

# マルチモーダルモデルの動向

上野

# ■ 目次

- ✓ 大規模言語モデル
- ✓ マルチモーダルモデル
  - シングルタスクを解くマルチモーダルモデル
  - マルチタスクを解くマルチモーダルモデル

**～大規模言語モデル～**

# GPT-4 Technical Report, arXiv, 2023

- ✓ OpenAIによって発表されたLLM
  - チャット形式で応答できるようにRLHFで学習され、一般に公開
    - RLHF：人間のフィードバックに基づく強化学習
- ✓ モデル構造・データなどはすべて非公開
- ✓ 特にテキスト応答の性能が圧倒的であり、GPT-4とのチャット履歴を用いて新たなデータセットが構築されている
  - ShareGPTなど

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)	WinoGrande [56] Commonsense reasoning around pronoun resolution	87.5% 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	<b>86.4%</b> 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]	HumanEval [43] Python coding tasks	<b>67.0%</b> 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [57]
HellaSwag [52] Commonsense reasoning around everyday events	<b>95.3%</b> 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [53]	DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	<b>88.4</b> QDGAT [59]
AI2 Reasoning Challenge (ARC) [54] Grade-school multiple choice science questions. Challenge-set.	<b>96.3%</b> 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [55]	86.5% ST-MOE [18]	GSM-8K [60] Grade-school mathematics questions	<b>92.0%*</b> 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [61]	87.3% Chinchilla + SFT+ORM-RL, ORM reranking [62]

# ■ LLaMA: Open and Efficient Foundation Language Models, 2023, arXiv

- ✓ Metaより発表されたLLM(ソースコード公開, 商用利用不可)
- ✓ 既存のLLMで提案されたアーキテクチャを組み合わせて高性能を達成した

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	<b>88.0</b>	82.3	-	83.4	<b>81.1</b>	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	<b>80.0</b>	<b>57.8</b>	58.6
	65B	85.3	<b>82.8</b>	<b>52.3</b>	<b>84.2</b>	77.0	78.9	56.0	<b>60.2</b>

# Llama 2: Open Foundation and Fine-Tuned Chat Model, 2024, arXiv

- ✓ LLaMAをベースとした商用利用可能なオープンソースのLLM
  - RLHFを用いてChatも可能なLLaMA2-Chatもリリース
- ✓ モデル構造はLLaMAからAttentionの構造のみ変更
  - Queryをグループ化して計算することで速度を犠牲にして性能を向上
- ✓ 2兆トークンのテキストデータで学習 (LLaMAは1.4兆)

Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
MPT	7B	20.5	57.4	41.0	57.5	4.9	26.8	31.0	23.5
	30B	28.9	64.9	50.0	64.7	9.1	46.9	38.0	33.8
Falcon	7B	5.6	56.1	42.8	36.0	4.6	26.2	28.0	21.2
	40B	15.2	69.2	56.7	65.7	12.6	55.4	37.1	37.0
LLAMA 1	7B	14.1	60.8	46.2	58.5	6.95	35.1	30.3	23.9
	13B	18.9	66.1	52.6	62.3	10.9	46.9	37.0	33.9
	33B	26.0	70.0	58.4	67.6	21.4	57.8	39.8	41.7
	65B	30.7	70.7	60.5	68.6	30.8	63.4	43.5	47.6
LLAMA 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
	34B	27.8	69.9	58.7	68.0	24.2	62.6	44.1	43.4
	70B	37.5	71.9	63.6	69.4	35.2	68.9	51.2	54.2

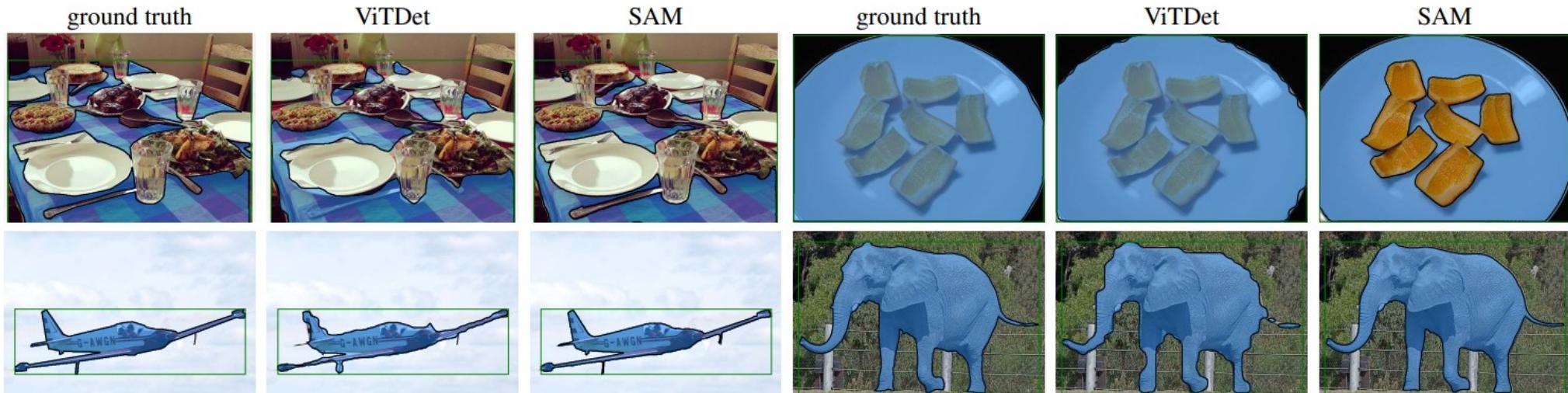
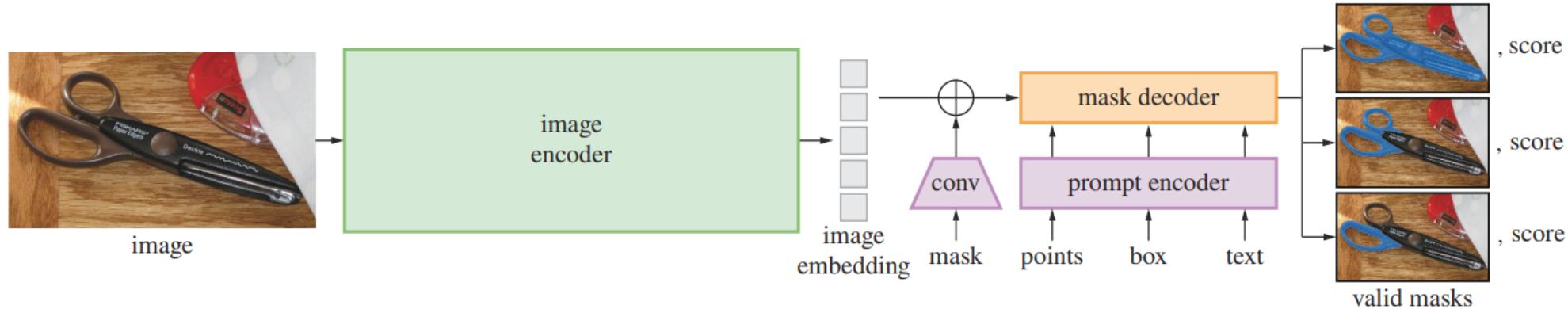
# Introducing Meta Llama 3: The most capable openly available LLM to date, 2024, Meta

- ✓ 8Bモデル、70Bモデルの構造はLLaMA2から変わらず
  - トークナイザは128Kにスケーリング
- ✓ 15兆トークンで学習（95%が英語、残りが多言語30種）
  - 学習に使用したGPUは24Kクラスタ×2
- ✓ 今後、400Bモデル・マルチリングルモデル・マルチモーダルモデルが登場予定
  - 同時にTechnical Reportも公開予定

～シングルタスクを解くマルチモーダルモデル～

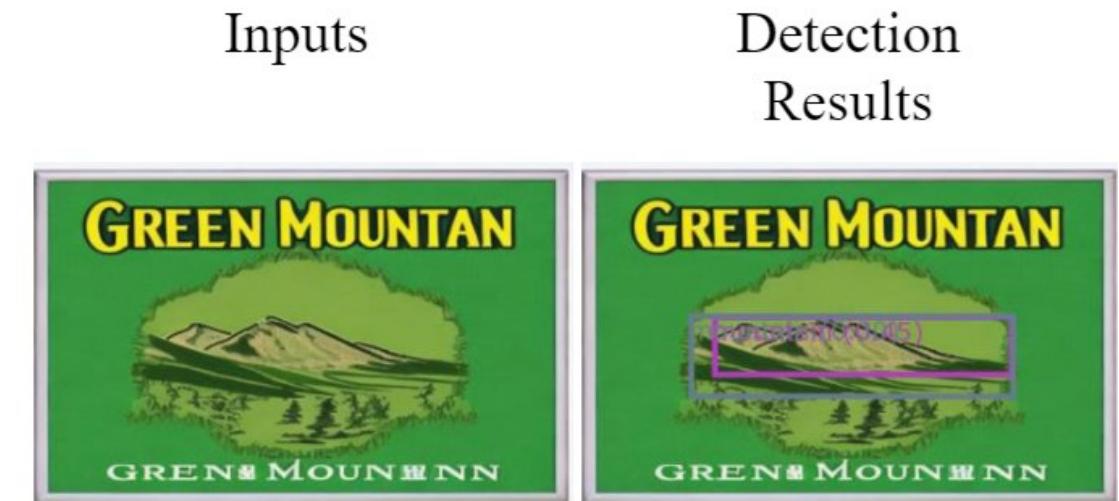
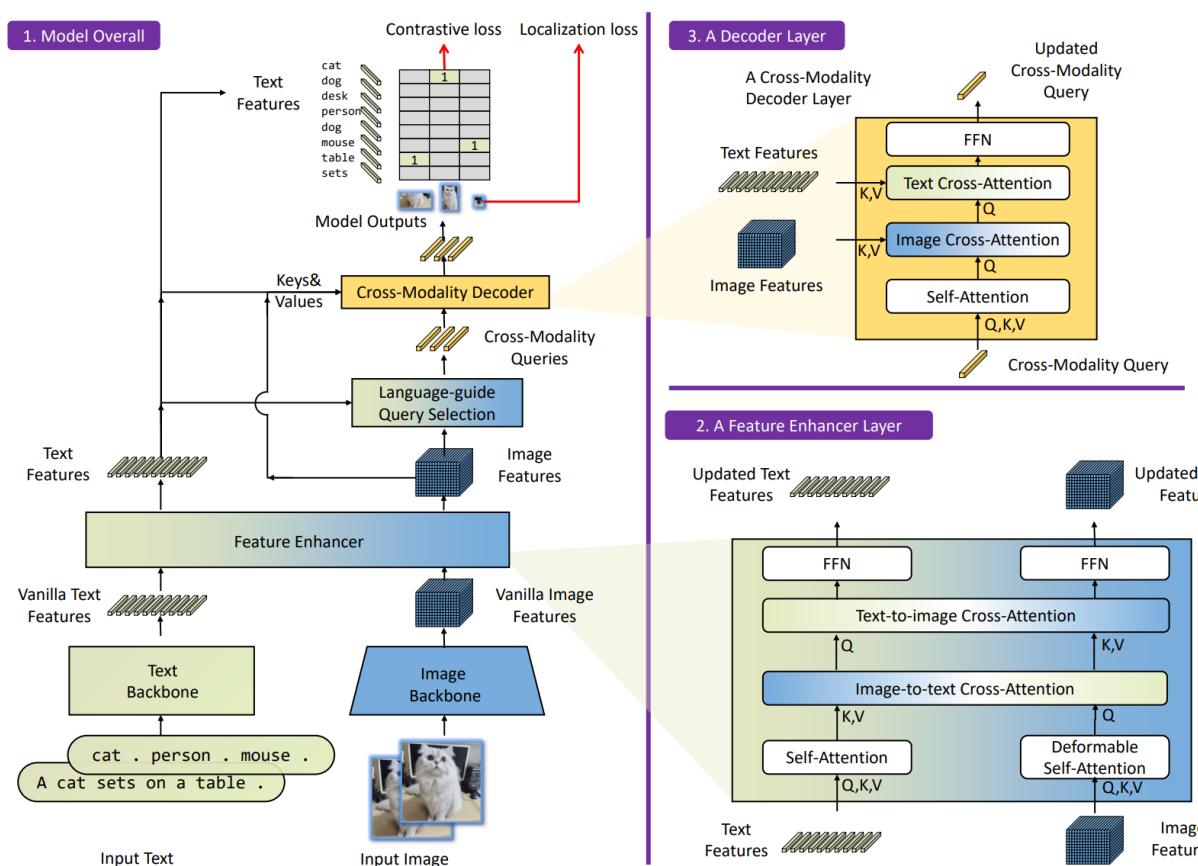
# Segment Anything, 2023, arXiv

- ✓ 言語や画像によってプロンプト可能なセグメンテーションモデル
  - 座標・Bounding Box・テキストなどが利用可能



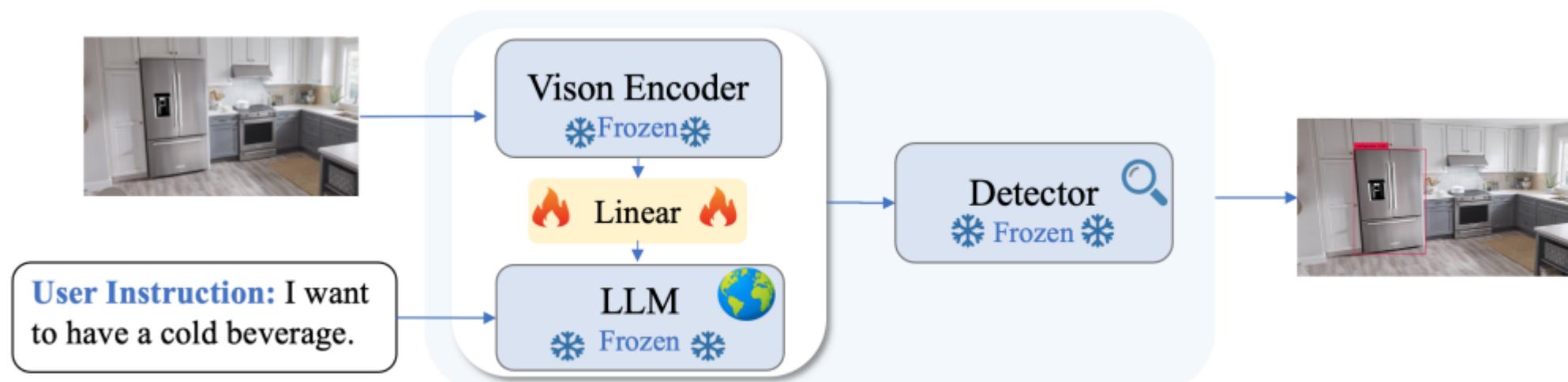
# Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, 2023, arXiv

- ✓ 物体検出モデルのDINOを拡張し、Groundingタスクを解けるようにテキストを処理するTransformerを導入
  - Grounding : 言語と対応する画像中の領域を参照するタスク



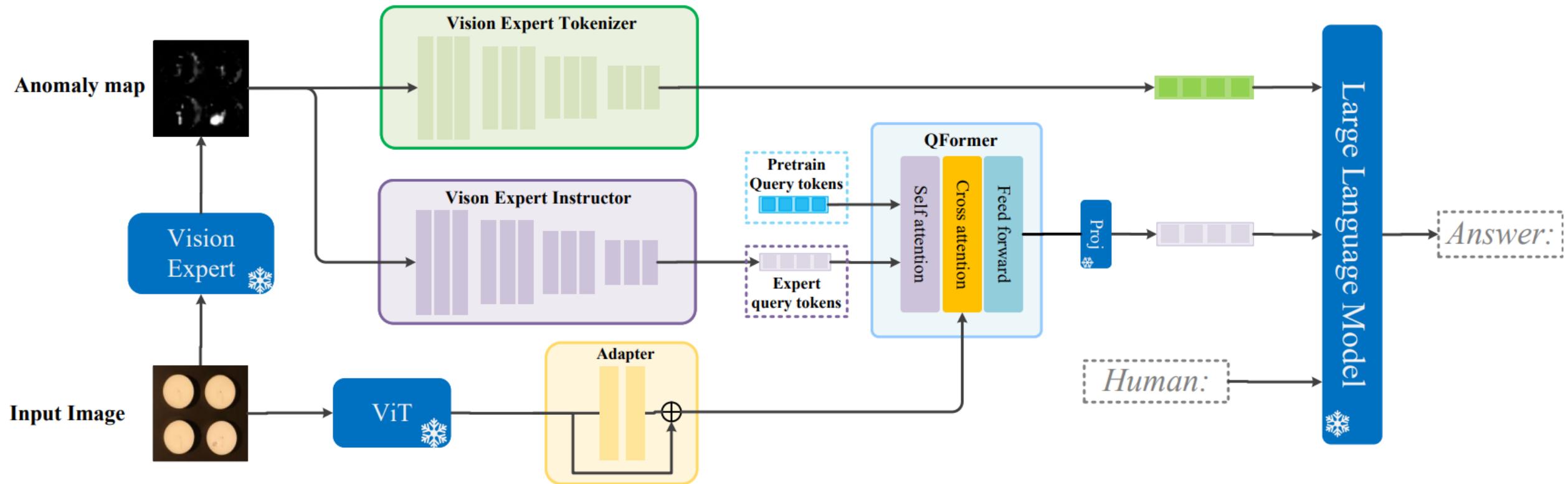
# ■ DetGPT: Detect What You Need via Reasoning, 2023, arXiv

- ✓ Groundingにおいて入力テキストが物体を明示していない場合も、推論によって正しい物体を検出可能なモデル
  - 「冷たい飲み物が欲しい」という入力から、冷たい飲み物が存在する(確率が高い)冷蔵庫を検出
- ✓ 推論能力を強化するためChatGPTとの会話ログをデータセットに使用



# Myriad: Large Multimodal Model by Applying Vision Experts for Industrial Anomaly Detection, 2023, arXiv

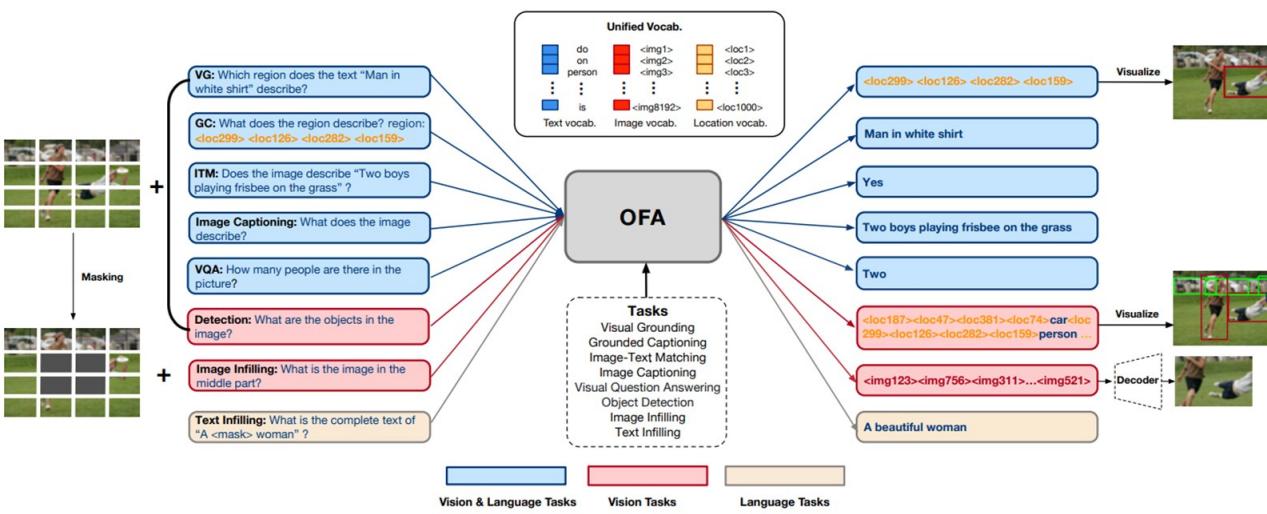
- ✓ 異常度マップとViT(視覚モデル)の特徴量をLLMに入力して扱うことで工業製品の外観検査とインタラクションが可能なモデル
  - 異常度マップはPadimやAnomalyGPT, PatchCoreなどの既存モデルを使用



～マルチタスクを解くマルチモーダルモデル～

# OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework, 2022, ICML

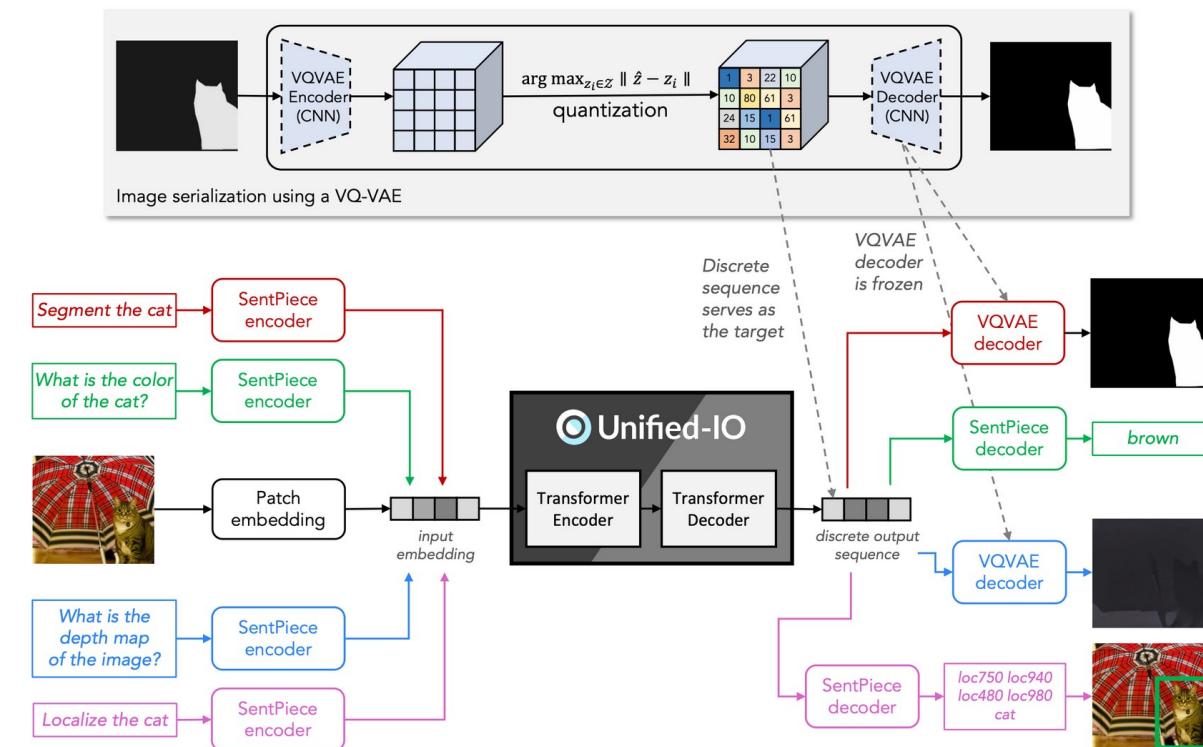
- ✓ 画像と言語をシーケンスとして扱うマルチモーダルモデル
- ✓ 物体検出などの画像認識から、VQAやキャプショニングなどの視覚言語タスクを同時に学習・推論可能



Model	SST-2	RTE	MRPC	QQP	MNLI	QNLI
<i>Multimodal Pretrained Baseline Models</i>						
VisualBERT [38]	89.4	56.6	71.9	89.4	81.6	87.0
UNITER [14]	89.7	55.6	69.3	89.2	80.9	86.0
VL-BERT [8]	89.8	55.7	70.6	89.0	81.2	86.3
VilBERT [13]	90.4	53.7	69.0	88.6	79.9	83.8
LXMERT [40]	90.2	57.2	69.8	75.3	80.4	84.2
Uni-Perceiver [61]	90.2	64.3	86.6	87.1	81.7	89.9
SimVLM [22]	90.9	63.9	75.2	90.4	83.4	88.6
FLAVA [60]	90.9	57.8	81.4	90.4	80.3	87.3
UNIMO [46]	96.8	-	-	-	89.8	-
<i>Natural-Language-Pretrained SOTA Models</i>						
BERT [2]	93.2	70.4	88.0	91.3	86.6	92.3
RoBERTa [28]	96.4	86.6	90.9	92.2	90.2	93.9
XLNet [25]	<b>97.0</b>	85.9	90.8	92.3	90.8	94.9
ELECTRA [82]	96.9	88.0	90.8	92.4	90.9	95.0
DeBERTa [83]	96.8	88.3	<b>91.9</b>	92.3	<b>91.1</b>	<b>95.3</b>
<i>Ours</i>						
OFA	96.6	<b>91.0</b>	91.7	<b>92.5</b>	90.2	94.8

# Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks, 2022, arXiv

- ✓ 単一のTransformerとVQ-VAEのヘッドを活用し、 CVとNLPの90を超える多様なデータセットで学習したモデル
  - ゼロショットでOFAよりも複数のタスクに対して高精度を達成



	restricted	params (M)	Categorization		Localization		VQA		Refexp		Segmentation		Keypoint	Normal
			same	new	same	new	same	new	same	new	same	new	same	new
0	NLL-AngMF	✓	72	-	-	-	-	-	-	-	-	-	-	<b>50.7</b>
1	Mask R-CNN	✓	58	-	-	51.9	40.8	-	-	-	44.9	0.3	<b>70.9</b>	-
2	GPV-1	✓	236	58.7	0.8	48.3	37.8	58.4	74.0	29.7	23.1	-	-	-
3	CLIP		302	49.1	46.7	-	-	-	-	-	-	-	-	-
4	OFA <sub>LARGE</sub>		473	28.9	15.8	-	-	74.9	88.6	63.4	58.5	-	-	-
5	GPV-2		370	<b>85.0</b>	13.5	54.6	54.2	69.8	81.7	57.8	48.3	-	-	-
6	UNIFIED-IO <sub>SMALL</sub>		71	52.9	31.9	47.5	61.5	59.0	72.5	54.2	45.7	37.4	48.5	46.6
7	UNIFIED-IO <sub>BASE</sub>		241	60.3	47.5	57.9	68.4	68.0	81.8	72.5	62.2	45.8	57.2	60.2
8	UNIFIED-IO <sub>LARGE</sub>		776	63.0	52.7	63.3	70.9	72.1	84.3	79.2	66.3	50.4	62.2	67.7
9	UNIFIED-IO <sub>XL</sub>		2925	66.1	<b>60.1</b>	<b>65.6</b>	<b>74.4</b>	<b>78.6</b>	<b>90.2</b>	<b>83.5</b>	<b>72.4</b>	<b>53.0</b>	<b>64.2</b>	68.2

# Flamingo: a Visual Language Model for Few-Shot Learning, 2022, NeurIPS

- ✓ 複数画像とそれに対応する言語入力を可能とし、対照学習や In-Context Learning（後述）を可能に

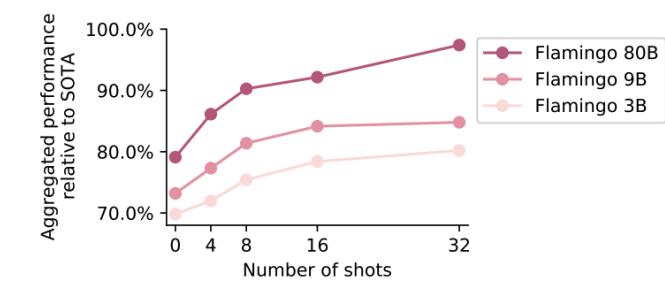
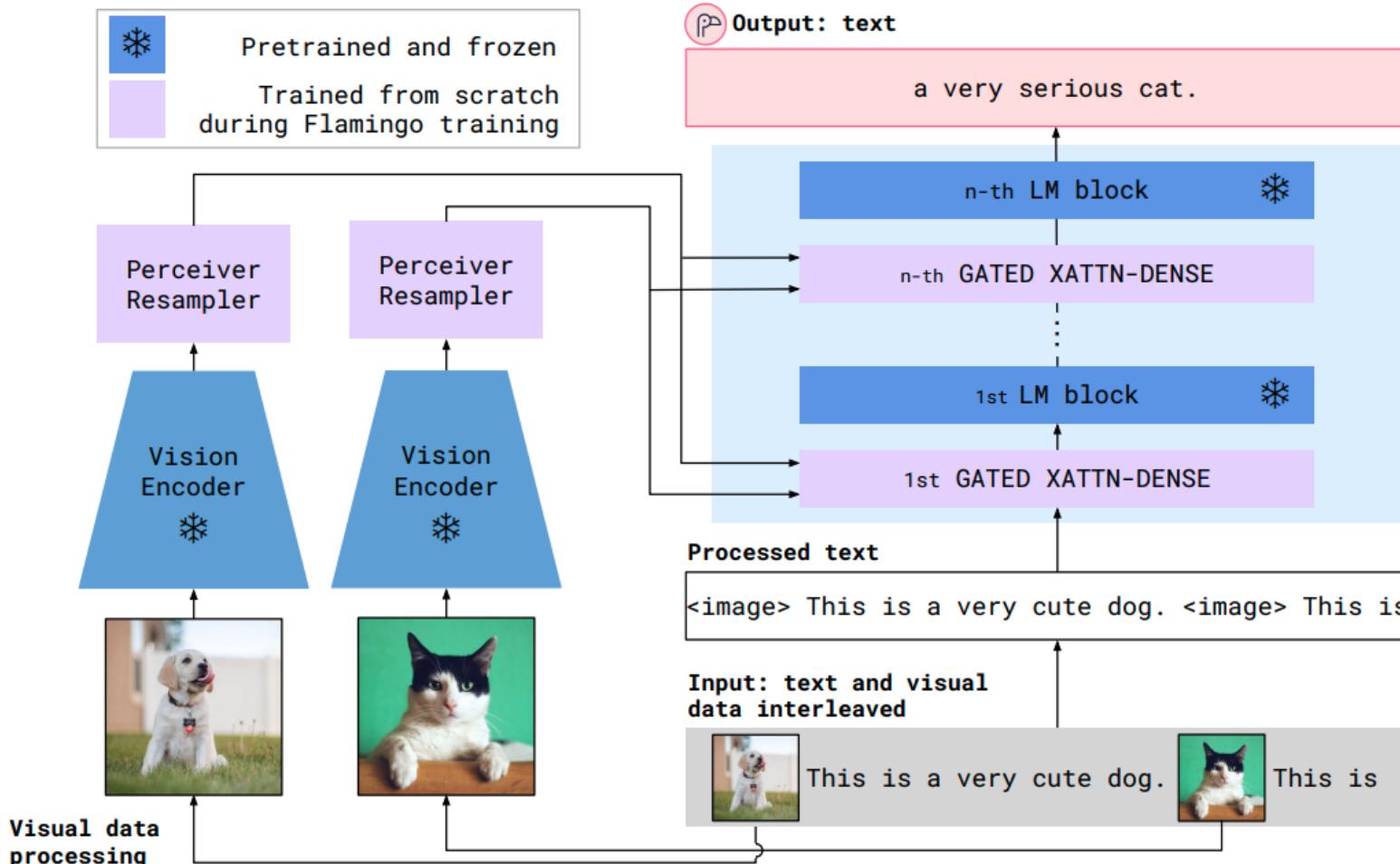


Figure 9 | Overall impact of model scaling and number of shots. The performance of Flamingo models increases with their parametric size and with the number of in-context shot. Performance is reported by averaging the SotA relative score across the 16 benchmarks from Section 4.2.1.

# (補足) ICL: In-Context Learning

- ✓ 入力と出力の例をプロンプトとして与えることで、モデルにタスクを理解させ出力を強制する技術
- ✓ GPT-3で提案されて以降、特にNLPの分野で使われてきた

入出力の例から文脈を理解して回答

$$\begin{aligned} 6 + 8 &= 14 \\ 8 + 2 &= 10 \\ 2 + 0 &= 2 \\ 3 + 4 &= ? \end{aligned} \quad \left. \right\} \text{例示}$$

→ 7 回答

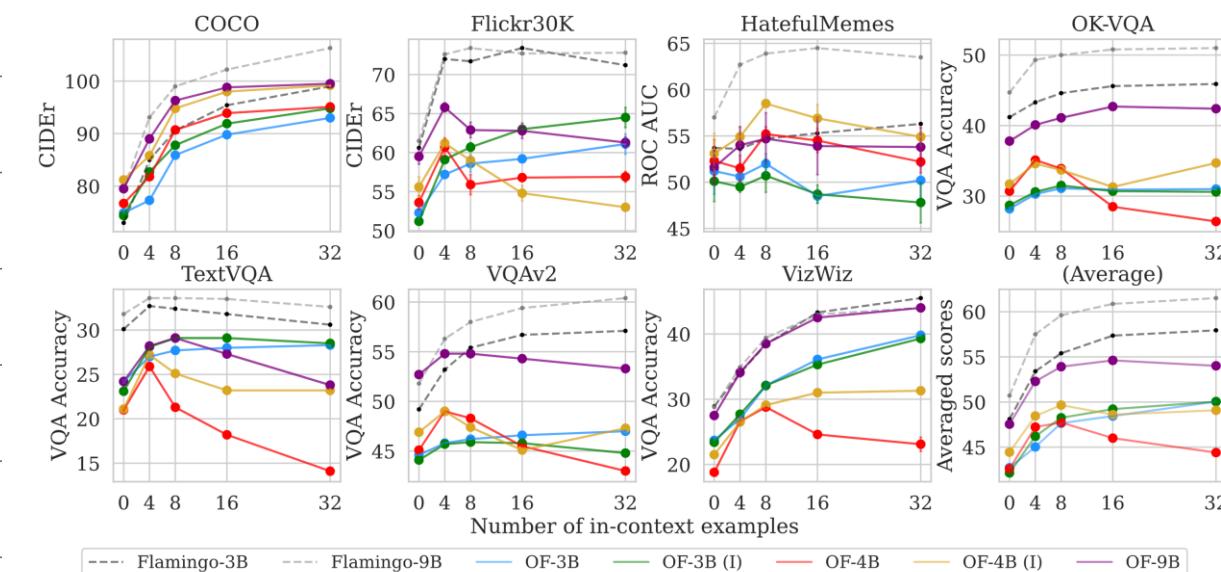
$$\begin{aligned} \text{Hello} &\rightarrow \text{こんにちは} \\ \text{World} &\rightarrow \text{世界} \\ \text{Dog} &\rightarrow \text{犬} \\ \text{Bread} &\rightarrow ? \end{aligned} \quad \left. \right\} \text{例示}$$

→ パン 回答

# OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models, 2023, arXiv

- ✓ Flamingoがコード・データとともに非公開だったため  
Flamingoを再現し、オープンソースで公開
  - 基本的な構造は全てFlamingoに準拠

Benchmark	Shots	Fl-3B	Fl-9B	OF-3B	OF-3B (I)	OF-4B	OF-4B (I)	OF-9B
COCO [7]	0	73.0	79.4	74.9 (0.2)	74.4 (0.6)	76.7 (0.2)	<b>81.2 (0.3)</b>	79.5 (0.2)
	4	85.0	<b>93.1</b>	77.3 (0.3)	82.7 (0.7)	81.8 (0.4)	85.8 (0.5)	89.0 (0.3)
	32	99.0	<b>106.3</b>	93.0 (0.6)	94.8 (0.3)	95.1 (0.3)	99.2 (0.3)	99.5 (0.1)
Flickr-30K [40]	0	60.6	<b>61.5</b>	52.3 (1.0)	51.2 (0.2)	53.6 (0.9)	55.6 (1.3)	59.5 (1.0)
	4	72.0	<b>72.6</b>	57.2 (0.4)	59.1 (0.3)	60.7 (1.2)	61.2 (0.5)	65.8 (0.6)
	32	71.2	<b>72.8</b>	61.1 (1.3)	64.5 (1.3)	56.9 (0.7)	53.0 (0.5)	61.3 (0.7)
VQAv2 [2]	0	49.2	51.8	44.6 (0.0)	44.1 (0.1)	45.1 (0.1)	46.9 (0.0)	<b>52.7 (0.2)</b>
	4	53.2	<b>56.3</b>	45.8 (0.0)	45.7 (0.1)	49.0 (0.0)	49.0 (0.0)	54.8 (0.0)
	32	57.1	<b>60.4</b>	47.0 (0.1)	44.8 (0.1)	43.0 (0.2)	47.3 (0.0)	53.3 (0.1)
OK-VQA [26]	0	41.2	<b>44.7</b>	28.2 (0.2)	28.7 (0.1)	30.7 (0.1)	31.7 (0.1)	37.8 (0.2)
	4	43.3	<b>49.3</b>	30.3 (0.5)	30.6 (0.2)	35.1 (0.0)	34.6 (0.0)	40.1 (0.1)
	32	45.9	<b>51.0</b>	31.0 (0.1)	30.6 (0.1)	26.4 (0.2)	34.7 (0.3)	42.4 (0.0)
TextVQA [33]	0	30.1	<b>31.8</b>	24.2 (0.2)	23.1 (0.2)	21.0 (0.3)	21.1 (0.4)	24.2 (0.5)
	4	32.7	<b>33.6</b>	27.0 (0.3)	28.1 (0.4)	25.9 (0.0)	27.2 (0.3)	28.2 (0.4)
	32	30.6	<b>32.6</b>	28.3 (0.2)	28.5 (0.1)	14.1 (0.2)	23.2 (0.2)	23.8 (0.2)
VizWiz [11]	0	<b>28.9</b>	28.8	23.7 (0.5)	23.4 (0.3)	18.8 (0.1)	21.5 (0.2)	27.5 (0.2)
	4	34.0	<b>34.9</b>	27.0 (0.3)	27.7 (0.1)	26.6 (0.5)	26.5 (0.4)	34.1 (0.7)
	32	<b>45.5</b>	44.0	39.8 (0.1)	39.3 (0.4)	23.1 (1.1)	31.3 (0.2)	44.0 (0.5)
HatefulMemes [15]	0	53.7	<b>57.0</b>	51.2 (2.5)	50.1 (2.2)	52.3 (2.3)	53.1 (2.2)	51.6 (1.8)
	4	53.6	<b>62.7</b>	50.6 (0.8)	49.5 (0.6)	51.5 (1.4)	54.9 (1.1)	54.0 (2.0)
	32	56.3	<b>63.5</b>	50.2 (1.8)	47.8 (2.2)	52.2 (1.2)	54.9 (1.1)	53.8 (2.1)



# Otter: A Multi-Modal Model with In-Context Instruction Tuning, 2023, arXiv

- ✓ OpenFlamingoをベースに、 In-Context Learning能力を強化する為のデータセットで追加学習されたモデル
  - モデルは (Open) Flamingoに準拠

User provided examples



**Instruction:** What is the danger of this sport?

**Answer:** The player may get hit by the tennis ball.

**Instruction:** What is the danger of this sport?

**Answer:** While chasing the baseball, the player may inadvertently collide with other players.

Otter reaction

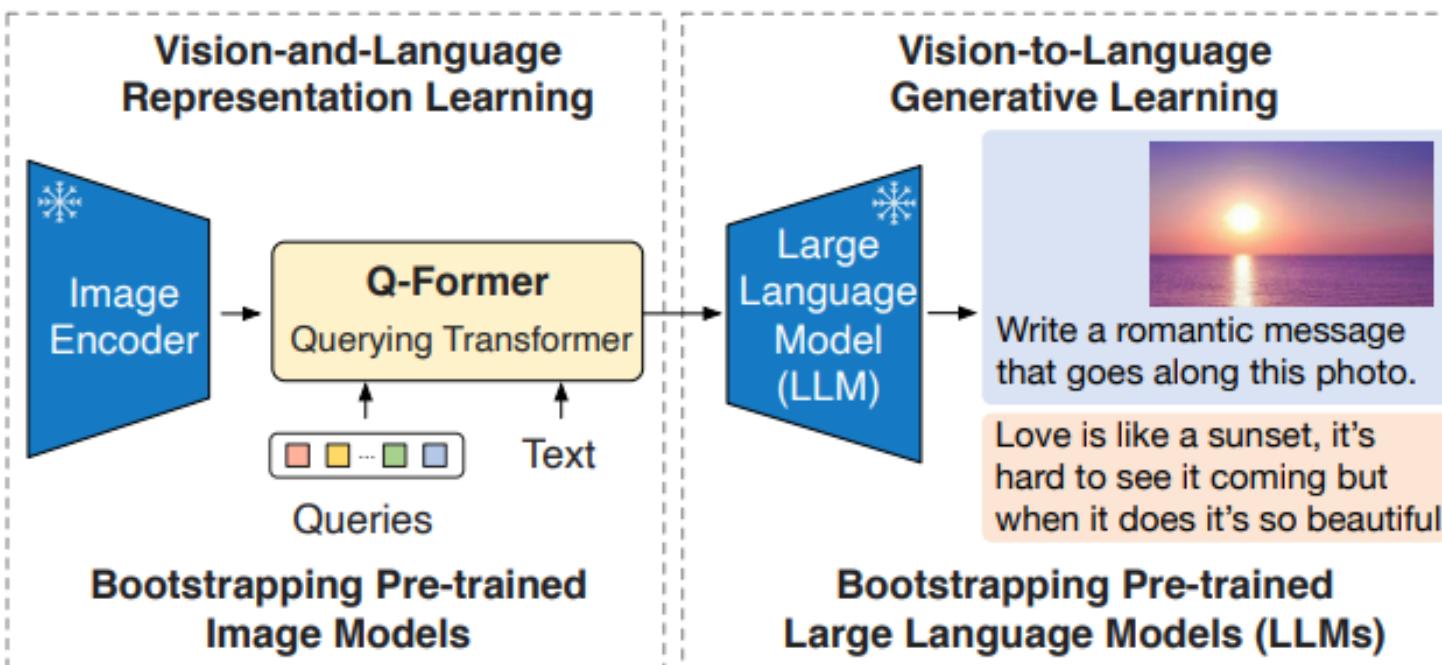


**Instruction:** What is the danger of this sport?

**Otter:** The sport involves players running and trying to catch the ball while others are standing in the grass, which can lead to collisions or accidents. These collisions can result in injuries to the players involved. It is essential for players to be aware of their surroundings and maintain a safe distance from each other to minimize the risk of accidents during the game.

# ■ BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, 2023, arXiv

- ✓ 画像特徴と言語特徴の融合に特化したQuerying Transformerを提案
  - それまでのマルチモーダルモデルでは、特徴量の融合はMLP, Attention, Concatなどが利用されてきた



Models	#Trainable Params	#Total Params	VQAv2		OK-VQA test	GQA test-dev
			val	test-dev		
VL-T5 <sub>no-vqa</sub>	224M	269M	13.5	-	5.8	6.3
FewVLM (Jin et al., 2022)	740M	785M	47.7	-	16.5	29.3
Frozen (Tsimpoukelli et al., 2021)	40M	7.1B	29.6	-	5.9	-
VLKD (Dai et al., 2022)	406M	832M	42.6	44.5	13.3	-
Flamingo3B (Alayrac et al., 2022)	1.4B	3.2B	-	49.2	41.2	-
Flamingo9B (Alayrac et al., 2022)	1.8B	9.3B	-	51.8	44.7	-
Flamingo80B (Alayrac et al., 2022)	10.2B	80B	-	56.3	<b>50.6</b>	-
BLIP-2 ViT-L OPT <sub>2.7B</sub>	104M	3.1B	50.1	49.7	30.2	33.9
BLIP-2 ViT-g OPT <sub>2.7B</sub>	107M	3.8B	53.5	52.3	31.7	34.6
BLIP-2 ViT-g OPT <sub>6.7B</sub>	108M	7.8B	54.3	52.6	36.4	36.4
BLIP-2 ViT-L FlanT5 <sub>XL</sub>	103M	3.4B	62.6	62.3	39.4	<u>44.4</u>
BLIP-2 ViT-g FlanT5 <sub>XL</sub>	107M	4.1B	<u>63.1</u>	<u>63.0</u>	40.7	44.2
BLIP-2 ViT-g FlanT5 <sub>XXL</sub>	108M	12.1B	<b>65.2</b>	<b>65.0</b>	<u>45.9</u>	<b>44.7</b>

# The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision), 2023, arXiv

- ✓ GPT-4に画像入力が可能となり、マルチモーダルモデルに進化
- ✓ モデル構造などはGPT-4と同様、非公開

## Prompt:

Describe the image



## GPT-4V:

The image shows a collage of 8 prominent personalities from different fields. On the top row, from left to right, there is singer Justin Bieber, footballer Lionel Messi, footballer Cristiano Ronaldo, and entrepreneur Elon Musk. On the bottom row, from left to right, there is model and reality TV star Kylie Jenner, singer Taylor Swift, actor Robert Downey Jr., and actor Scarlett Johansson. All individuals are shown in different poses and outfits, with various backgrounds.

## Zero-shot (w/ text instructions, think step-by-step)

### Prompt:

What is the read of the speed meter? To read this speed meter, observe the needle on the outer edge of the dial which points to your current speed represented in miles per hour. Yellow is the color of the needle. Let's think step-by-step.



Note: Same query image as shown above, resized to save space

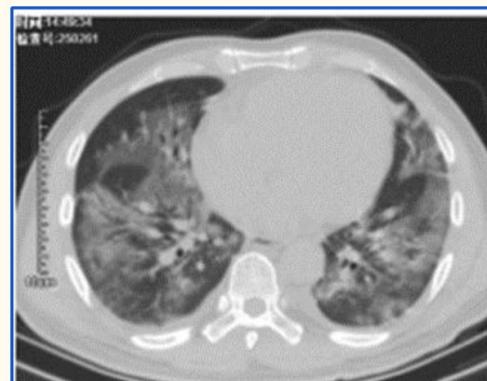
### GPT-4V:

1. First, look at the outer edge of the dial and find the numbers that represent miles per hour.
2. Next, find the yellow needle that points to the current speed.
3. Observe where the needle is pointing and read the number on the outer edge of the dial.
4. This number is the current speed in miles per hour.

In this example, the needle is pointing to approximately 40 mph, so the current speed is 40 miles per hour.

### Prompt:

Look at the CT scan of lung in the image below, tell me what's wrong.



### GPT-4V:

There appears to be multiple areas of consolidation and ground-glass opacities in both lungs, which could indicate a lung infection or inflammation. There is also a possible mass or nodule in the right upper lobe. Further evaluation and clinical correlation is needed to determine the exact diagnosis.

# (LLaVA) Visual Instruction Tuning, 2023, NeurIPS

- ✓ GPT-4を用いて作成したInstruction Following Datasetで学習したモデル
- ✓ モデル構造はLLaMA(言語モデル)×CLIP ViT(視覚モデル)×線形射影(融合)



Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

User  
LLaVA

What is unusual about this image?

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.

[Start a new conversation, and clear the history]

What's happening in the scene?

The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.

User  
LLaVA

Visual features	Before	Last
Best variant	90.92	89.96 (-0.96)
Predict answer first	-	89.77 (-1.15)
Training from scratch	85.81 (-5.11)	-
7B model size	89.84 (-1.08)	-

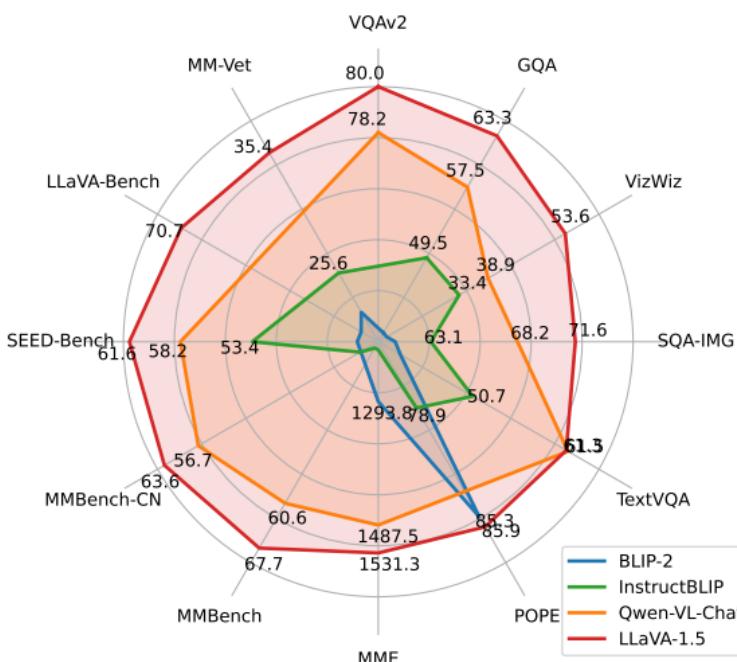
Table 7: Design choice ablations (%). The difference with the best variant is reported in red text.

# Improved Baselines with Visual Instruction

## Tuning : LLaVA 1.5, 2023, arXiv

✓ LLaVAから4つの変更を施し性能向上

- プロンプトの調整
- 線形射影を2層MLPに変更してモダリティを融合
- アカデミックタスク指向のデータセット, ShareGPTのデータセットを追加
- 画像サイズを224から336にスケーリング
- LLMを13Bにスケーリング



Visual input example, Tricky Question:



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

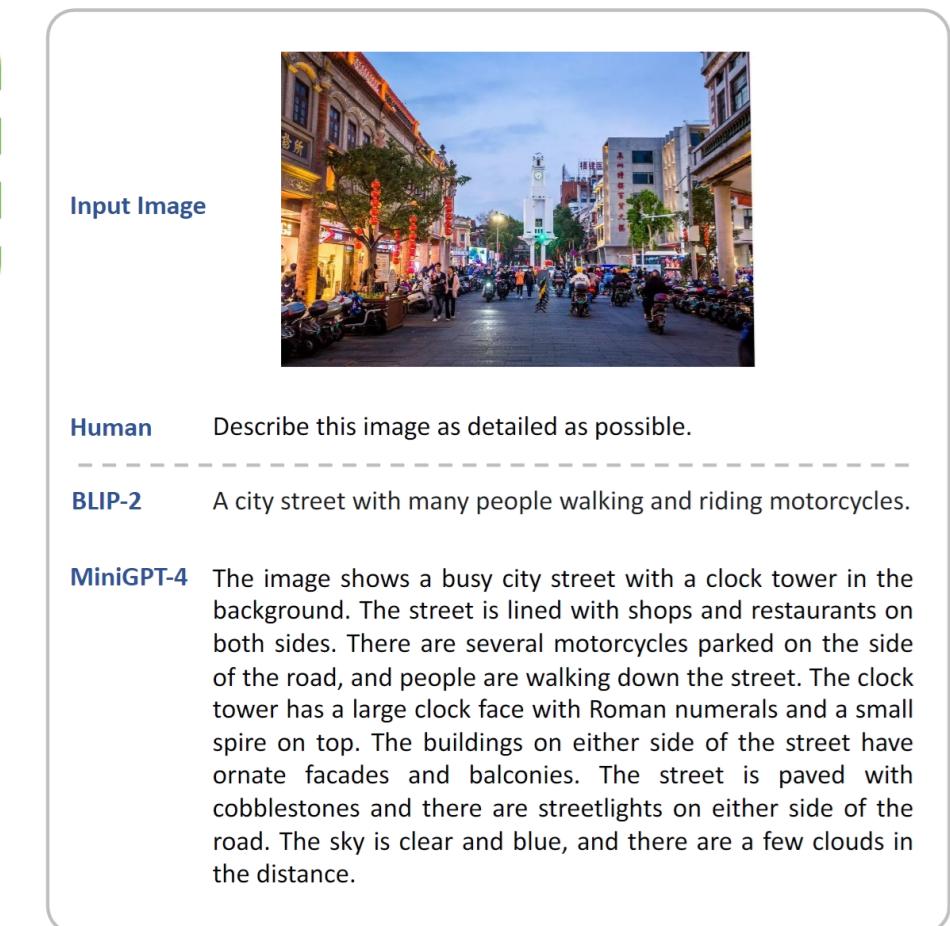
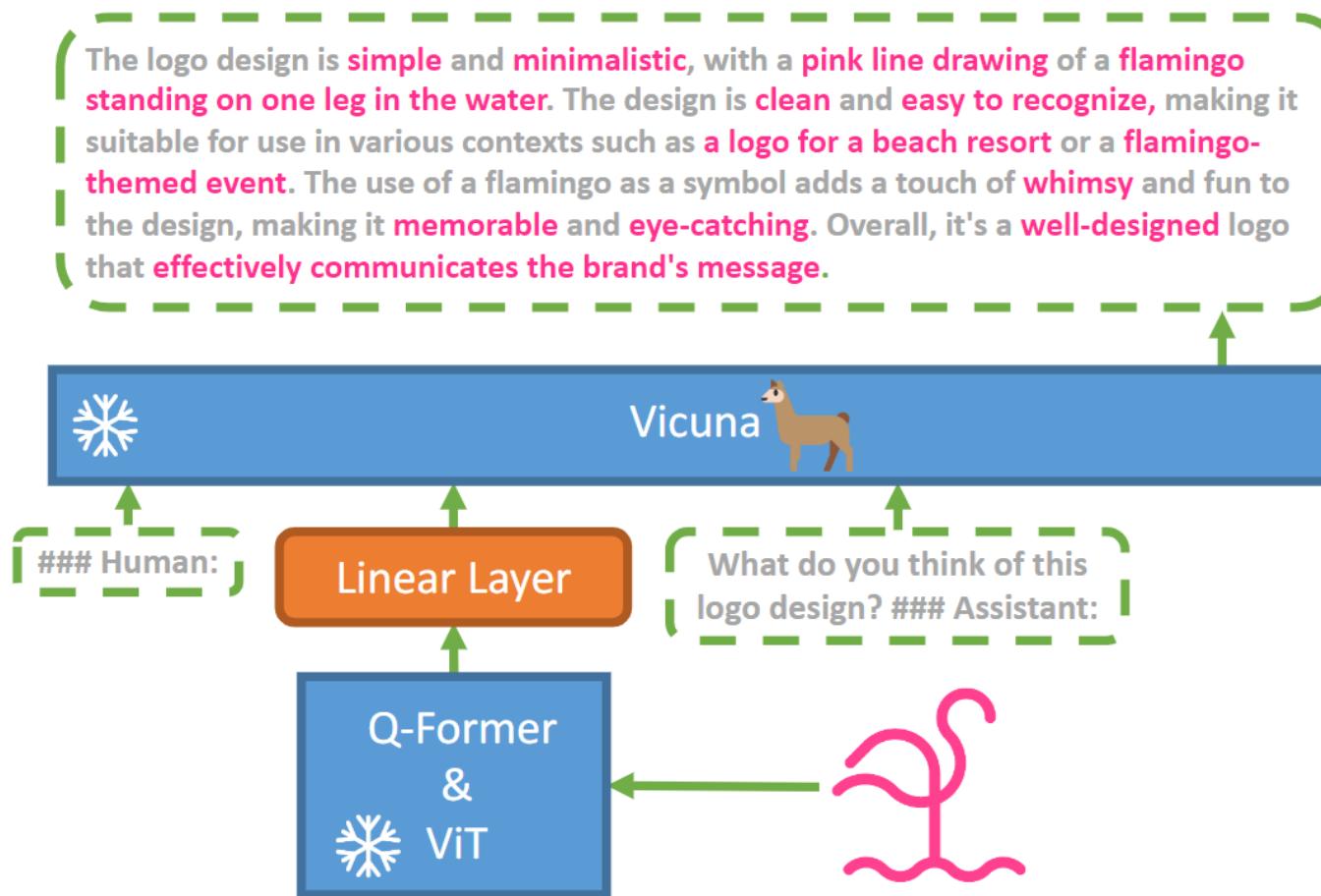
GPT-4V Sorry, I cannot answer that question as there is no context about which desert you are referring to.

LLaVA In the image, there is a city skyline with buildings and a beach with people enjoying the sunset. The scene is set in a desert, which is unusual for a beach setting.

LLaVA-1.5 There are no deserts in the image. The image features a beach with palm trees, a city skyline, and a large body of water.

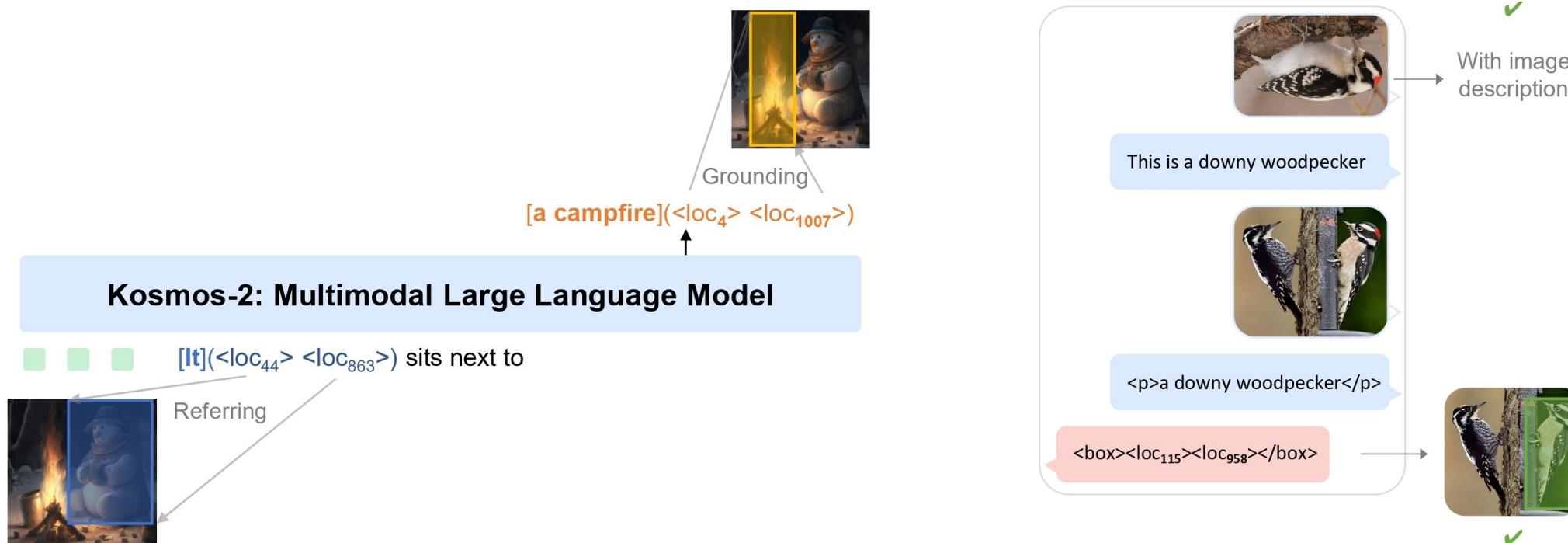
# MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models, 2023, arXiv

- ✓ GPT-4を小規模で再現することに取り組んだモデル
- ✓ モデルはVicuna(言語モデル)×ViT(視覚モデル)×Q-Former(融合)



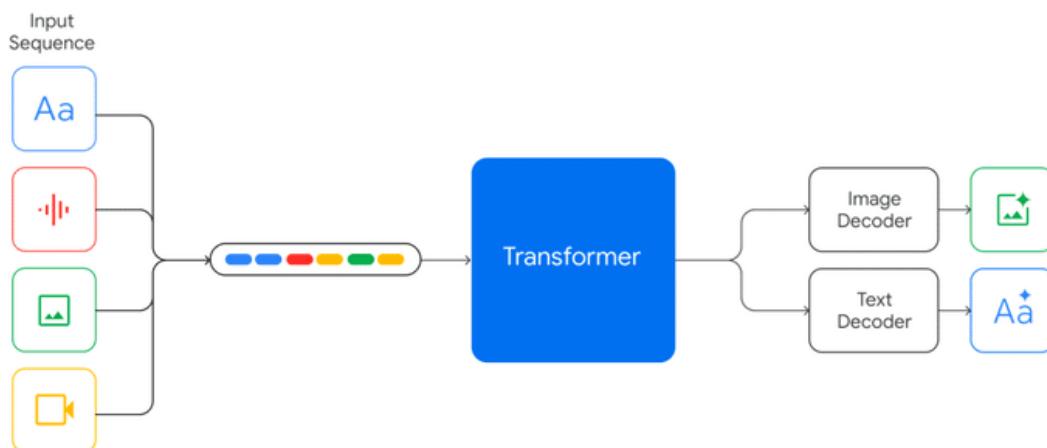
# Kosmos-2: Grounding Multimodal Large Language Models to the World, 2023, arXiv

- ✓ テキスト・画像・オブジェクトの領域を扱うモデル
- ✓ ゼロショットでGroundingが可能
- ✓ In-Context Learningを用いることも可能



# Gemini: A Family of Highly Capable Multimodal Models, 2023, arXiv

- ✓ Googleによって公開されたマルチモーダルモデル
- ✓ モデル構造は公開（？）されている
- ✓ Ultra, Pro, Nanoの3サイズが存在し、UltraはGPT-4を上回る、とも



	Gemini Ultra	Gemini Pro	GPT-4
<b>MMLU</b> Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	<b>90.04%</b> CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)
	83.7% 5-shot	71.8% 5-shot	86.4% 5-shot (reported)

# ■ 結論

- ✓ 視覚モデルとLLMの融合による大規模視覚言語モデルが登場
  - クラス識別などの物体認識からキャプショニングなどが高性能
- ✓ モデルサイズ・学習データの増大による更なる性能向上が期待

## 今後の展望

- ✓ 高度な専門知識の獲得
  - 外観検査や医用画像認識に特化したモデルは未登場
- ✓ 下流タスクへの効率的な汎化
  - 既存の大規模視覚言語モデルでは追加学習も困難
  - データセット面： Few-Shot Learning
  - リソース面： In-Context Learning