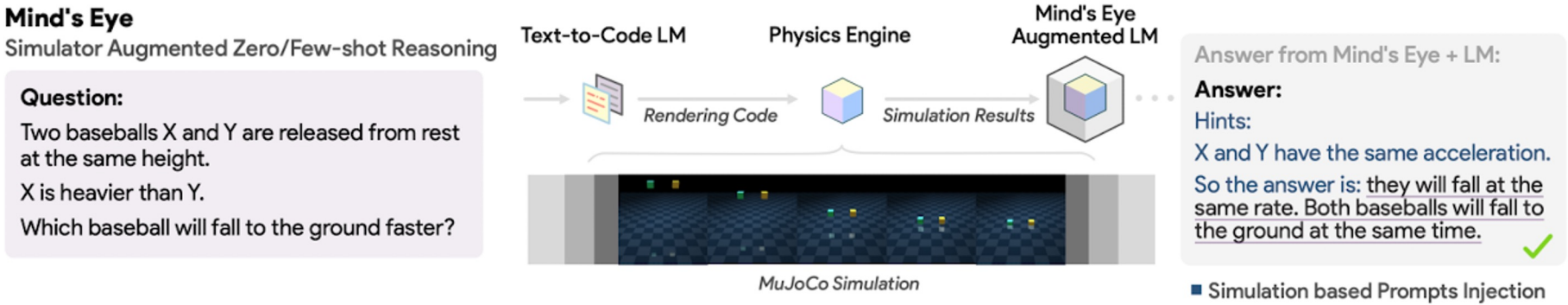
- **theory and principle:** 情報がどうやって膨大なパラメータを持つLLMsの内部に組織・利用されているのか、emergent abilitiesはどうして出現したか？などを検討する。
- **model architecture:** モデルの効率・性能・容量を高める研究。
- **model training:** システム的、経済的な事前学習、モデルの早期問題診断、ハードウェアリソースの配分などを検討。
- **model utilization:** 現在promptの設計にはlabor costが高い。また、複雑な数式・数値計算など言語で表しにくい知識のprompt、より複雑な問題を解決するために対話のような形式でpromptの検討など。
- **safety and alignment:** LLMsの信頼度・透明性を高める、有害情報を防ぐなどの検討。
- **application and ecosystem:** 様々な応用・ツールを融合したLLMsを利用したエコシステムの検討。

まとめBY: Qiu Yue

Mind's Eye: Grounded Language Model Reasoning through Simulation (arXiv 2022)

背景：現在のLLMは物理世界の理解を必要とする質問に対しては誤った回答をすることがある。
データセット：教科書レベルの物理学に関する質問応答データセットUTOPIA
結果：Step-by-StepやChain-of-Thoughtのようなプロンプトベースの方法と比べて正しく推論することができた。
シミュレーション結果をわざと間違えたり、問題と関係ない項目をプロンプトに含めたりすると精度が下がった。

Ablation Settings	Zero-shot	Few-shot
Mind's Eye (*default*)	51.9 $_{1.73}$	84.2 $_{0.79}$
Mismatched simulation results	30.4 $_{1.65}$	54.3 $_{1.28}$
Missing trigger words	51.0 $_{1.12}$	83.0 $_{0.73}$
Incorrect simulation	24.6 $_{1.73}$	39.6 $_{2.89}$

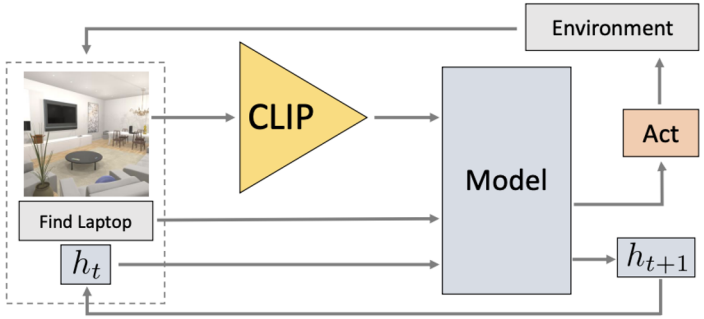


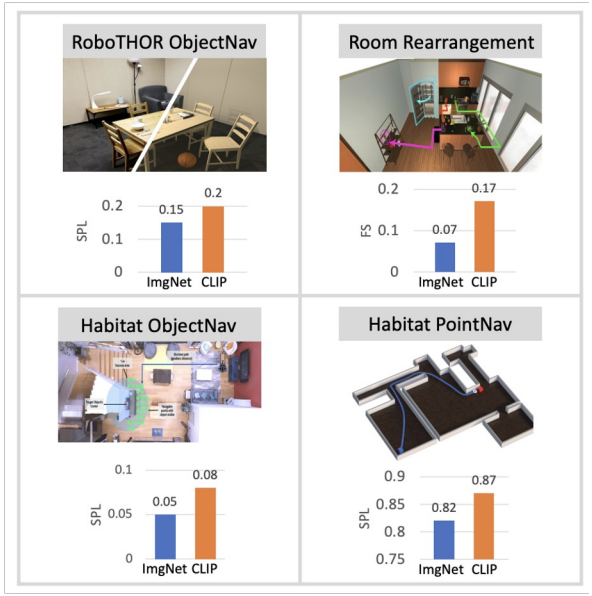
まとめBY: Fumiya Matsuzawa

マルチモーダル*LLM*関連

Simple but Effective: CLIP Embeddings for Embodied AI (Allen2AI, CVPR2022)

- 概要：**Embodied AIタスク**において、大規模言語・画像事前学習モデル（**CLIP**）の有用性の検証実験。
- 実験内容：既存のEmbodied AIタスクの画像特徴BackboneをResNetから、CLIPに変更した実験を行った（左図）。この際にRGB画像のみ使用し、CLIPのFinetuneをせず、タスクごとのAuxiliary Lossなども追加しない、**Simpleなモデル**を使用。
- 結果：4つのEmbodied AIタスクでSOTAを達成（右図）





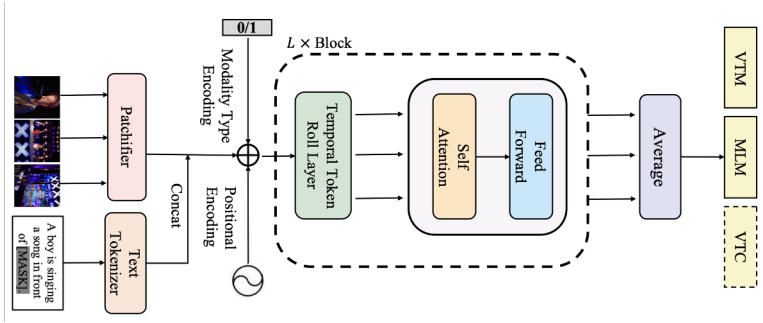
Simple but Effectiveで、Embodied AIタスクのBaselineとして活用できる！

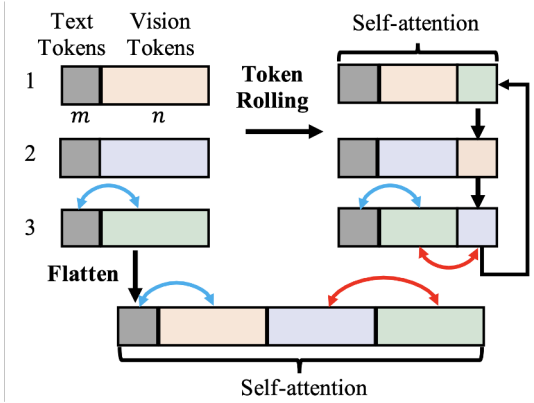
まとめBY: Qiu Yue

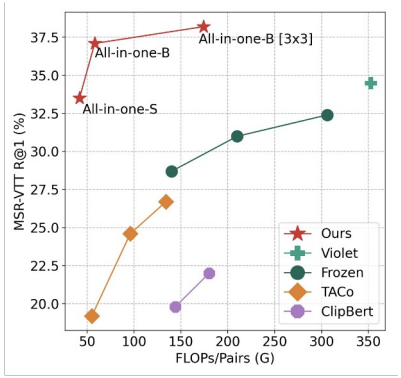
All in one: Exploring Unified Video-Language Pre-training (CVPR 2023)

- 概要：既存研究では、videoとlanguageをそれぞれ別々のencoderで特徴を抽出してから、videoとtextのfusionを行う。提案手法では、videoとtextの特有のencoderを用いずに、videoとlanguageの関係性を学習するAll in one Transformer構造の提案（左図）。
- 手法：non-parametric的なtoken rolling operation（中図）を提案。テキストとsingle imageのトークン間の関係性を学習。１つのビデオから複数のimagesをサンプリングし、imagesのトークンのrollingを行う（毎回一枚のimage内の一部のトークンを次のimageにシャッフルする）。
- 結果：提案のTransformer構造で学習したvideoとlanguageの特徴表現がいくつかのvideoとlanguageのタスクで最も高い精度を実現した。さらに、精度が高い他、計算のコストも削減できた（右図）。

Simple and effective and lightweight





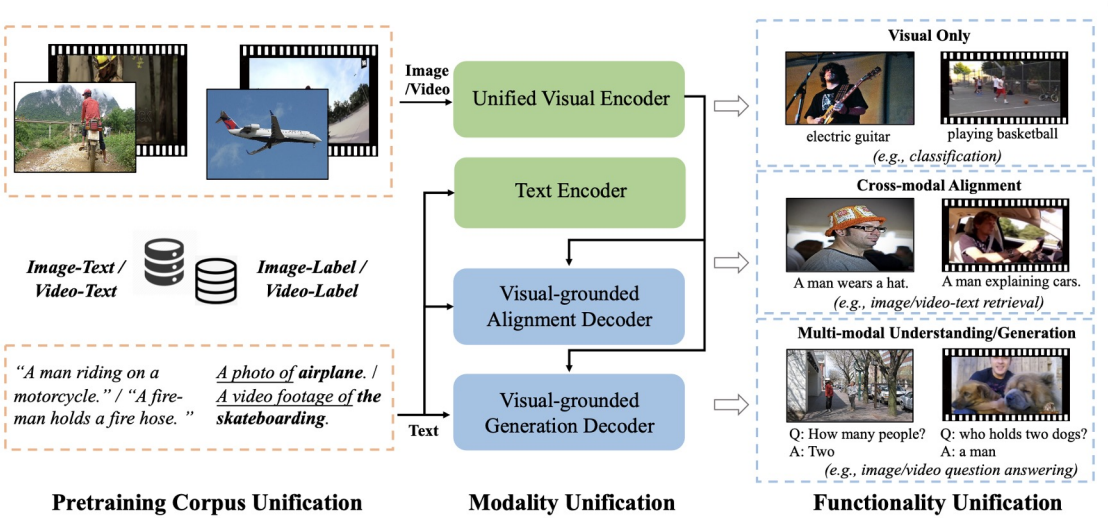


まとめBY: Qiu Yue

OmniVL: One Foundation Model for Image-Language and Video-Language Tasks (NeurIPS 2022)

- 概要：既存研究では、imageとlanguageもしくはvideoとlanguageの関係性を学習。この論文では、１つのモデルで複数のImage-languageとvideo-languageタスクを同時に行えるfoundation modelの提案（左図、右図）。
- 手法：decoupled joint pre-trainingを提案し、まずimage language pretrainingでspatial情報を得る。次に、image+video languageの学習を行い、temporal特徴表現を得る。また、既存のunified image language contrastive lossをimage+video+languageh絵拡張。
- 結果：同時に複数のタスクが行える(image/video recognition, image/video-text retrieval, image/video question answering, captioning）。また、上記のタスクにおいて、SOTAもしくは高い精度を達成。

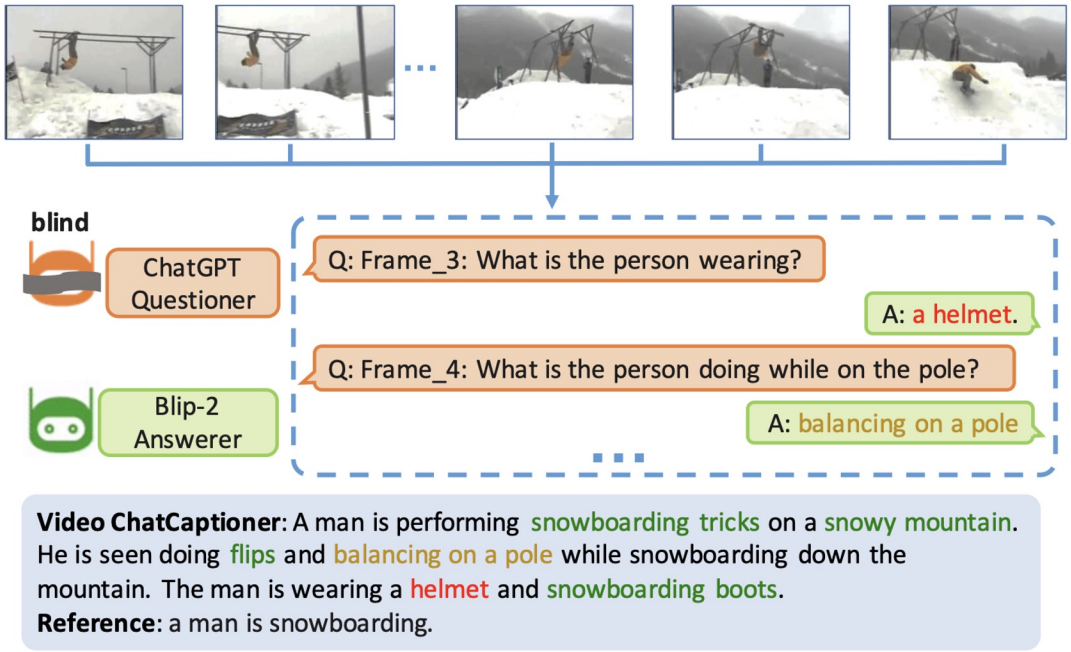
Method	Modality Unification		Functionality Unification		Data Unification			
	ILP	VLP	Non-Gen	Gen	I-T	I-L	V-T	V-L
CLIP [53]	✓		✓		✓			
ALIGN [30]	✓		✓		✓			
VLMO [62]	✓		✓		✓			
ALBEF [38]	✓		✓		✓			
SIMVLM [63]	✓		✓	✓	✓			
UniVLP [75]	✓		✓	✓	✓			
BLIP [37]	✓		✓	✓	✓			
FiT [6]		✓	✓		✓		✓	
ALPRO [36]		✓	✓		✓		✓	
VIOLET [23]		✓	✓		✓		✓	
FLAVA [55]	✓		✓		✓			
Florence [71]	✓		✓		✓	✓		
OmniVL (Ours)	✓	✓	✓	✓	✓	✓	✓	✓



まとめBY: Qiu Yue

Video ChatCaptioner: Towards Enriched Spatiotemporal Descriptions (arXiv 2023)

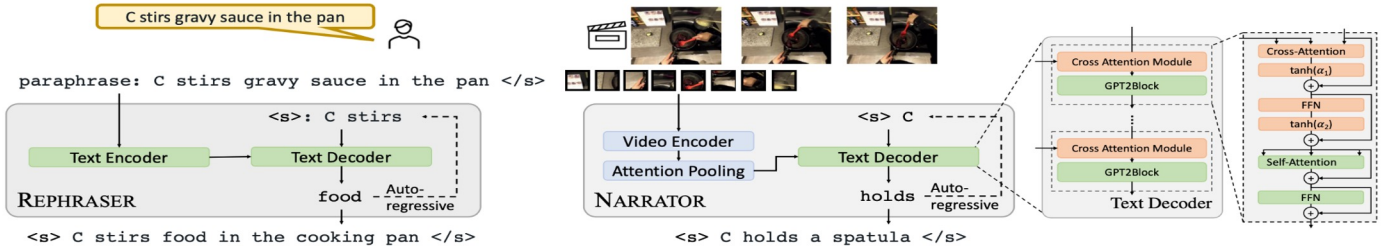
- 概要：LLMとマルチモーダルLLMを利用した、ビデオ内容を詳細的に説明する文章・対話を生成する手法の提案。
- 手法：提案手法は、ChatGPTとBLIP2を利用する。ChatGPTによりフレーム内容を理解するための質問を生成し、BLIP2により回答をする。最後にChatGPTによりビデオ内容の詳細説明文を生成する。
- 結果：提案のモデルを用いて、有効的にビデオ内容を詳細説明する文章を生成できる。全体的には、アプリケーションよりだが、ビデオ内容の認識を中心に有用な説明データ生成に活用できる（少し汚いデータかもしれませんが）。



まとめBY: Qiu Yue

Learning Video Representations from Large Language Models (META CVPR2023)

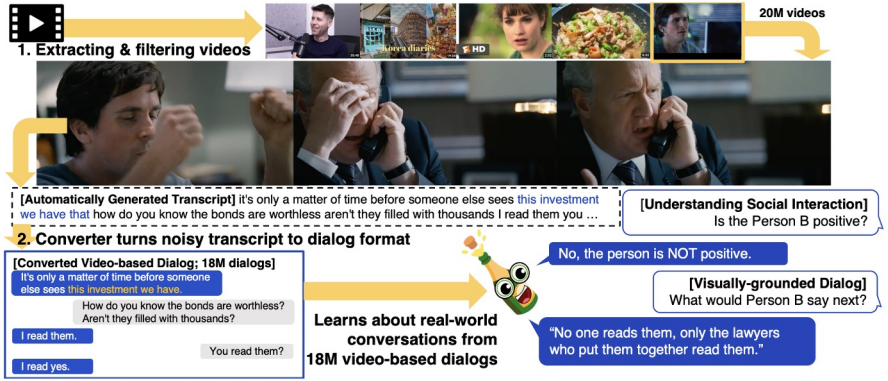
- 概要：LLMを使用し、videoとlanguageのrepresentation学習用のdata augmentationを行う手法を提案。大規模言語corpusと比べ、videoとlanguageのペアが比較的に少ない傾向になる。ここでvideo特徴をLLMに入力しnarrationsを生成する、データ拡張の手法を提案。
- 手法（下図）：論文では２種類のデータ拡張を提案。**Narrator**ではvideoをTimeSformerにより特徴抽出し、cross-attention moduleを挟んで、GPT2（Freeze）によりNarrationsを生成。cross-attention moduleは学習される。学習済みのNarratorにvideo clipsを入力することで新しいvideoとnarrationsのペアが得られる。**Rephraser**は既存のLLMモデルT5-largeを使用し、学習データの中のテキスト情報を別のセンテンスで言い換えることによりデータを拡張。
- 結果：提案手法により生成したnarrationsを学習することで、いくつかのvideoとlanguageのベンチマークで既存の再高精度より大幅に性能向上ができた。また、同レベル程度の精度を達成するため必要な人工データのサイズも大幅に下げた。
- 感想：精度が良かったが、手法的な新規性がほぼない？



まとめBY: Qiu Yue

CHAMPAGNE: Learning Real-world Conversation from Large-Scale Web Videos (arXiv2023)

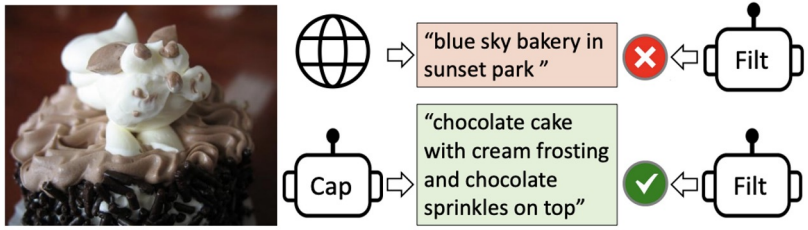
- 概要：既存の対話生成の研究はテキストベースのものが多い。しかし、実世界のconverstationでは視覚情報（例：social interactions, 表情・姿勢・行動などなど）の内容も重要になる。この論文では**視覚コンテキストと対話をトータルで理解すること**を検討。
- データセット：視覚コンテキスト（video）とそれと関連する対話を理解するために、新たなvideo+conversationnのデータセットYTD-18Mを提案（下は概要図）。YTD-18Mはこれまで**最大のreal-world conversation videoデータセット**となる。YTD-18MはYouTube動画を収集し、その中のnoisy transcriptsをLLMによりクリーンな対話に直して作成されている。
- 手法：ビデオ・過去の対話内容から、最後の対話に関してのresponseを生成するモデルCHAMPAGNEを提案（Unified-IO手法をベースとしている）。
- 結果：提案データセットで事前学習したCHAMPAGNEモデルが4つのreal-world conversation ベンチマークにおいて最も高い精度を達成。

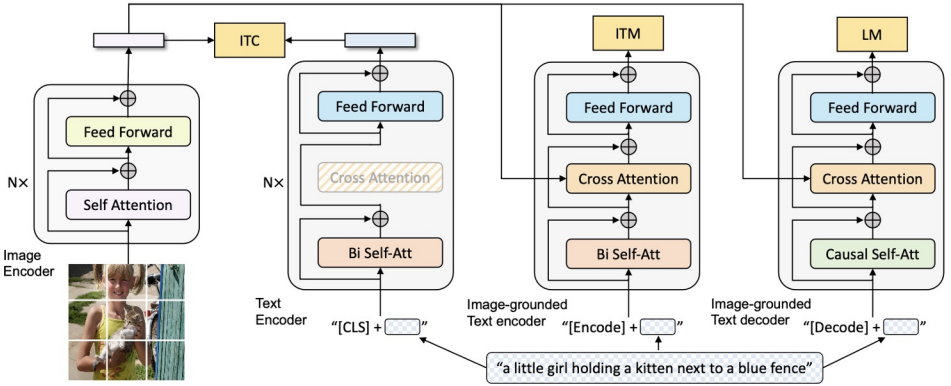


まとめBY: Qiu Yue

BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (ICML 2022)

- 概要：**認識と生成両方行えるvision-language**モデルBLIPを提案。これまでのvision-languageモデルは、認識と生成の片方のみ対応するモデルが多かった。また、既存の大規模事前学習モデルはノイズが含まれるウェブデータを使用し、データの質が結果を影響することがある。
- 手法：まず、パラメータ共有する**3つのモジュールを提案(右下図):1.テキストのencoding, 2.画像特徴を考慮したテキストのencoding / 3.decoding。**またロストしては画像・テキストのcontrastive loss, matching loss, 及び言語生成ロスでモデルを学習する。次に、CaptioningとFiltering（左下図）の2つのモジュールによりデータを拡張・クレーンジングする手法を提案。学習データの質を向上。
- 結果：論文の時点では、image-text retrieval、image captioning、VQAタスクのベンチマークで最も高い精度を達成。video-languageタスクへゼロショートの効果もあった。

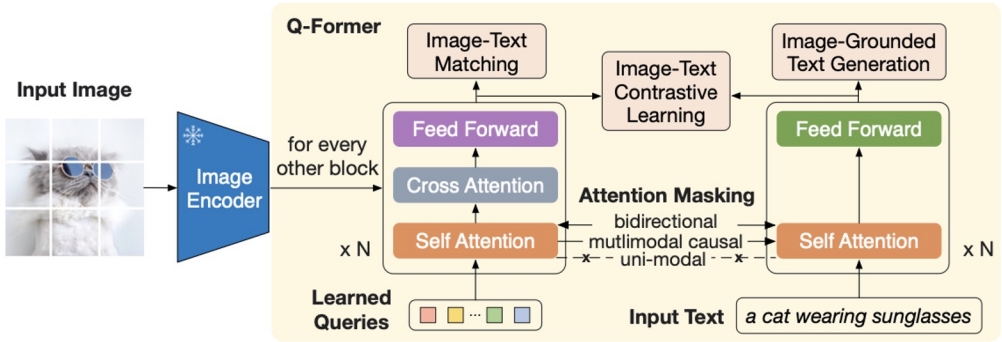


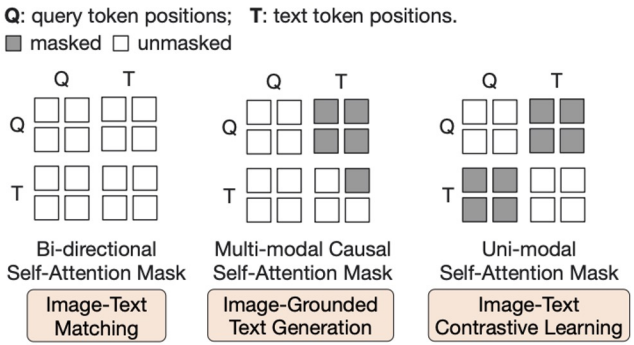


まとめBY: Qiu Yue

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

- 概要：**学習済みのvisual encoderとLLMをvision and language representationに有効的に活用**できる仕組みBLIP2（下図）を提案。
- 手法：BLIP2が学習済みのvisual encoderとLLM（両方とも再学習しない）及び、比較的に軽量的なquerying transformerから構成される。querying transformerの事前学習が２ステップに分ける。ステップ１では画像特徴を相関のテキスト特徴量と近づけるようにする。ステップ２ではLLMに解釈可能な画像特徴量の学習をする。
- 結果：既存のvision and languageモデルと比べて比較的にパラメータ数が少ない一方、いくつかのタスクにおいて最も高い精度を達成。更に、BLIP2構造は様々な事前学習visual encoderとLLM（encoder-decoder構造、decoder-only構造両方適応可能）に適応できる。BLIP2がzero-shot image-to-text生成も高い性能を示した。



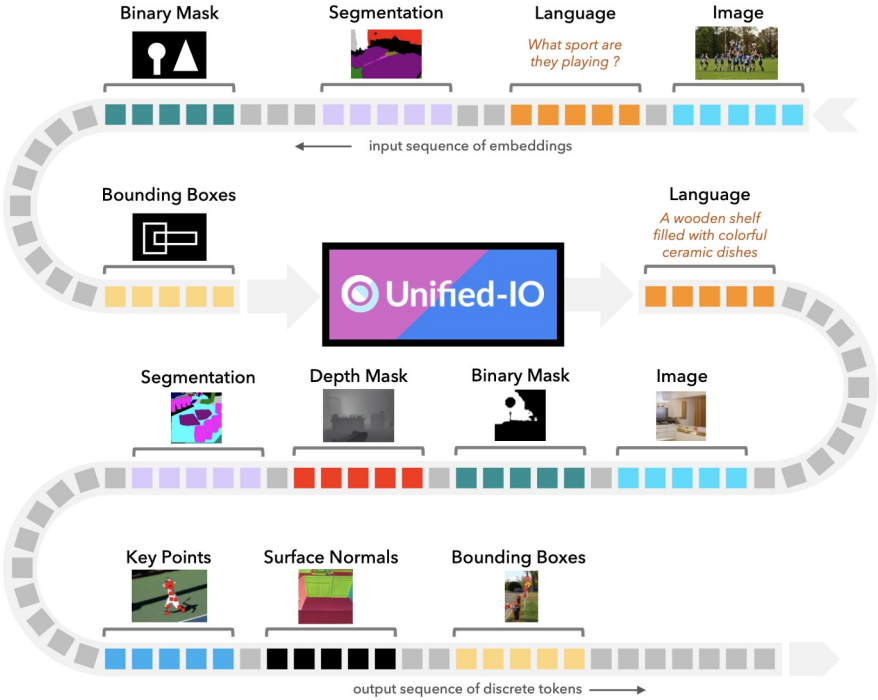


まとめBY: Qiu Yue

UNIFIED-IO: A Unified Model for Vision, Language, and Multi-modal Tasks (ICLR 2023, Allen2AI)

- 概要：**様々なimage/text recognition/generationタスクを１つに統一した構造**Unified IO(右)を提案。
- 新規性：認識のみではなく、**画像生成、デプスマップ推定、物体検出などのタスクも対応可能。**また、**タスク特化した画像encoderやdecoderが必要なし。**

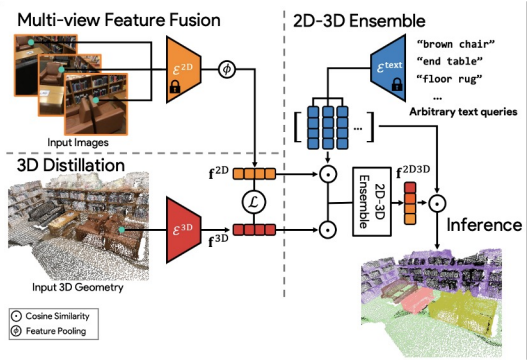
- 手法：Unified IOがTransformer Encoder-Decoderの構造となる。入力する際に、タスクごとに”Predict the depth map of the input image”のようなタスクPromptを一緒に入力。また、Unified IOに入力する前に、画像やテキストをパッチ化しLinear Projectionにより Encodeする。出力する際に、画像・デプスマップなどのようなdenseな出力をVQ GANのCode tokenにする。
- 結果：GRITというマルチタスクベンチマークに含まれるすべてのタスクに対応可能であり、平均点数が64.3となり、２番目の手法より30%以上の性能向上。また、タスクごとにFine-tuneをせずに様々なタスクで高い精度を達成。

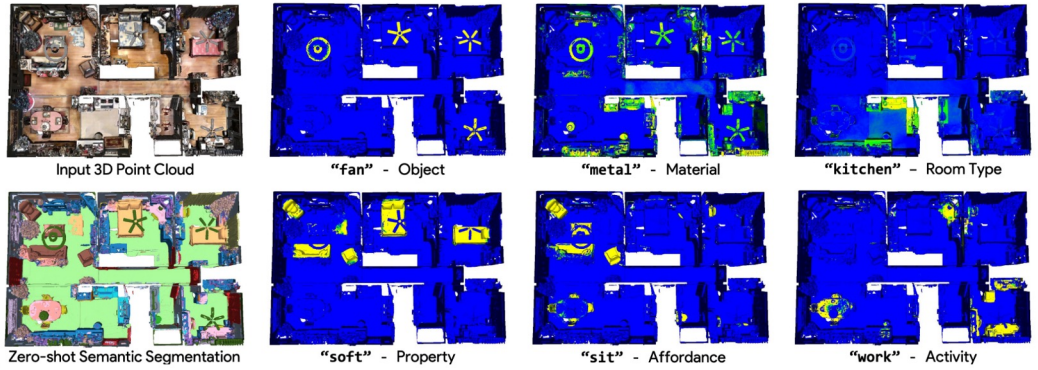


まとめBY: Qiu Yue

OpenScene: 3D Scene Understanding with Open Vocabularies (CVPR2023, Google)

- 概要：**CLIP特徴量を室内3次元環境のzero-shot認識**に適応した手法を提案。
- 新規性：この論文までには、CLIPを3次元環境のdenseな認識タスクへ適応したものがなかった。さらに、CLIPの特徴表現を適応することで、Open vacabularyのQueryを3次元環境において検出することが可能（右下図に結果例）。
- 手法：まず、3次元環境の点群とマルチ視点から撮影したシーンの画像の特徴量が類似となるように学習される。次に、3次元特徴と2次元特徴をアンサブリングし、アンサンブルされた特徴量とCLIPテキスト特徴量の類似度を計算することで、点群の点ごとのラベル推定ができる（与えられたqueryと類似すればqueryのラベルが当てはまる）。（左下図）
- 結果：**Zero-shotで**複数の既存の3次元タスクで**Full supervised 学習した手法より高い精度**を達成。さらに、提案のOpenScene手法はOpen vacabularyの3次元シーン認識ができる。

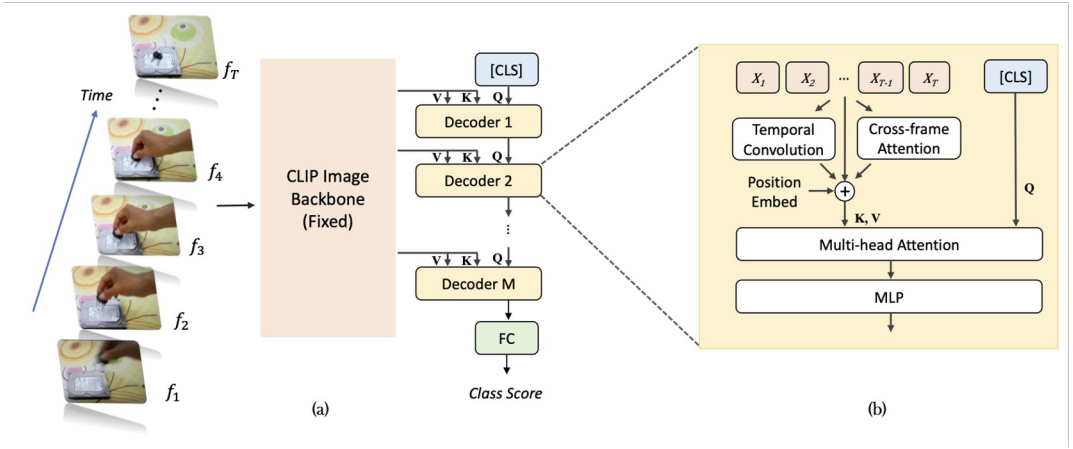




まとめBY: Qiu Yue

Frozen CLIP Models are Efficient Video Learners (ECCV2022)

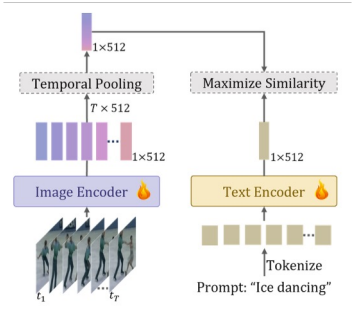
- 概要：**CLIP画像特徴量（Frozen）を動画認識に適応した。**
- 新規性：CLIPのような画像とテキストの特徴量を動画像認識に適応したものがこれまでに少なかった。また、既存の手法では画像特徴量を動画像認識に適応する際にFinetuneするものが多い。この論文ではCLIP特徴量Finetuneしない設定をしている。
- 手法：ビデオの画像フレームをCLIPモデルに入力し、Frameごとの特徴量を得る（Frozen）。またCLIP特徴量の上にTransformer Decoderを導入し、画像の間の時系列情報を学習。（下図）
- 結果：まず、CLIPの重みが学習時に更新しない（Frozen）のため、提案手法EVLの計算コストが比較的に小さい。更に、Kinetics400、Something-Something-v2などのデータセットで高い精度を達成。

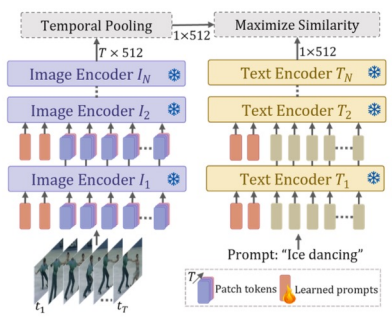


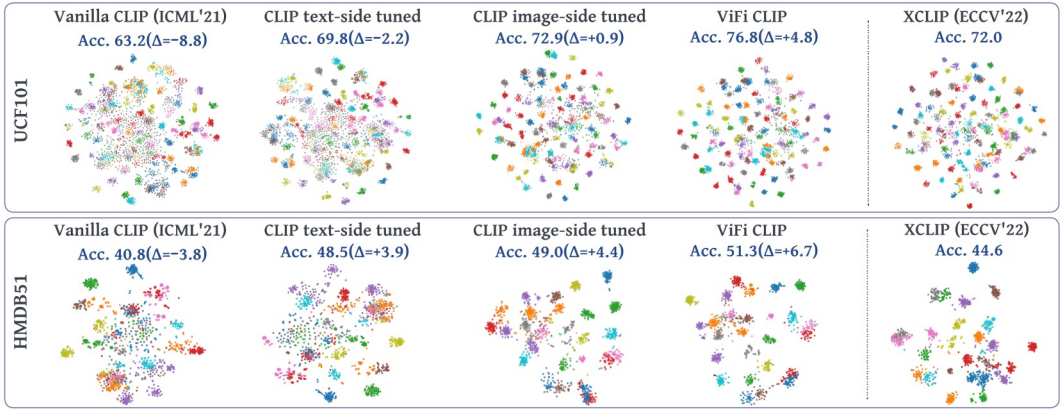
まとめBY: Qiu Yue

Fine-tuned CLIP Models are Efficient Video Learners (CVPR2023)

- 概要：**CLIP画像特徴量（Fine-tune）を動画認識に適応した。**
- 新規性：既存手法はCLIP特徴量を動画認識に適応する際に、CLIPの画像特徴抽出の部分をfreezeする手法が多かった。ここでは、**frozen CLIP特徴量の汎化性能が低下**することを主張し、Fine-tune CLIP特徴量を動画像認識に適応した。
- 手法：左下図に示すように、提案手法がCLIPの画像とテキスト両方を動画像データセットでfine-tuneして動画像とテキストの関係性を学習する。そして、各タスクへ転移学習する際に、Promptsを利用して（真ん中図）、タスクに適応した。
- 結果：右下図で示すように、既存の手法と比べて、同画像人式ベンチマークUCF101やHMDB51などにおいて最も高い精度を得られた。更に、学習した特徴量を可視化した結果、提案手法がうまくクラスタリングできることがわかった。



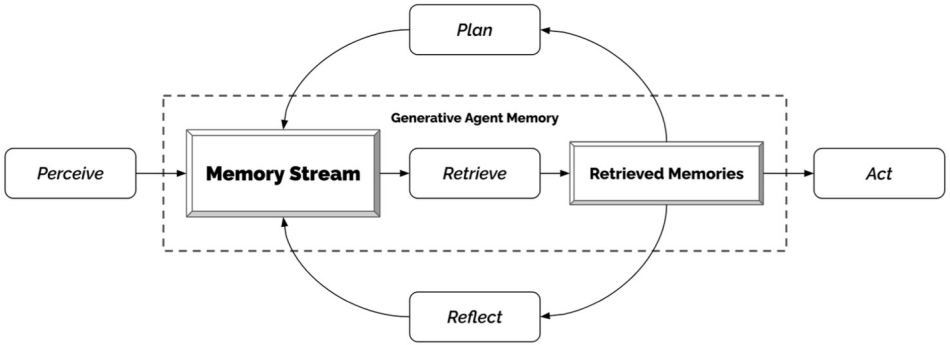




まとめBY: Qiu Yue

Generative Agents: Interactive Simulacra of Human Behavior (Google Research, Stanford University 2023)

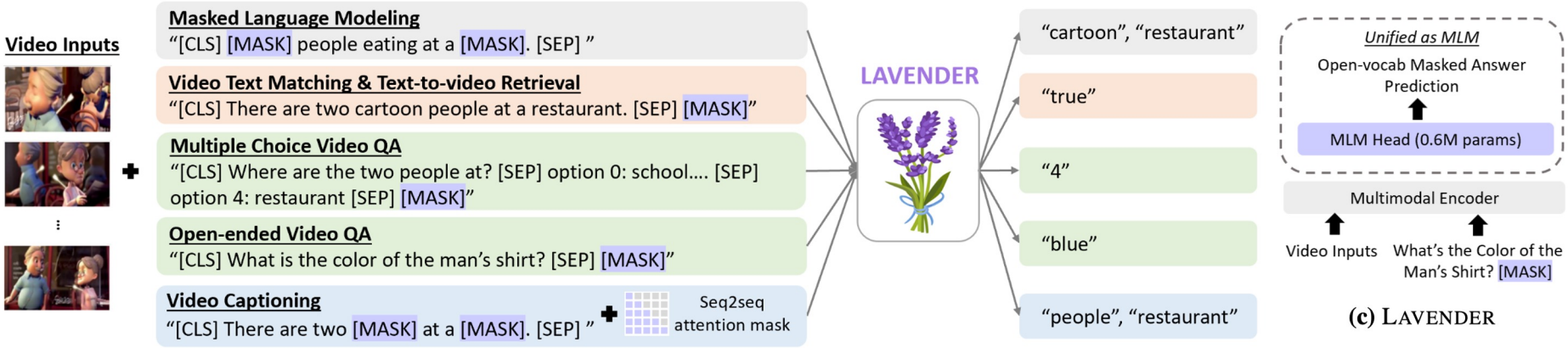
- 概要：大規模言語モデル**LLM**を利用しつつ、**Human social activitiesをシミュレーション**するフレームワークGenerative Agentsを提案。**LLM /ChatGPTを利用してゲームのNPCを自動生成**するみたいな感じ（左図）。
- 新規性：これまでの研究では短いタイムスパンでのHuman activitiesをシミュレーションできたが、Group human activitiesを長いタイムスパン（二日）でシミュレーションする研究はこれまでになかった。また、LLMをHuman activitiesのシミュレーションへ適応する研究もこれまでになかった。
- 手法：提案手法が３つのモジュールから構成される。まず、memory streamは自然言語でAgentsの行動リストを記述する（LLMベースのRetrievalモデル）。次に、reflectionでは行動リストを高レベルで認識し、重要度などにより行動リストを認識・まとめる。最後に、planningではReflectしたMemoryから、行動の計画を行う。また、選択された行動によりmemoryを更新していく。（右図）
- 結果：実験評価により、提案のGenerative Agents（25Agents）が有効的にリアルなIndividualとGroup Social Activitiesをシミュレーションできることを示した。例えば、25Agentの中の一人のAgentに”バレンターンディー”のイベントをあげたい”のみを伝えて、それに沿って”イベントをあげる”情報が全てのAgentsに流れたり、各々のAgentが準備したりするリアルなactivitiesのシミュレーションに成功。



まとめBY: Qiu Yue

LAVENDER: Unifying Video-Language Understanding as Masked Language Modeling (Microsoft, CoRR 2022)

- 概要：**あらゆるVideo＋Language→Languageモデルを Masked LMに統一した**モデル提案。
- 既存手法：まずタスクspecific headやlossが必要な手法がある。また、複数タスクをseq2seqなencoder-decoderへ統一する手法もある（decoder側パラメータ数が多い）。手法：提案手法ではMasked LMの形式でVideo＋Languageタスクを統一する。また、事前学習（MLMタスクとVideo Text Matchingタスクで事前学習）・finetune時は同じ構造を使う。このようにspecific headやlossの設計が必要ない、かつ、encoder-decoder構造よりパラメータ数が少ない。
- 結果：提案のmasked LMベースの手法は比較的にモデルが軽量的であり、14種類のVideo ＋Language ベンチマークデータセットにおいて高い精度を達成（うち12種類はSOTA）。更にfew-shotやzero-shot タスクにおいても活用できる。

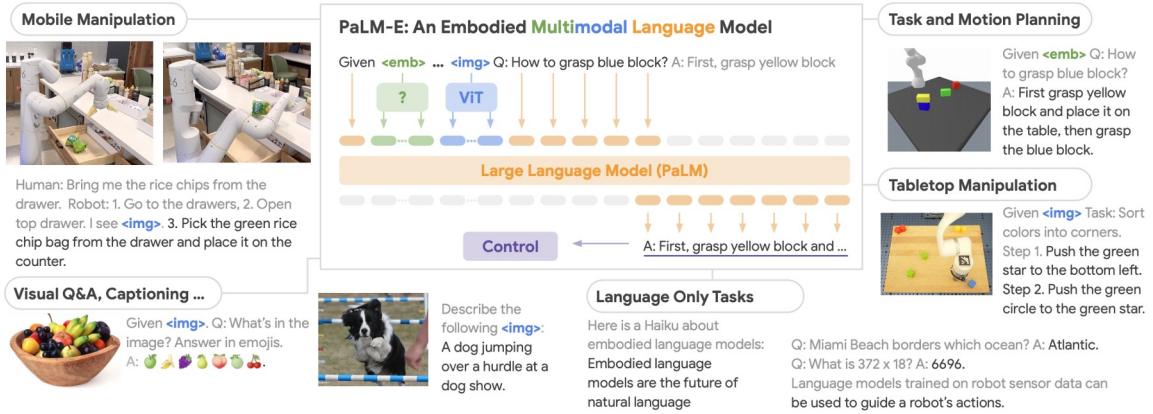


まとめBY: Qiu Yue

PaLM-E: An Embodied Multimodal Language Model (Google, 2023)

- 概要：**Vision and Languageタスク、Language-onlyタスク、Embodied AIタスク、実際のロボットマニピュレーションができる、embodied languageモデル**を提案。
- 既存研究の問題点：既存のVandL研究が大規模ベンチマークで精度が良いが、実世界ロボットに応用するレベルに達していない。また、実際にロボットの動的な視覚観測・物理センサーなどのデータを扱える大規模LLM、Multi-modalがほとんどなかった。
- 提案の**PaLM-E（562B）**：PaLMをベースとして、画像・物理センサーのStateなどをTransformer encoder-decoderの形式で言語生成し（これを用いてロボットの操縦が可能）、画像などの情報をLLMの言語と同じembeddingにアラインさせる。
- 結果：既存のVandL、Embodied AIタスクのZero-shot 性能が高い一方、OCR-free math reasoning, multi-image reasoning, そして
  Multi-modal chain-of-thought 能力が出現（emergent ability）。

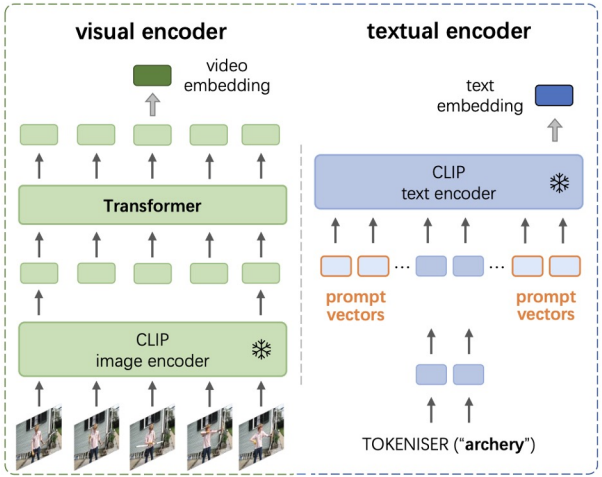
実世界ロボットと繋ぐMulti-Modal Embodied LLM

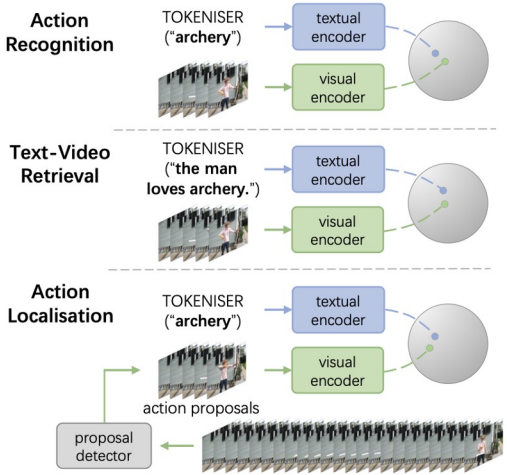


まとめBY: Qiu Yue

Prompting Visual-Language Models for Efficient Video Understanding (ECCV, 2022)

- 概要：**CLIPを動画認識・Video＋Languageタスクに適応した**手法。
- 手法：CLIPのテキスト側にPrompt vectorsを追加し、タスクごとの特徴量を学習させる。また、CLIP学習済みの画像特徴抽出器にTransformer構造を追加しフレーム間のTemporal関係を学習する。学習する際に**prompt vectorsとtemporal transformer**のみ学習。（下図）
- 精度：既存手法より少ないパラメータ数で10つ既存の動画像ベンチマーク（動画タスクとVideo＋Languageタスク）で高い精度を得られた。

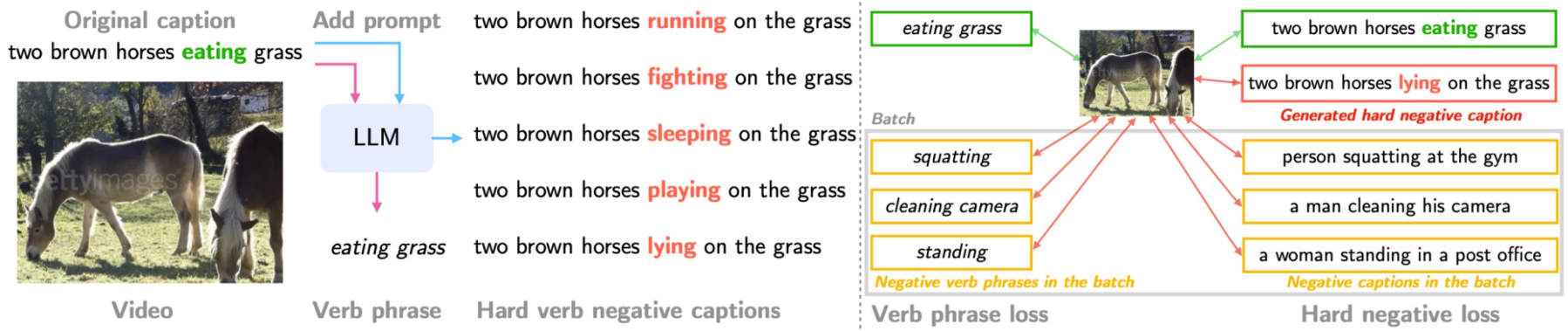




まとめBY: Qiu Yue

Verbs in Action: Improving verb understanding in video-language models (Google, 2023)

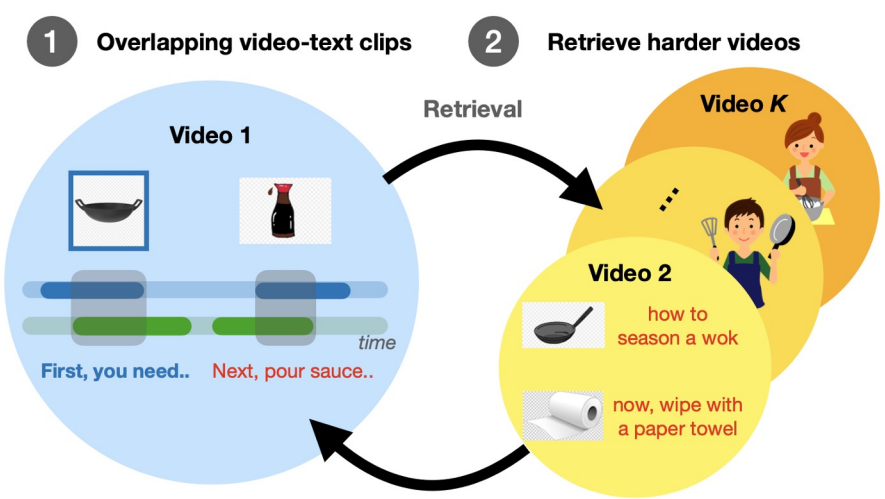
- 概要：**動詞にフォーカスした、videoとtextのcontrastive learning手法の提案。**
- 新規性：画像とTextのCLIPモデルを動画像に適応した手法が多い。この文章では、動詞に着目し、動画とテキストのcontrastive Learning手法を提案。
- 手法：まず動画像と動画内容のcaptionから、LLMを使用し、captionの動詞を異なるものに変更するhard negative captionsを用意。また、captionをまとめたverb phraseを生成。そして、２種類のcontrastiveロス（verb phrase lossとhard negative caption loss）を用いて動画像とテキストのペア関係を学習する。
- 精度：３つの下流タスク（video-text matching, video QA, 動画像認識）において最も高いゼロショート精度を達成。



まとめBY: Qiu Yue

VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding (Meta, EMNLP2021)

- **概要：videoとtextのcontrastive learning手法**の提案。下流タスクで**高いzero-shot精度**が得られた。
- **学習データ：**まず、pre-training時に既存のベンチマークHowTo100Mを使用。実験で、HowTo100Mをダイレクトで使用するとzero-shot精度が比較的に高くないと示しながら、もっと**fine-grained的なvideoとtextのassociationsが必要**と主張。
- **モデル：**まず、temporally overlapped的なvideo-text aignmentを使用し、異なる長さのセンテンスと対応するvideoの対応を可能にした。contrastive learning時に、既存の手法では同じ動画内でhard negative examplesを集めることが多い。ここでは、retrievalベースの手法でvideoをクラスタリングし、学習時に同じクラスターの中の類似した他のビデオからhard negative examplesを集める。（右図）
- **精度：**複数の下流video-languageタスクで最も高いzero-shot精度を達成。また、既存のsupervised学習の精度を超えたケースもある。

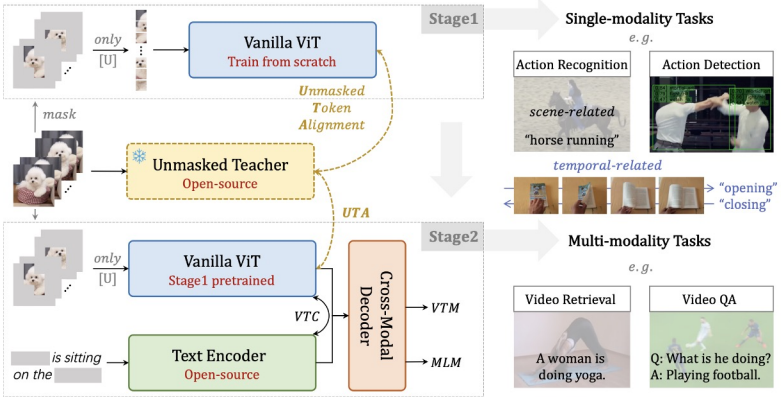


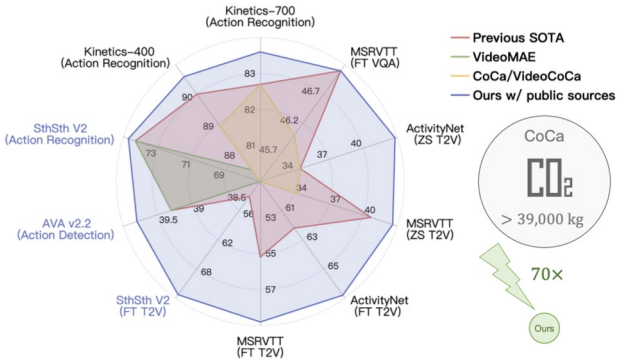
VideoCLIP: Contrastive learning with hard-retrieved negatives and overlapping positives for video-text pre-training.

まとめBY: Qiu Yue

Unmasked Teacher: Towards Training-Efficient Video Foundation Models (arXiv 2023)

- 概要：**画像のfoundation modelを教師とした効率的なvideo foundationモデルの提案。**
- 既存手法：既存のvideo foundationモデルはimage foundation model（IFM）をベースとしたものが多く、temporal情報認識が劣る。また、videoMAE手法がtemporal情報を習得できる一方、計算コストが高く、高レベルのセマンティック情報が必要となるvideo+languageタスクにおいての性能が高くない。
- 提案手法：提案手法はvideoMAEとIFMのCLIPを利用した。具体的に、CLIPの出力を教師としてvideoMAEモデルを学習させる。その際にCLIPの高レベルセマンティック情報を利用できるつつ、オリジナルのvideoMAEの重いdecoderの部分を避ける。また２段階で、videoMAEの部分とvideo+languageの部分をそれぞれ学習することで、学習効率を高めながら低レベルから高レベルなセマンティック情報の習得が可能になる。（左図）
- 精度：**32枚のA100 GPUで６日間で学習した**提案手法が複数のvideo-only, video+languageタスクで**最も高い精度**を実現。（右図）

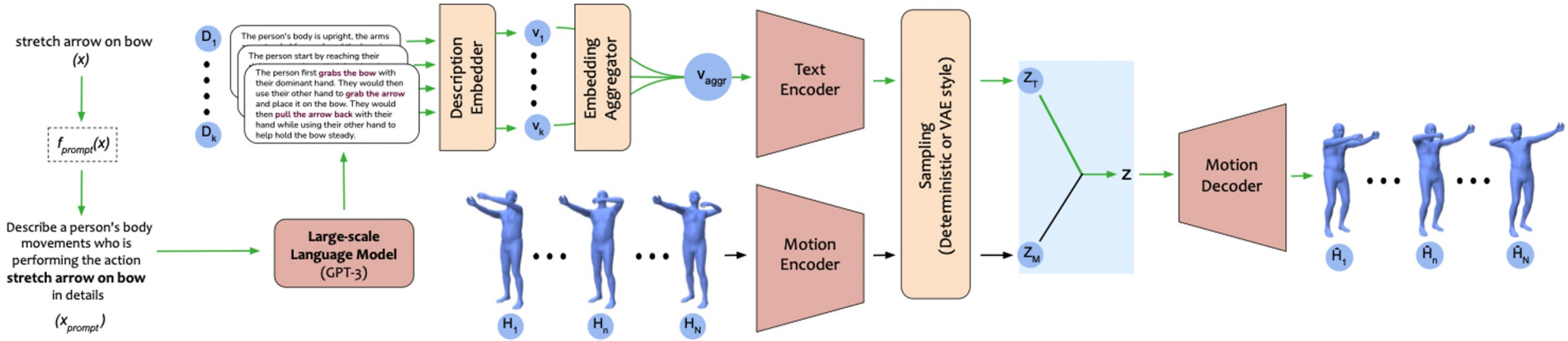




まとめBY: Qiu Yue

Action-GPT: Leveraging Large-scale Language Models for Improved and Generalized Action Generation

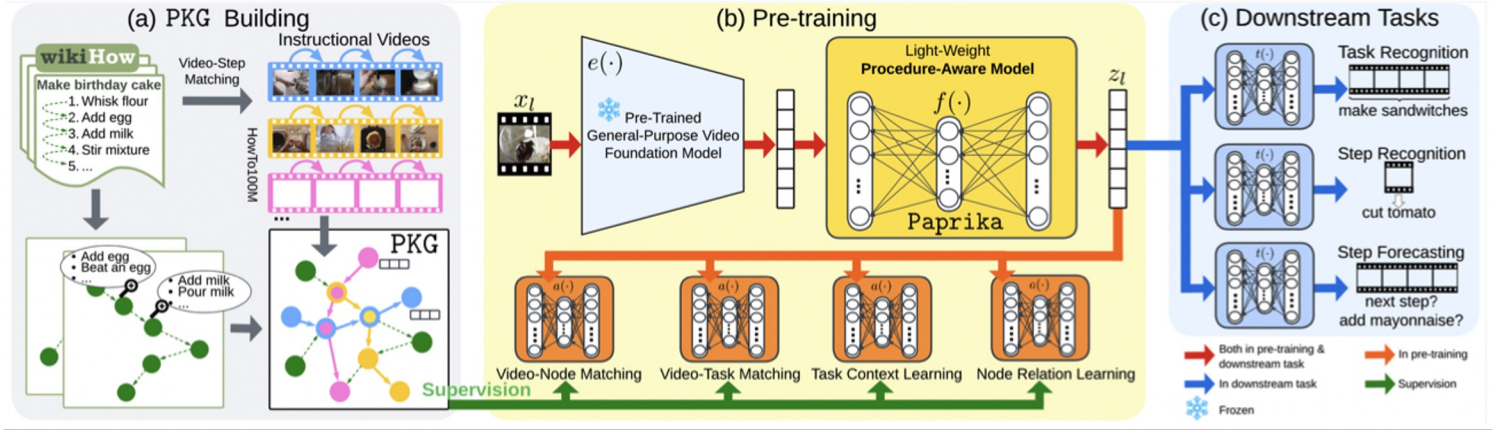
- 概要：既存のデータセットの行動の記述をプロンプトとしてLLMを用いて詳細な記述に変換することで行動モーション生成を品質を改善.
- 新規性：LLMのテキストからのモーション生成への応用，適切なプロンプトを生成する関数の設計.
- 結果：従来の行動記述を用いた学習に対して，設計した関数によるプロンプトに基づいて生成したLLMによる記述を用いた学習により，高品質なモーションの生成ができることを実証.



まとめBY: Kensho Hara

Procedure-Aware Pretraining for Instructional Video Understanding

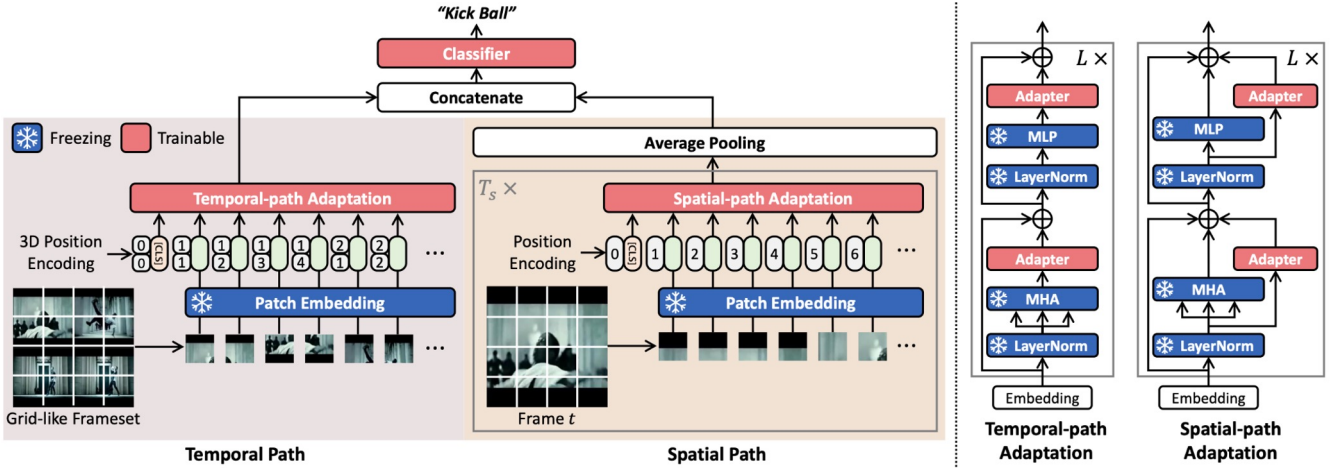
- 概要：インストラクション動画の理解に向けたPre-training手法の提案．wikiHowのデータから得た手順データとインストラクション動画から抽出した手順の遷移に基づいて作業手順の知識グラフ (PKG) を構築しPKGを教師としてモデルを学習．タスク認識，ステップ認識，ステップ予測の各タスクにおいて性能を向上．
- 新規性：PKGの構築とPKGによるPre-training手法の提案．
- 結果：PKGを利用した4つのPre-training Taskを全て用いる提案手法が，COIN, CrossTaskをDownstream Taskとした評価実験において最も高い性能を達成．



まとめBY: Kensho Hara

Dual-path Adaptation from Image to Video Transformers

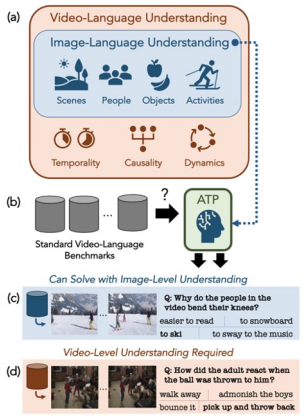
- 概要：画像で学習した基盤モデル (Transformer)を動画に効率的に適用するための手法を提案．従来手法とは異なり時間と空間に分割するDual-path構造を採用し，各pathでそれぞれ軽量なAdapterを挿入．Adapterのみを学習することで効率的な学習を実現.
- 新規性：Dual-path構造のAdaptation.
- 結果：4つのデータセットで従来のAdaptation手法を超える性能を達成.

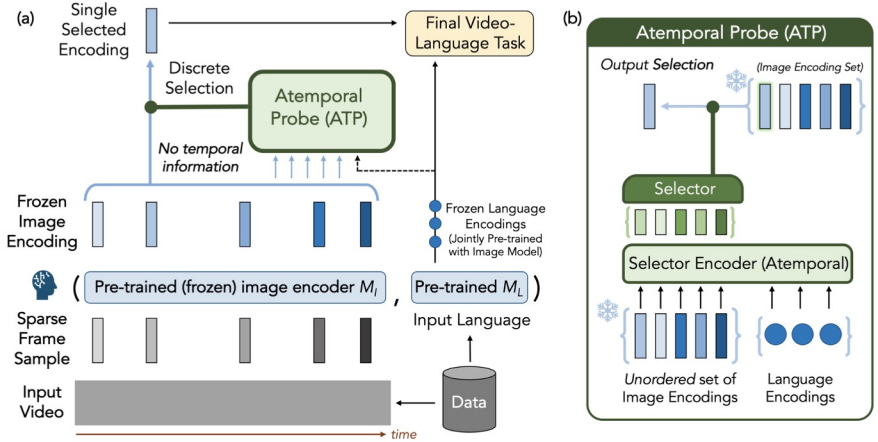


まとめBY: Kensho Hara

Revisiting the “Video” in Video-Language Understanding

- 概要・結果：動画-言語理解において「動画」の必要性を分析．画像のみを用いる強いベースラインとしてATPを提案．動画にはノイズとなるフレームも混ざっているため，複数フレームから抽出した特徴量（学習済みの画像-言語モデルを利用）の中で適切なものを選択し利用するためのモジュール．ATPにより静止画-言語モデルの表現だけでSOTAに匹敵・超える性能が確認できた他，limiationも分析．
- 新規性：ATPを提案し動画入力を最大限に利用した静止画特徴による性能を評価して「動画」の必要性を詳細に分析した点．

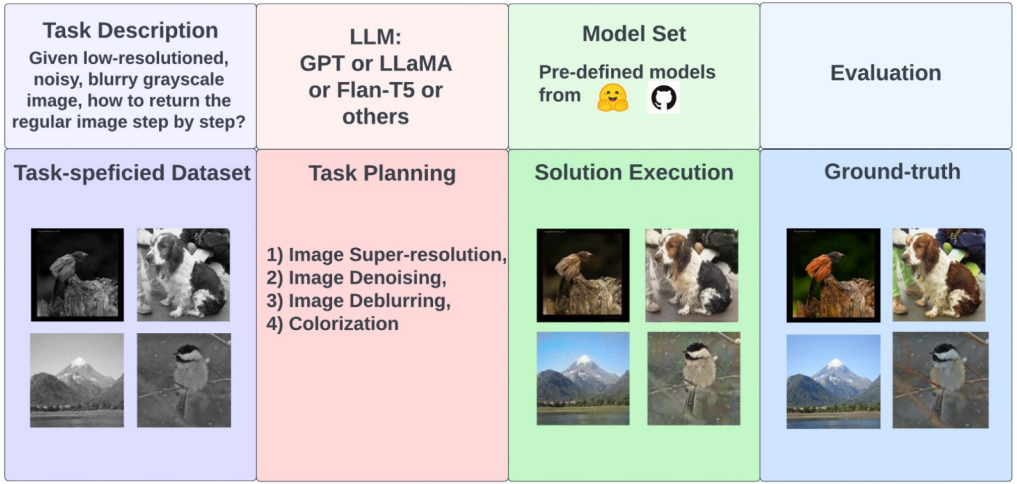




まとめBY: Kensho Hara

OpenAGI: When LLM Meets Domain Experts (arxiv)

- 概要：複数の複雑なタスクを解決するため，自然言語クエリを入力として，LLMが外部で提供するモデルを選択，合成，実行してタスクを行うOpenAGI（左図）を提案した.
- 新規性：Reinforcement Learning from Task Feedback(RLTF)メカニズムで，LLMがタスクの解決結果をフィードバックとして使用し，LLMの能力を向上させる.
- 結果：強化学習からのタスクフィードバック（RLTF）は、特定のタスクやドメインに対するLLMの適応能力を大幅に向上させることができる(右図)

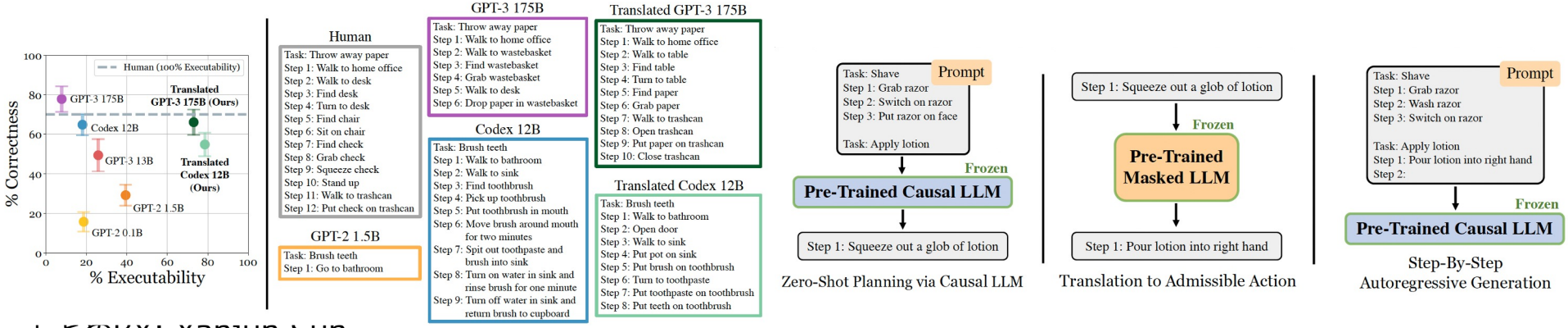


Task description	Input Sample	Output Sample
Given low-resolutioned noisy blurry grayscale image, how to return the regular image step by step?		
Given low-resolutioned noisy blurry grayscale image, how to return the object names in English step by step?		bear
Given clozed English text, how to translate the text in German step by step?	A big burly grizzly bear is show [Mask] grass in the background.	Ein kräftiger Grizzly Bär ist im Hintergrund mit Gras zu sehen.
Given noisy blurry grayscale image and clozed English query, how to answer the question in English step by step?	 **Question**: what number is [Mask] the player's jersey?	22
Given clozed English document and clozed English query, how to answer the question in German step by step?	**Context**: Super Bowl 5 was an American football game to determine the champion of the National... **Question**: What was the theme of Super [Mask] 50?	Goldener Jahrestag

まとめBY：Yanjun Sun

Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents (ICML)

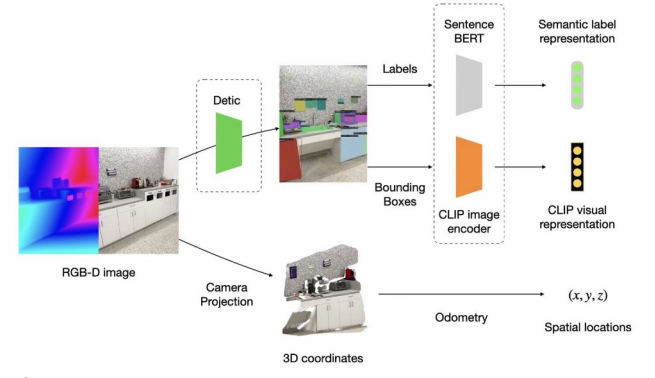
- 概要：既存のLLMはタスクに基づいて一連の行動計画を作成することが可能ですが、これらの行動の実行可能性（左図参照）が低い点に着目した。LLMの出力と環境が許容する行動をマップすることで、LLMがEmbodied環境において高い実行可能性を持つ行動計画を作成することが可能となりました。
- 新規性：モデルが低いレベルのタスクを実行可能にするため、モデルを変更せずに、一連のツール（右図参照）を提案した。
- 結果: Human Evaluationで提案したTranslated LLMは、生成されたアクションの実行可能性と精度を向上させました。

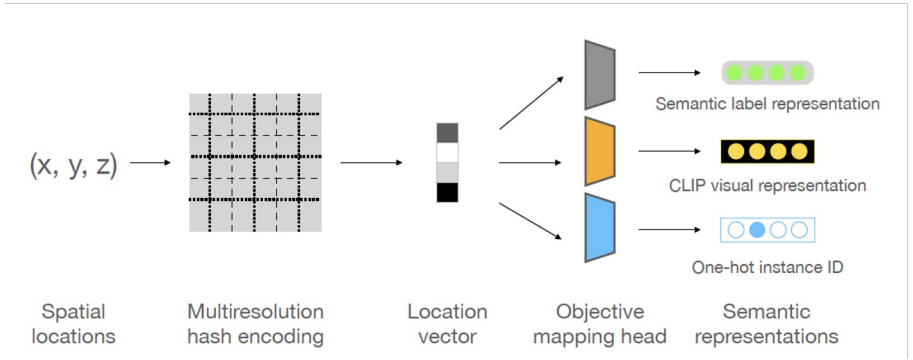


まとめBY：Yanjun Sun

CLIP-Fields: Weakly Supervised Semantic Fieldsfor Robotic Memory

- 概要：CLIPとNeRFを組み合わせることで、空間上のセマンティックな表現を得られて、いろんなロボティクスのタスクに応用できる。
- 新規性：
  - ○ 学習済みの大規模モデル（CLIP、Detic、SentenceBERT）から得たセマンティックな表現を使用するため、人間によるラベル付けなくても良い
  - ○ セマンティックな表現と空間的位置の対応を学習した、
- 結果：セグメンテーションとナビゲーションの実験で検証した。
  - インスタンスセグメンテーション、セグメンテーション両方とも向上した。
  - 抽象的な指示与えても、ナビゲーションも成功。

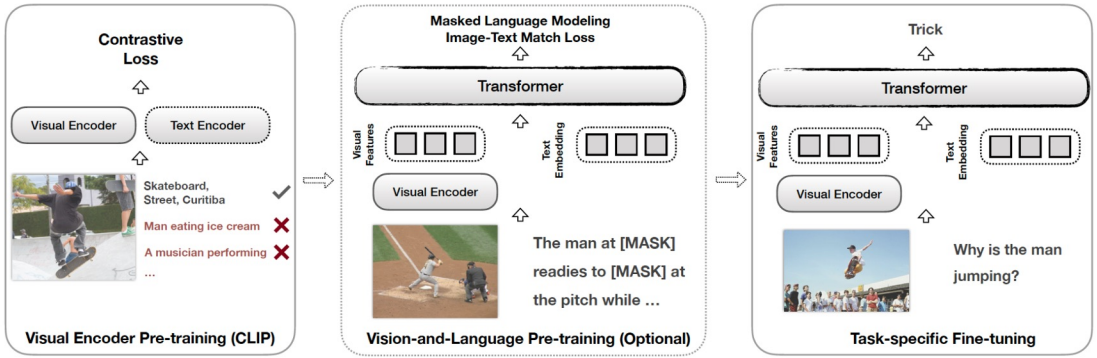




まとめBY: Yanjun Sun

How Much Can CLIP Benefit Vision-and-Language Tasks?(ICLR)

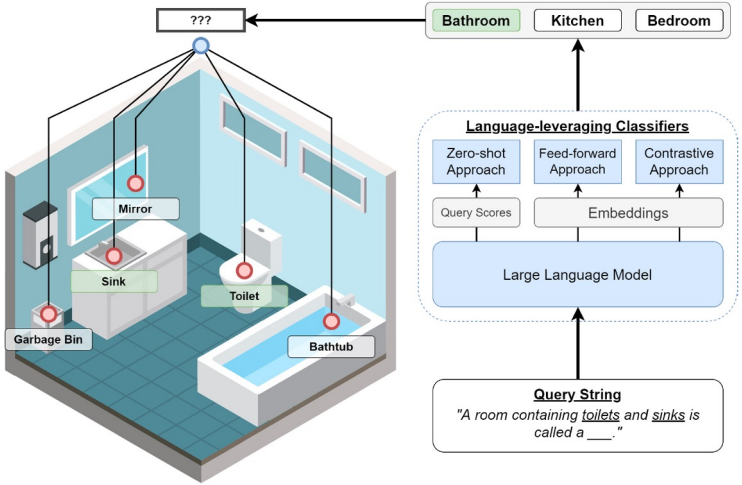
- 概要：V&LタスクにおけるボトルネックとされるVisual Encoderの性能を背景に、CLIPは膨大なデータで学習されたモデルとして優れたVisual Encoderを備えている。本研究では、CLIPがV&Lタスクにどれほど貢献できるのかを検証するために、既存のV&LタスクのVisual EncoderをCLIPのものに置き換え、多数の実験を通じてその性能を評価した。
- 実験：①CLIPのvisual encoderを既存のVQA、Image caption、VLNモデルに置き換える。②ペアのデータでもう一度事前学習を行ったあと、タスクに応じてfine-tuneする。
- 結果：①番目の手法では顕著な精度向上はない。それは、ViTが物体の位置検知精度がよくないからである。②番目の結果ではSoTAに達成。



まとめBY: Yanjun Sun

Leveraging Large Language Models for Robot 3D Scene Understanding(ICRA)

- 概要：ロボットが一般的な人間の家庭用品や場所に関する常識的な知識を持つことはまだ遠いため、LLMを使用してシーン理解に共通の感覚を与えることができる手法を提案した。具体的には、シーンから３Dシーングラフを構築し、３DシーングラフをLLMで部屋を分類することを行った。
- 新規性：LLMでシーンの理解を改善した。
- 結果：



まとめBY: Yanjun Sun

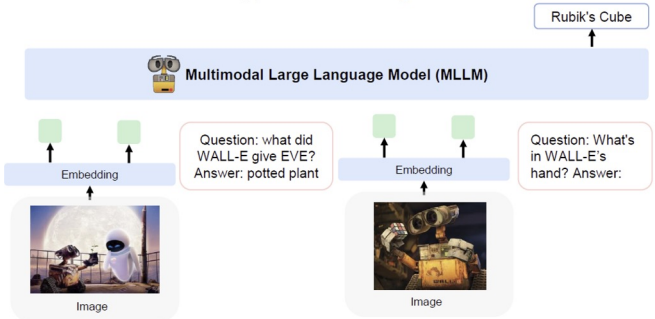
Language Is Not All You Need: Aligning Perception with Language Models

会議：ArXiv 2023　In-context Learningに画像や音声を自然に取り込める
著者：Microsoft　マルチモーダル大規模言語モデル KOSMOS-1を提案

- 概要：画像や音声などの情報を <image>埋め込み</image>といったトークンに囲まれた埋め込みベクトルとして表現し任意の場所でテキスト以外の情報を受け取れるように
- 新規性：画像-テキストペアだけでなくインターネット上の画像を含むテキストを訓練データとして学習、IQテストを含む様々な画像・言語タスクで性能を評価
- まず言語のみのLLMを訓練してからマルチモーダルデータでfinetune
- 訓練データ
  - テキスト：The PileやCCなど約2800億トークン
  - 画像-テキスト：LAION-2Bなど数十億画像
  - 不連続データ：CCサイトから取得した約20億の画像を含む英語ページ

Datasets	Format Examples
Text	<s> KOSMOS-1 can perceive multimodal input, learn in context, and generate output. </s>
Image-Caption	<s> <image> Image Embedding </image> WALL-E giving potted plant to EVE. </s>
Multimodal	<s> <image> Image Embedding </image> This is WALL-E. <image> Image Embedding </image> This is EVE. </s>

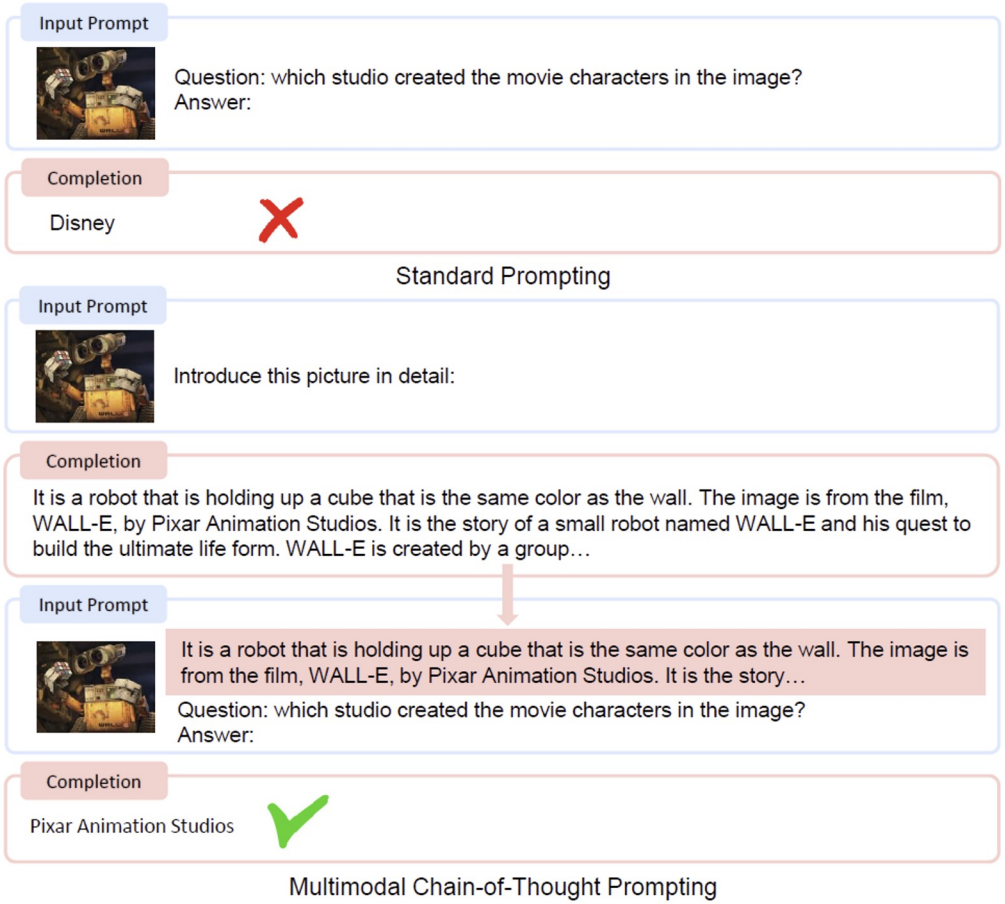
Table 21: The examples of the data format to train the KOSMOS-1 model.



LLMの枠組みをできる限り崩さないようにしながら
巨大データで訓練、コーパスのクレンジングが肝？

まとめBY: Takuma Yagi

Language Is Not All You Need: Aligning Perception with Language Models



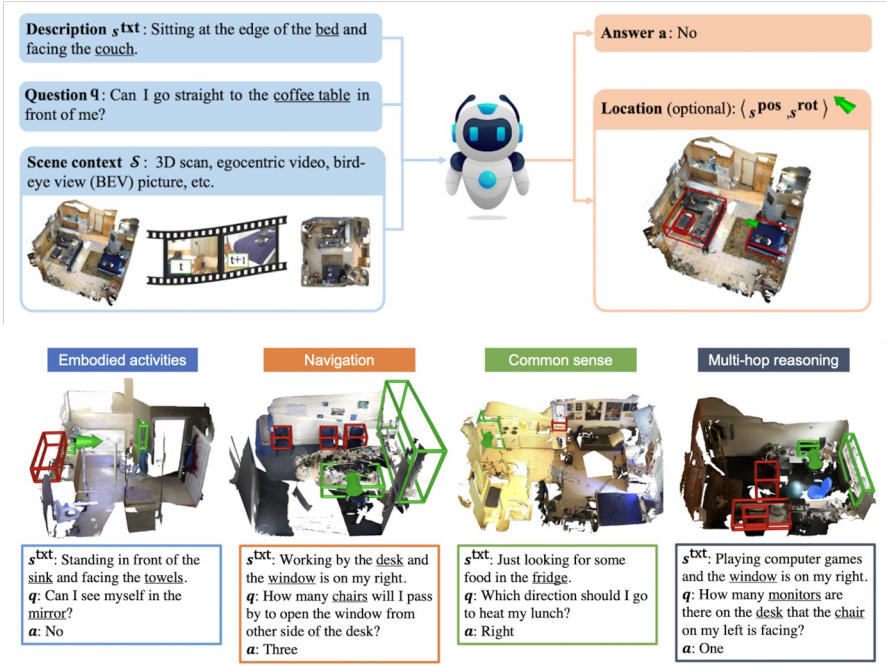
- 拡張性の高い（任意のモダリティを埋め込める）アーキテクチャのおかげで画像や音声などを取り込みながら質問応答が可能
- Chain-of-Thought Prompting（まず画像の内容を説明させて根拠（rationale）を生成、それを再入力して質問に回答）がMLLMの枠組みでも有効であることを確認

SQA3D: SITUATED QUESTION ANSWERING IN 3D SCENES

会議：ICLR 2023

著者：Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, Siyuan Huang

- 概要：エージェントに位置や方向などの情報をテキストで与え、それに基づいたQAを行う。
- 貢献：AMTにより大規模データセットを収集
- 実験：点群入力（ScanQAモデル）、動画・鳥瞰画像入力（ClipBERT）、点群入力（Scan2Cap+GPT-3 Zero-shot）など複数パターン
- 結果：ScanQAモデルが高精度だが人間の精度の半分くらい



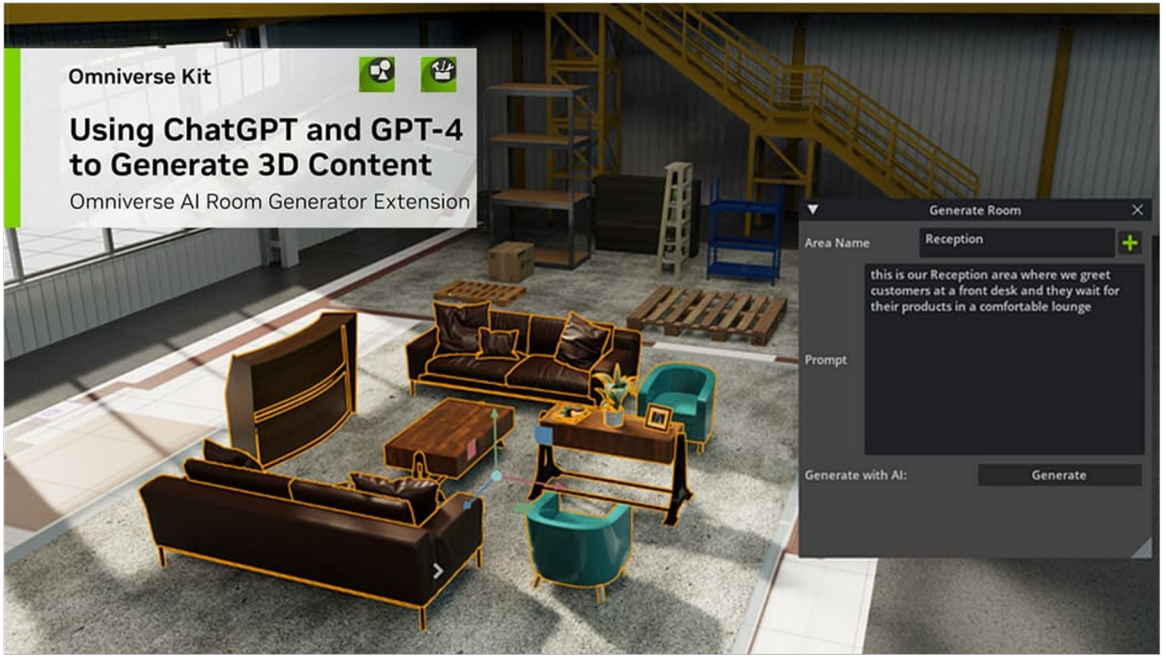
scan2capのGTキャプション＋GPT-3 Zero-Shotの場合どのくらいの精度になるのか気になる。

まとめBY: Fumiya Matsuzawa

Using ChatGPT and GPT-4 to Generate 3D Content

URL：https://medium.com/@nvidiaomniverse/chatgpt-and-gpt-4-for-3d-content-generation-9cbe5d17ec15

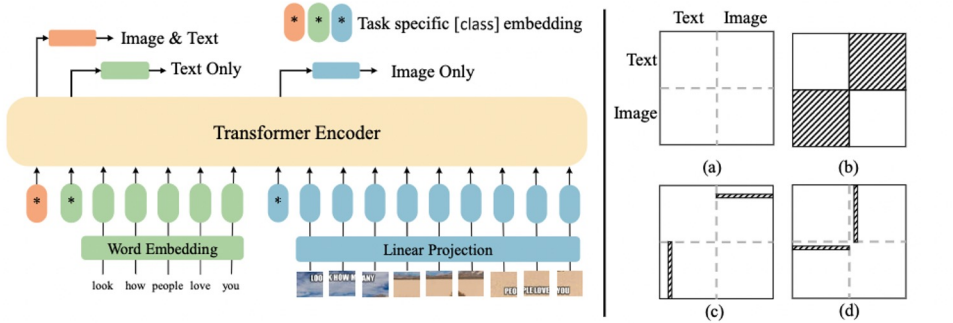
テキストベースで三次元シーン（家具の配置など）を自動生成
各家具のアフォーダンスを理解しリアルな配置が可能
（この家具とこの家具は近くにある。この位置関係の時、家具の向きは〇〇である。など）



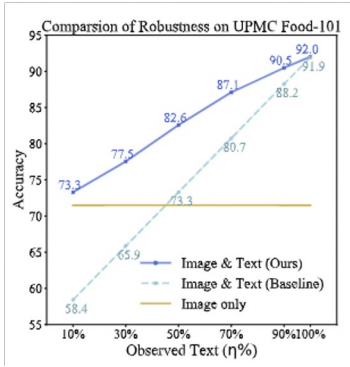
まとめBY: Fumiya Matsuzawa

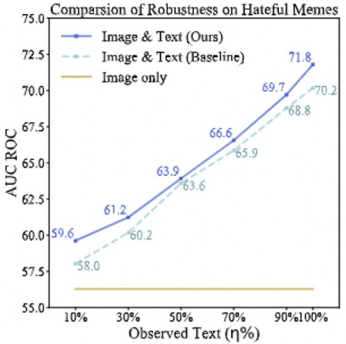
Are Multimodal Transformers Robust to Missing Modality?

- 概要：マルチモーダルなデータが一部欠損しているデータセットに対して、ベースラインよりも高い精度を達成できるマルチモーダルなTransformerの学習を提案
- 提案手法：image-only-task, text-only-task, image+text-taskそれぞれを合算したもの損失関数として定義、また、early fusion, late fusionどちらを用いた構造にするのか学習し最適化
- 実験内容：欠損したテストデータセットでベースラインとの比較
- 結果：シングルモダリティーで学習したTransformerよりも高い精度を達成(右図)



$$\mathcal{L} = \lambda_1 \mathcal{L}_{img}(\boldsymbol{x}^1; \boldsymbol{\theta}) + \lambda_2 \mathcal{L}_{txt}(\boldsymbol{x}^2; \boldsymbol{\theta}) + \lambda_3 \mathcal{L}_{it}(\boldsymbol{x}^1, \boldsymbol{x}^2; \boldsymbol{\theta}),$$





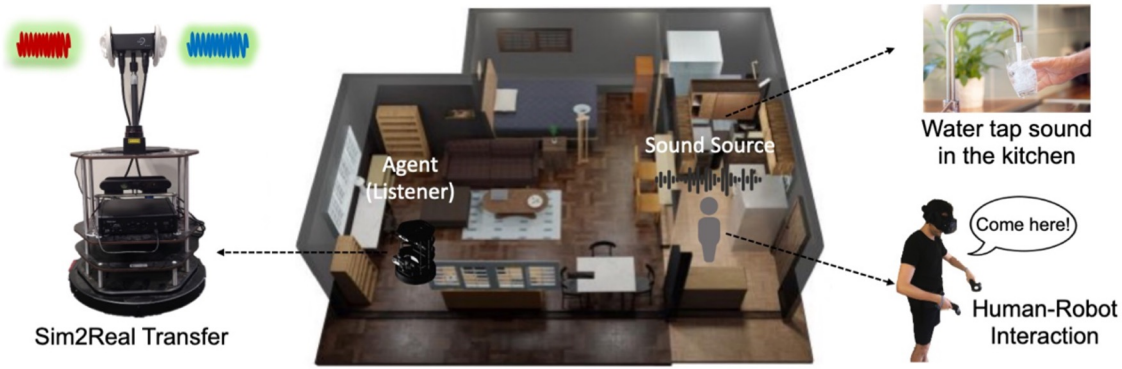
まとめBY: Okamoto

LLM+Robotics in ICRA 2023

Sonicverse: A Multisensory Simulation Platform for Embodied Household Agents That See and Hear

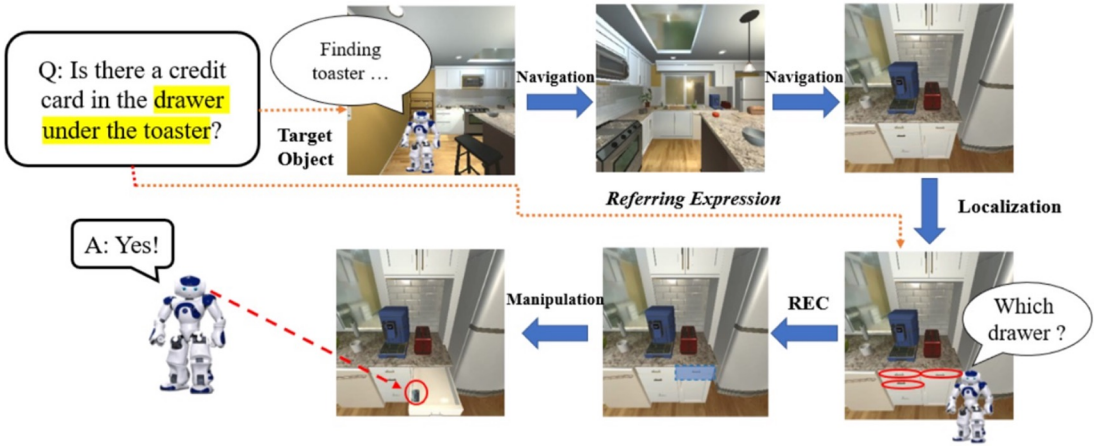
オーディオとビジョンによるMultimodal Embodied シミュレーションプラットフォームを提案。

Sim2Realでも効果を実証。VRゴーグルによってシミュレーション環境でインタラクションすることも可能。



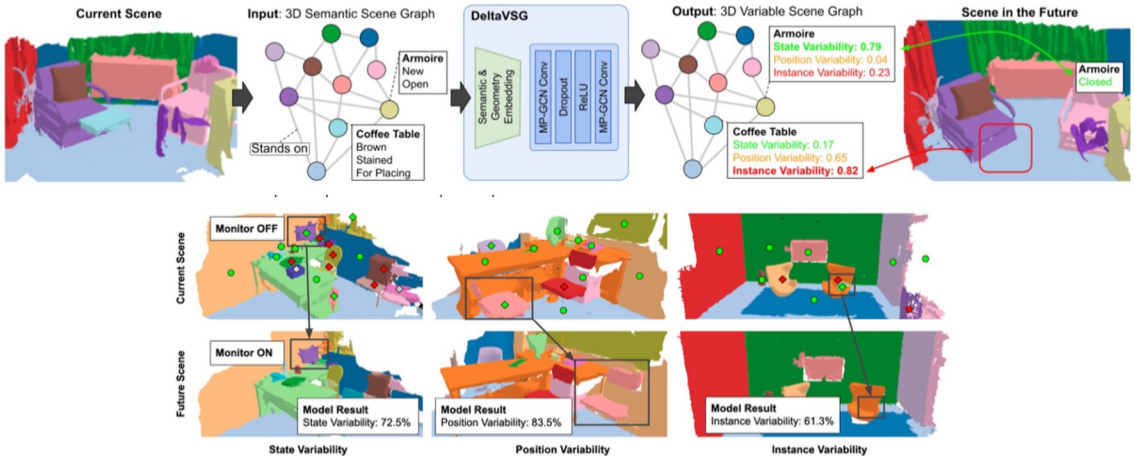
Embodied Referring Expression for Manipulation Question Answering in Interactive Environment

オブジェクトを操作しないと答えられないEmbodied QA タスクを提案。



3D-VSG: Long-Term Semantic Scene Change Prediction through 3D Variable Scene Graphs

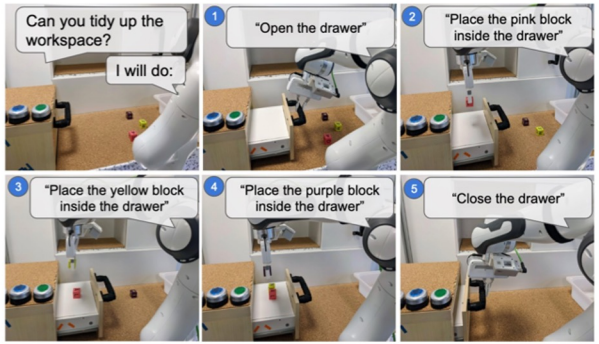
既存の3Dシーングラフに新たに可変性属性を追加した3DVSGを提案。モデルは現在のシーングラフを入力とし、それぞれのオブジェクトの将来の存在・状態・位置の変化可能性を予測？

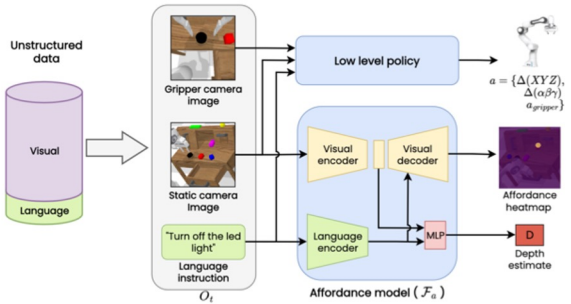


Grounding Language with Visual Affordances over Unstructured Data

①言語指示に対する人間のプレイデータから指示された物体の位置予測モデルを学習する
②言語条件付き模倣学習エージェントにより方策を学習
③LLMを用いて抽象指示をサブタスクに分解

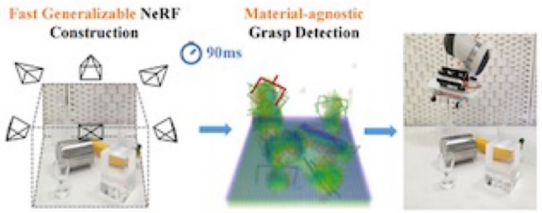
学習に必要な言語アノテーションが全データの1%ですむ（？）



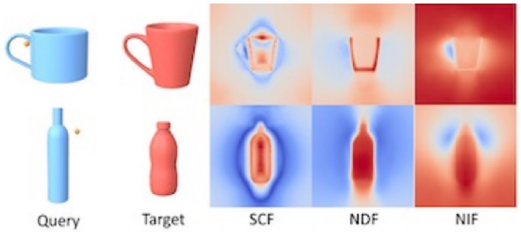


Neural FieldsをGraspタスクに応用

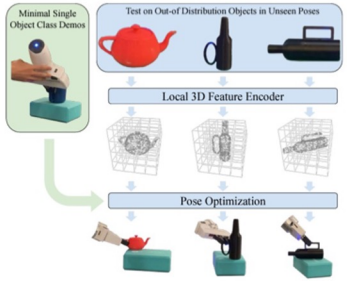
GraspNeRF



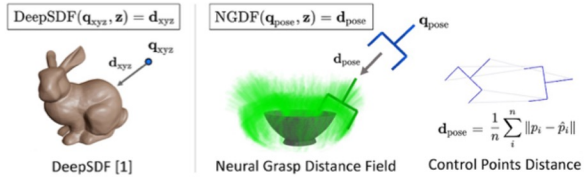
Neural Interaction Field



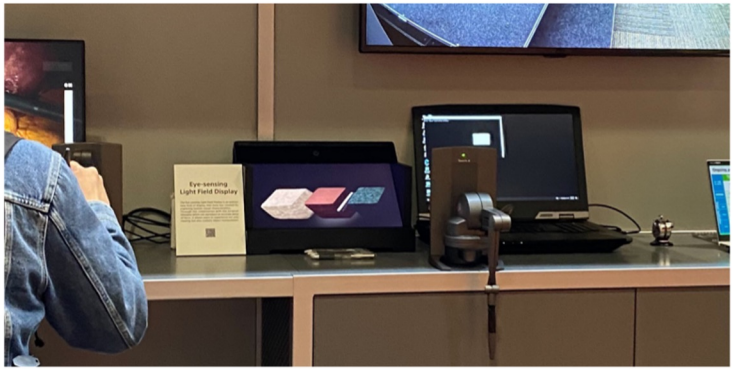
Local Neural Descriptor Fields

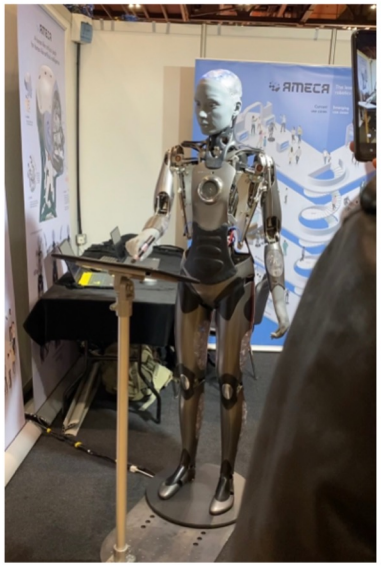


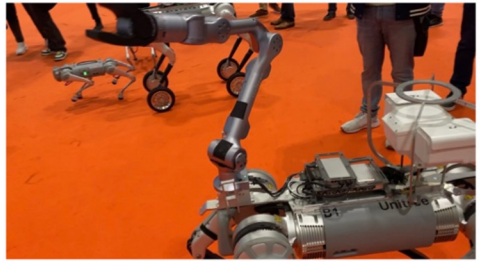
Neural Grasp Distance Field



企業展示







国・組織の声明

欧州のAI規制案(2021.4.21)

- 基本的人権や社会的弱者を守る上でのAIのリスクとその規制基準と方法について
- ヒトの潜在意識を制御を制御しようとしたり、身体的または精神的危害を与えうる行為は基本的に禁止

禁止度合は4つのレベルに分かれている

- 許容できないリスク：上記基本的人権を脅かし人に危害を加えうるAIは禁止
- 高リスク：顔認識による生体認証やインフラAI、テストの採点、労働者の雇用管理など、人の生活や安全に密接に関わるAIは高リスク扱い
- 低リスク：人間と対話するシステム、感情認識器、ディープフェイクなどのコンテンツ編集システムは生成物についての透明性（AI生成物であることの周知）のみを重視
- 最小リスク：その他のAI（スパムフィルタなど）

- AI悪用によるリスクの範囲は広く、人の生活の役に立つAIのほとんどが高リスク扱い
- 知的財産権の保護については思ったよりも簡素というのはちょっと意外

まとめBY: Seitaro Shinagawa