Multimodality: Language & Vision

CS-552: Modern Natural Language Processing
22.05.23

Syrielle Montariol

EPFL

nlp lab

Announcements

- M2 deadline pushed back 2 days to 5/28/2024
- No lectures next week! Work on your project!
- A3 grades are out!
  - Procedure for grading reviews outlined on Ed !

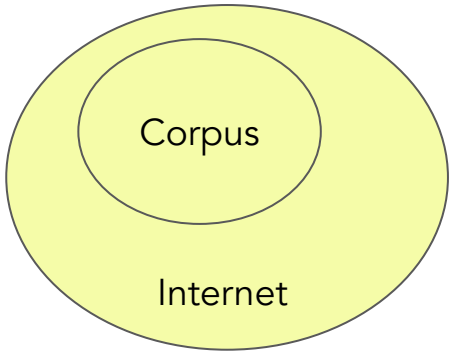
How to train a VLM?

(and why)

What is AI doing with language right now?

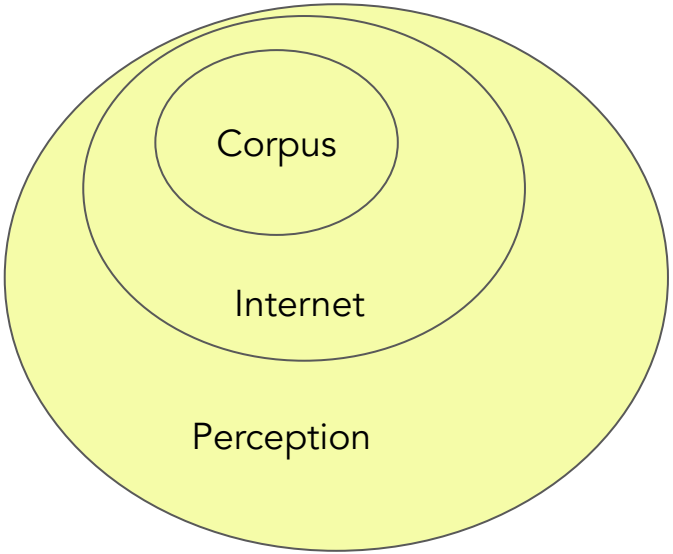
You can’t learn language from the radio.

Can you perform a task without perception ? (visual, auditory…)



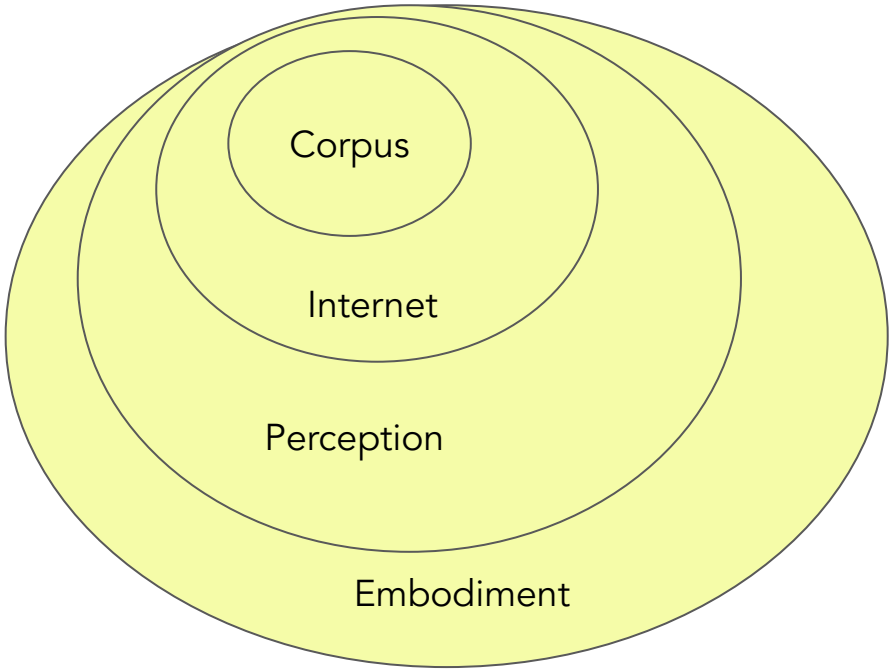
You can’t learn language from the television.

Do you have a limited action space?



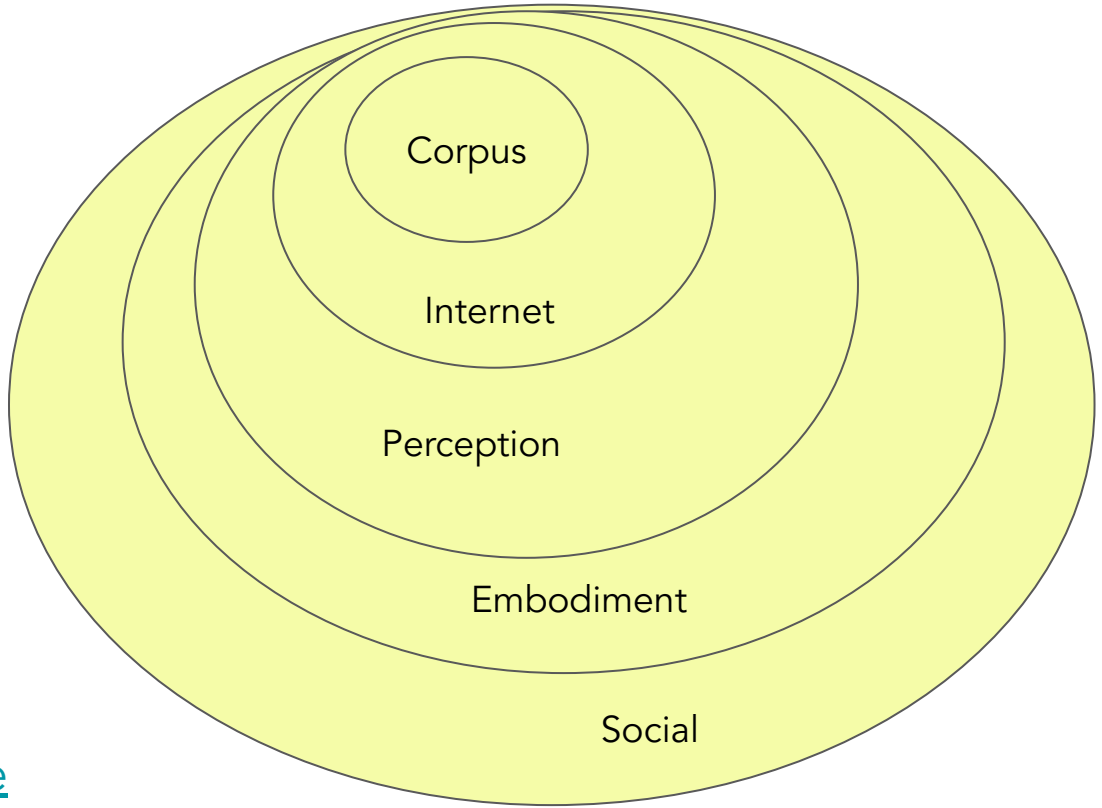
You can’t learn language by yourself.

Can you cooperate with humans to achieve your goal?



You can’t learn language…

- From the radio (Internet)
- From a television
- By yourself



[Experience Grounds Language]

Why Multimodality?

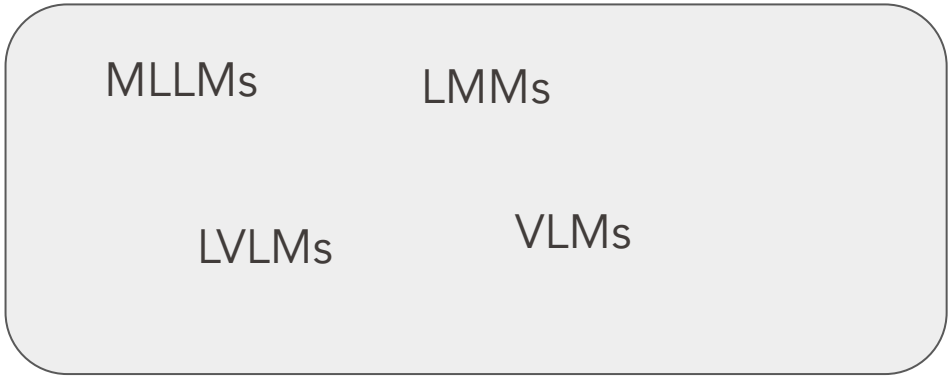
- Human learning & experience is multimodal
- Multimodal data is richer than language
- More data is usually better, and the amount of available textual data is limited

Let’s train a VLM !

What I won’t talk about

- VLM that generate or retrieve images in any way
- Diffusion models
- Any modality other than static images

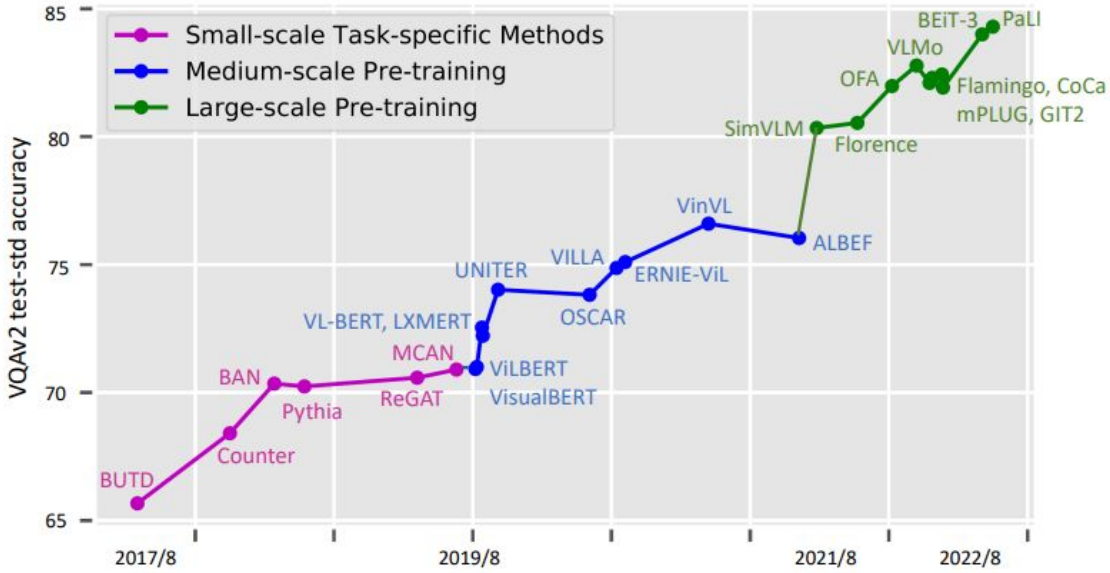
In the literature, they are called…



LLMs

Evolution of Vision and Language Models (VLMs)

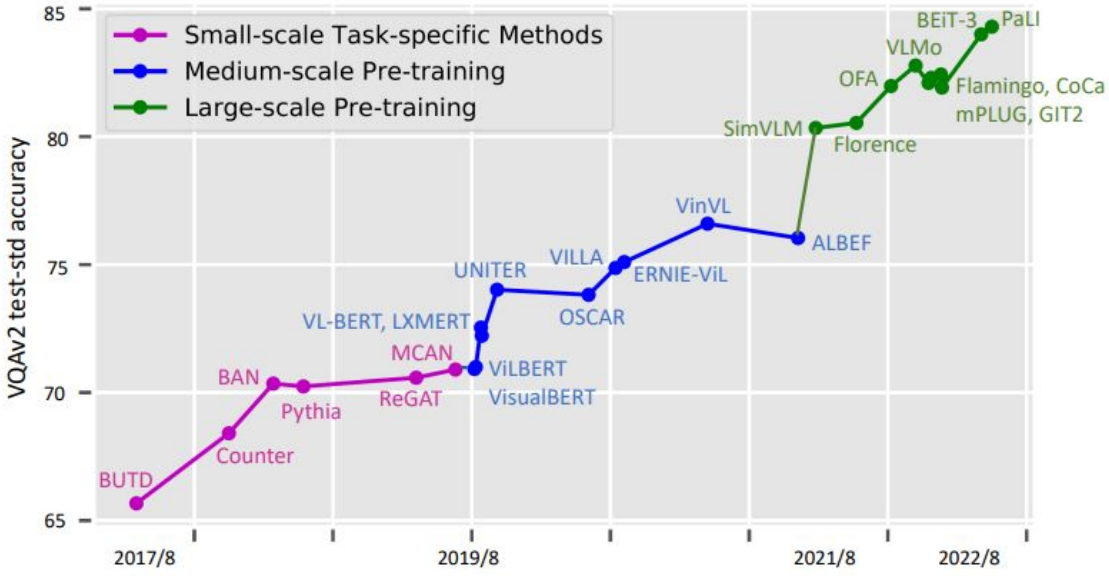
1. Small-scale, task-specific methods: ResNet & FasterRCNN, Glove & Word2Vec



Evolution of Vision and Language Models (VLMs)

1. Small-scale, task-specific methods: ResNet & FasterRCNN, Glove & Word2Vec

2. Medium-scale pre-training, since 2019 (up to 340M parameters with BERT-Large): Inspired by BERT, transformers-based multi-modal fusion.

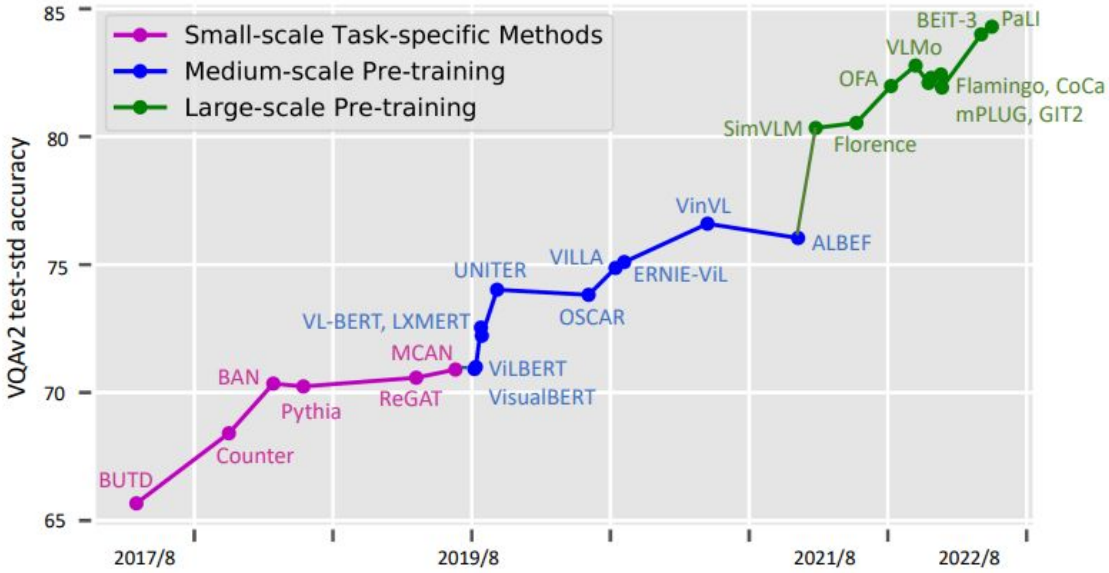


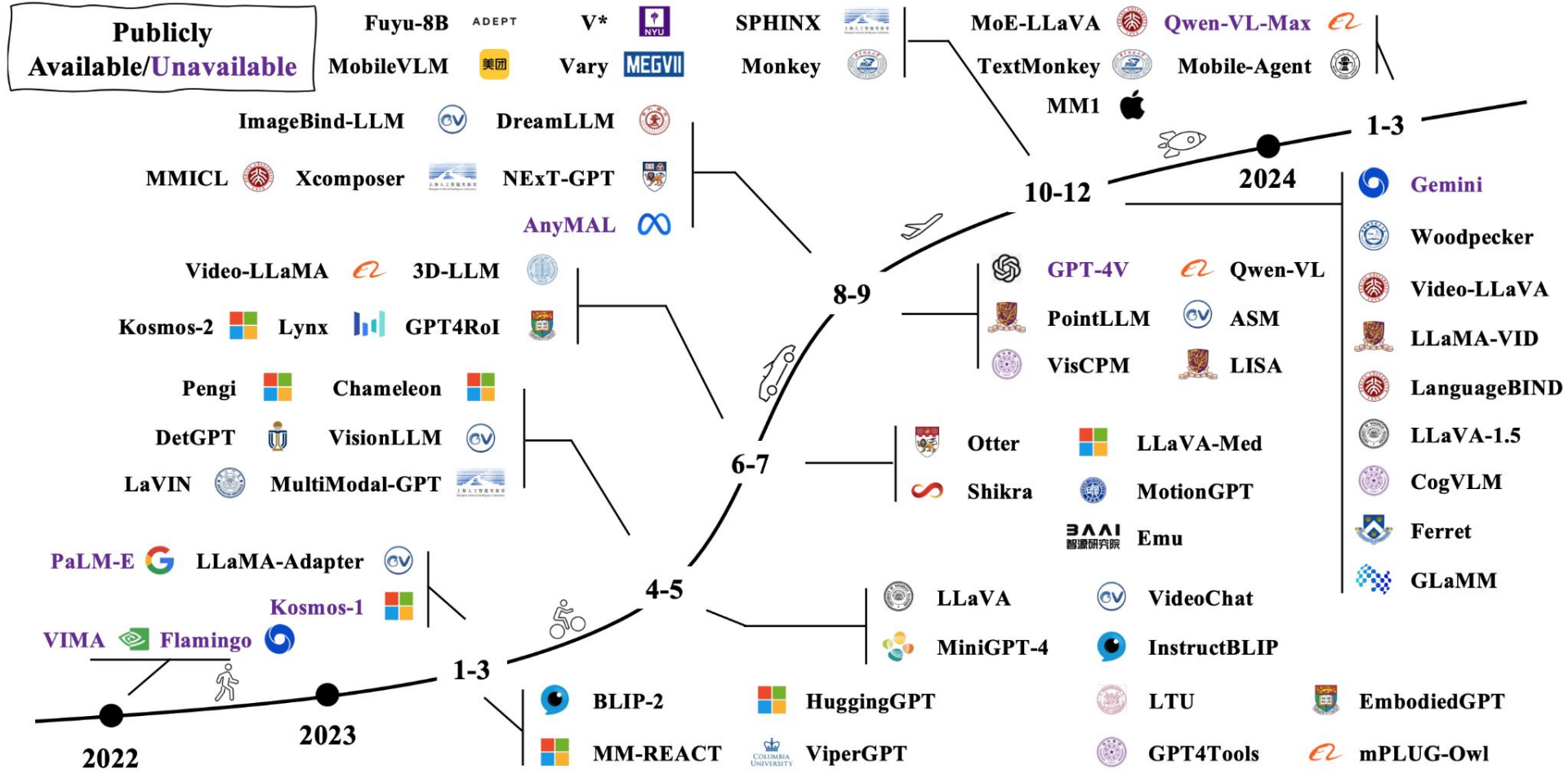
Evolution of Vision and Language Models (VLMs)

1. Small-scale, task-specific methods: ResNet & FasterRCNN, Glove & Word2Vec

2. Medium-scale pre-training, since 2019 (up to 340M parameters with BERT-Large): Inspired by BERT, transformers-based multi-modal fusion.

3. Large-scale pre-training, since 2021: starting with CLIP, then adapting pre-trained LLMs.





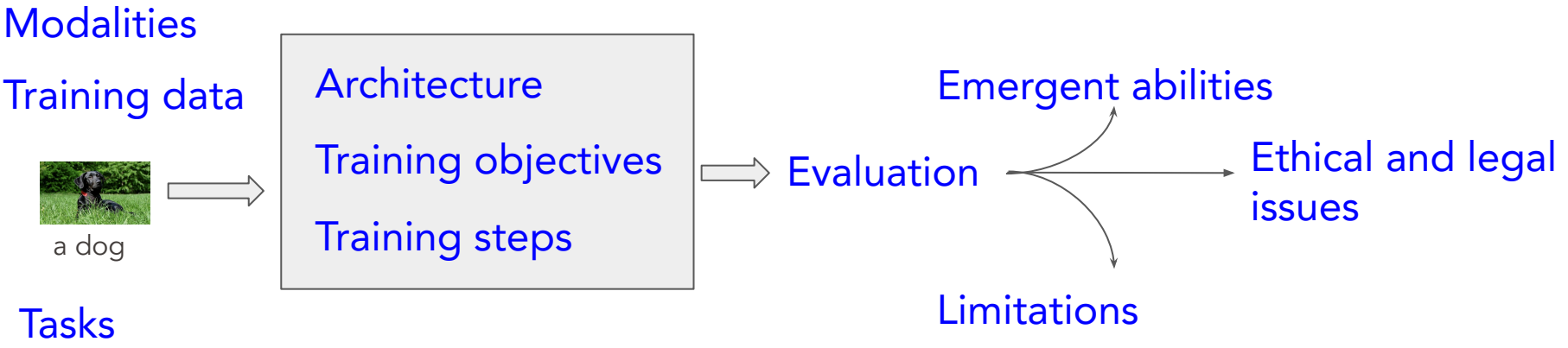
https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models

“Main” ones as of today

The ones you will see in most leaderboard:

- LLaVA family (1, 1.5, 1.6=Next): LLaMA, Vicuna, Mistral
- Flamingo and OpenFlamingo
- GPT-4v and -4o
- BLIP family: FLAN-T5
- IDEFICS 1 and 2: Mistral
- Claude family

How to train a VLM: main steps



How to train a VLM: main steps

1. Modalities
2. Tasks
3. Training objectives (contrastive, masked, autoregressive)
4. Architecture (encoder, decoder, projection)
5. Training steps
6. Training data (pre-training, fine-tuning, alignment)
7. Evaluation
8. Emergent abilities
9. Limitations: Robustness, bias, hallucinations
10. Ethical and legal issues

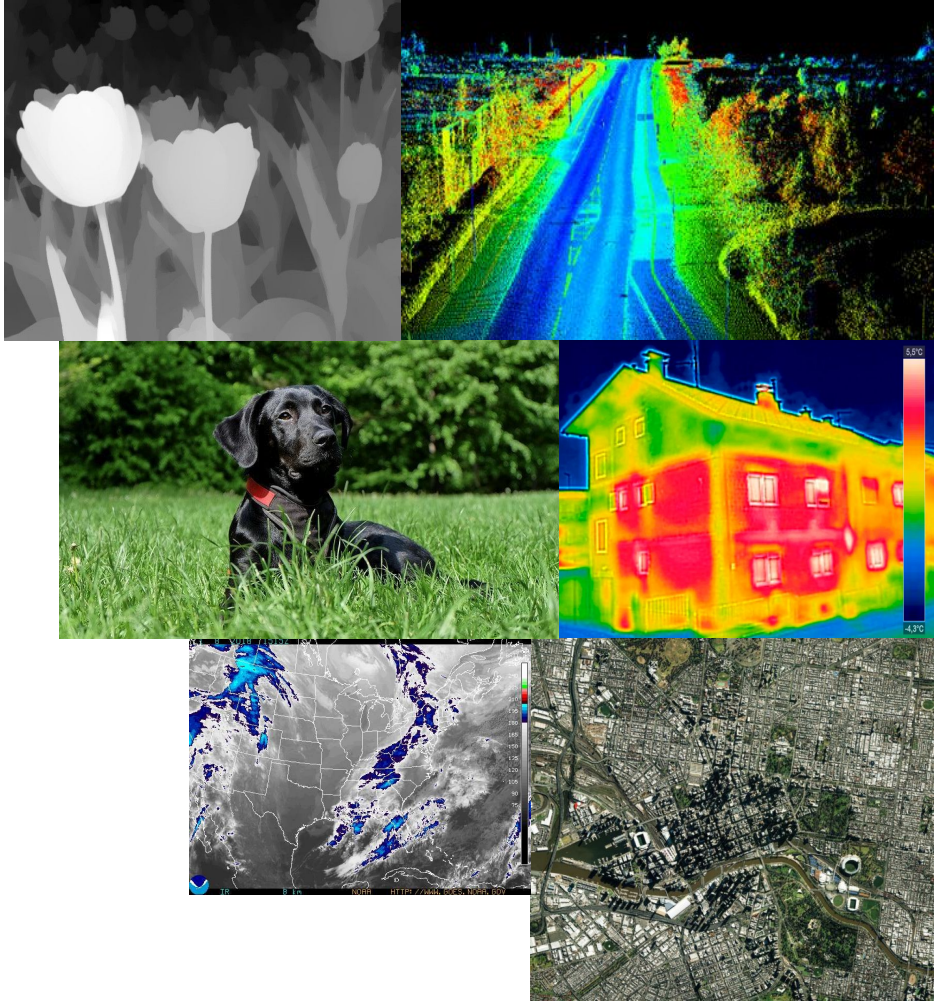
Modalities

Many types of modalities

- text
- image
- video
- knowledge graphs
- audio
- body gestures
- proteins
- physiological signals
- …

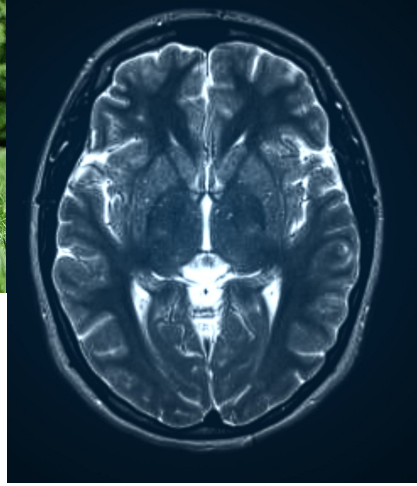
Many types of (static) images

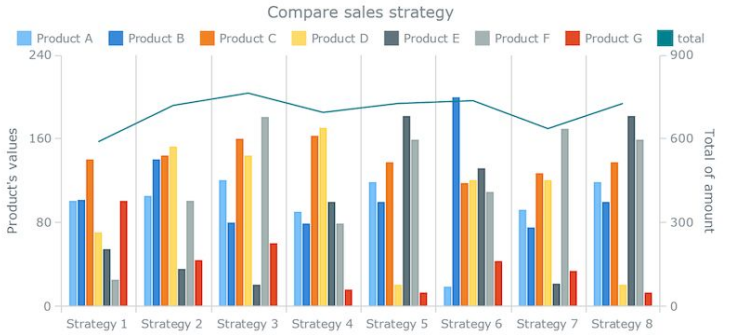
- Pictures: grayscale or color (RGB)
- 3D Point Clouds, typically obtained using depth sensors like LiDAR (autonomous driving, robotics)
- Depth Maps (3D reconstruction, scene understanding, virtual reality)
- Thermal Images (applications: night vision, surveillance, and thermal anomaly detection)
- Satellite images: visible, infrared, water vapor, NDVI (Normalized Difference Vegetation Index)
- …

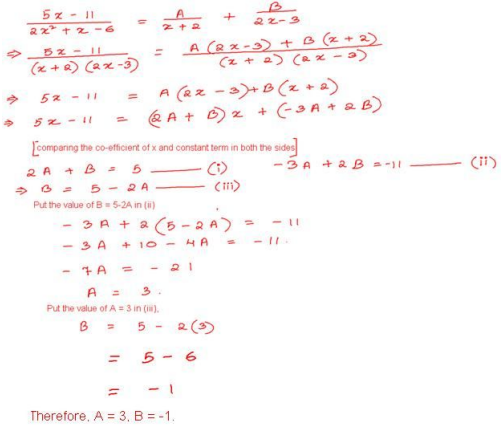


Many types of pictures (RGB)

- Photo
- Document scan
- Flowchart
- Brain scan
- …







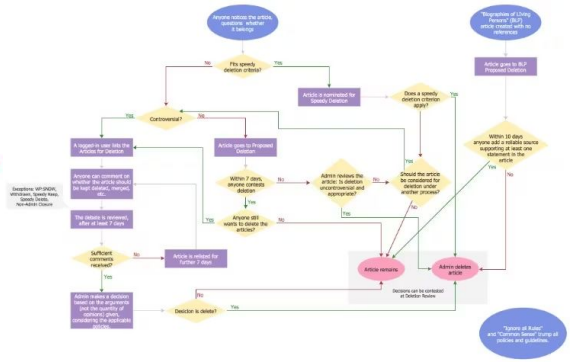
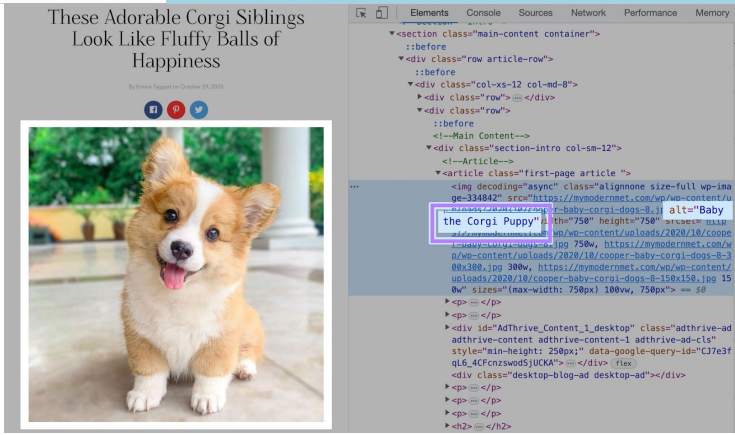


Image and text in the wild

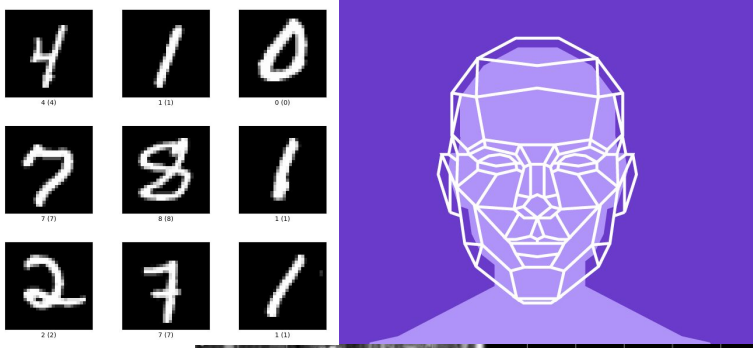
- alt-text of images
- Illustrations (process, article…)
- Titles (painting, graphic…)
- …

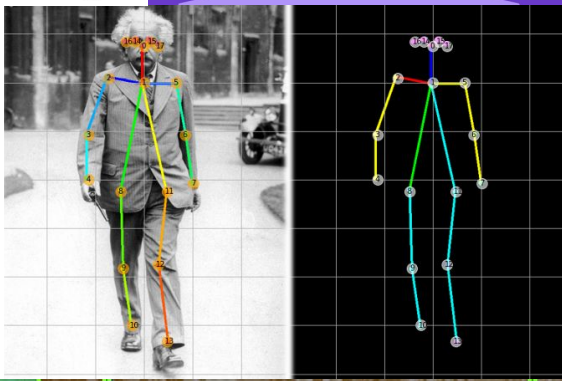


Tasks

Classical image-only tasks

- Face Recognition
- Image Classification
- Object Detection
- Depth Estimation
- Pose Estimation
- Image Segmentation
- Scene Understanding
- Action Recognition
- ...



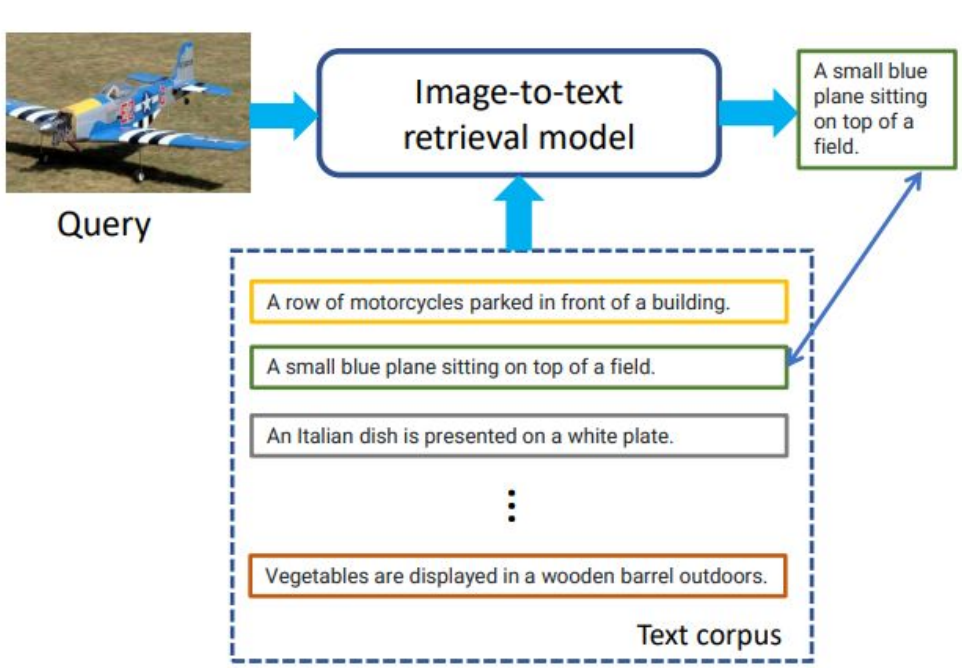


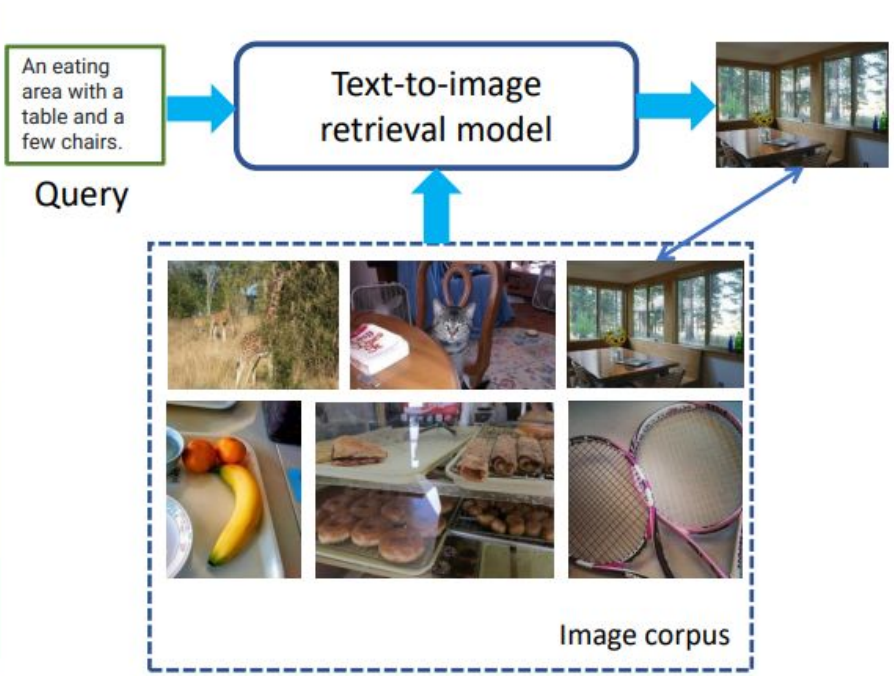
Multimodal tasks

A. Image / text retrieval

B. Image / text generation

Image / text retrieval





Generation tasks

Input / Output	Image	Text	Image + text
Image (Pixel prediction)			
Text (token prediction)			
Image + text			

Generation task examples

Input / Output	Image	Text	Image + text
Image (Pixel prediction)	Image translation (Colorization, Inpainting, Uncropping…)	Text-to-image synthesis	Text-guided image editing, phrase grounding
Text (token prediction)	Conditional text generation (Image captioning)	Question answering, summarization…	Visual question answering
Image + text		Visual dialogue	Visual dialogue

Generation task examples

Input / Output	Image	Text	Image + text
Image (Pixel prediction)	Image translation (Colorization, Inpainting, Uncropping…)	Text-to-image synthesis	Text-guided image editing, phrase grounding
Text (next token prediction)	Conditional text generation (Image captioning)	Question answering, summarization…	Visual question answering, grounding
Image + text		Visual dialogue	Visual dialogue

Image captioning

Generating a textual description of an image.

COCO style caption: “Single black dog sitting on the grass”

Narratives style caption: “The dog is black and brown. His collar is red and black. [...] The dog is laying on the grass. There are trees behind.”

Genome style caption: “Black dog”, “Red collar”.

SBU-style caption: (Noisy, web-scraped) “My labrador Jackie in the park near the house.”

Visual question answering

Answering a question about an image.

Where is the dog?

What color is the dog’s collar?

What is the animal?

Visual question answering

VQAv2:
General Visual Reasoning

Q: What brand name is on cooler to the left? A: Coca-Cola

TextVQA:
Text Reasoning

Q: What is the price of the bananas per kg? A: $11.98

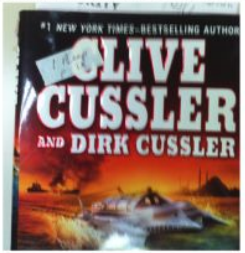
POPE:
Probing Hallucination

Q: Is there a person in the image? No

GQA:
Spatial Reasoning

Q: What animal is sitting on the sidewalk? A: Bear

VizWiz:
Unanswerable Questions



Q: What is the name of this book? A: Unanswerable

TallyQA:
Counting

Q: How many dogs are to the left of the person? One

Visual Dialogue

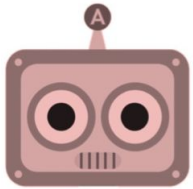
C: A dog with goggles is in a motorcycle side car.
Q: Is motorcycle moving or still?
A: It's parked
Q: What kind of dog is it?
A: Looks like beautiful pit bull mix

Q: What color is it?

Image

Dialog history

Question



Visual Dialog model

Answer

A: Light tan with white patch that runs up to bottom of his chin

https://visualdialog.org/

New trend: Image tasks with VLMs

- Object detection
- Object classification
- Image segmentation
- etc…

→ Referring and Grounding tasks

Referring tasks: understanding the content of an input region

Input:

Output:

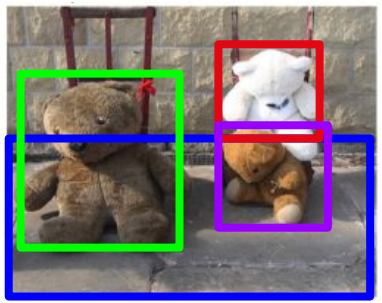
- region captioning:
  a white teddy bear
- referring expression generation (REG):
  the plushie on the top right
- region-level question answering:
  Q: What color is that objects?
  A: White.

Grounding tasks: localizing regions from a textual description

Input:

- referring expression comprehension (REC) and segmentation (RES)
  a white teddy bear
- phrase grounding:
  the plushie on the top right
- grounded captioning
  A dark brown, a light brown and a white teddy bear on a sidewalk.

Output:

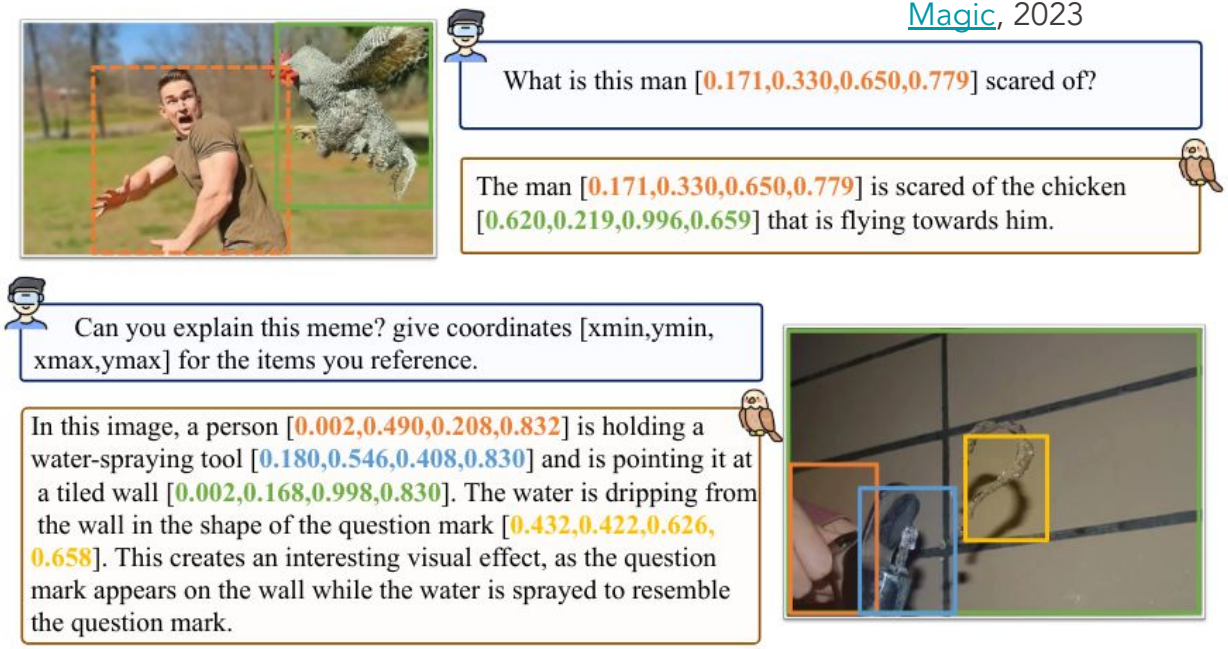


How to solve these tasks with text?

Need region-to-sequence and sequence-to-region method:

[Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic](), 2023

- regions as text: series of coordinates as numbers or special tokens
- regions as embeddings: train region encoders (more versatile, for pixel-level / segmentation)

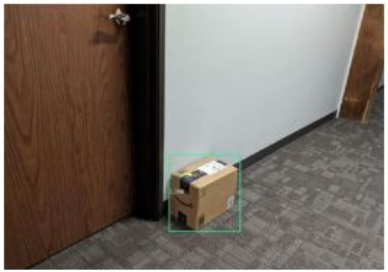


→More real-life applications?

Classical VLM tasks in real life

- Hateful memes detection
- Visual question answering for blind people
- Object detection:

Mask Wearing



Packages

Pistols

Potholes

New usages of VLMs

- Turning screenshots of webpages into code
- Explaining charts or diagrams
- Retrieving information in a scanned PDF
- …

Training Objectives

Training schemes

1. CLIP-like Contrastive learning
2. BERT-like Masked Modeling
3. GPT-like Next-token prediction

Contrastive Learning

Image-text matching objective

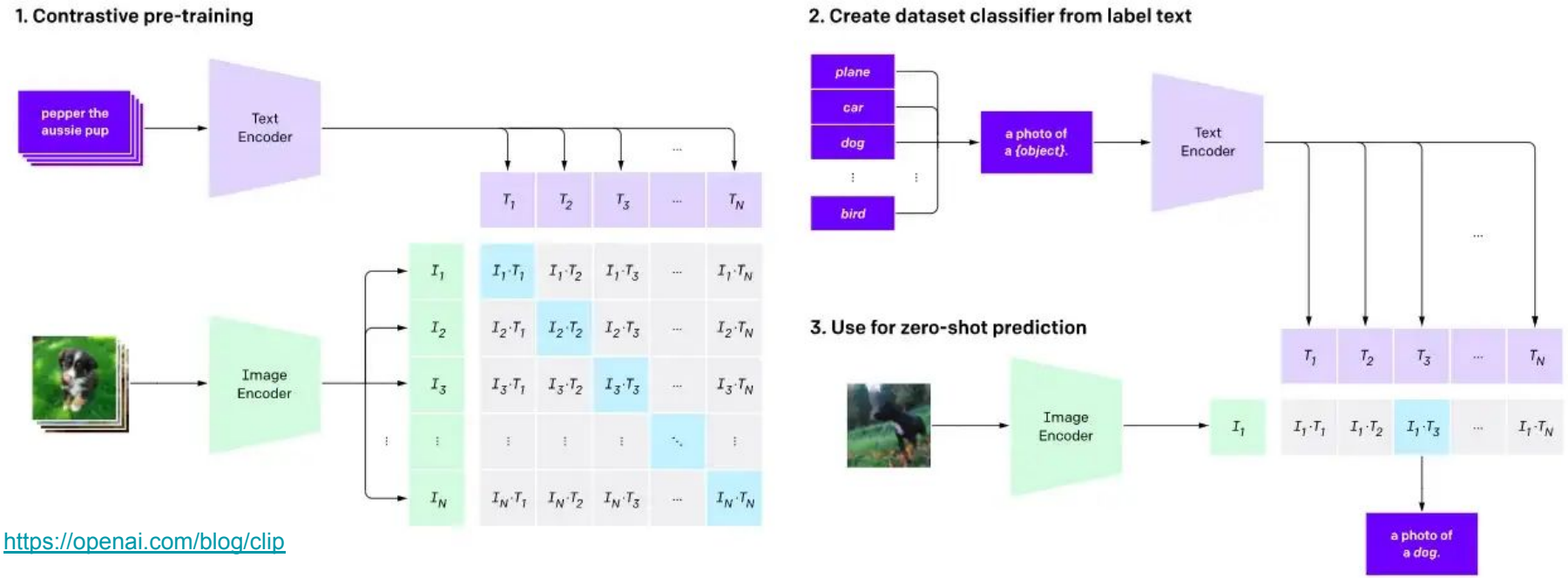
Goal: Represent the image and the text in the same space with a contrastive loss

Architecture: image and text encoder

Models: CLIP, ALIGN, DeCLIP, FLAVA

Contrastive Learning

Aligning images and texts to a joint feature space in a contrastive manner



https://openai.com/blog/clip

Training schemes

1. CLIP-style Contrastive learning
2. BERT-like Masked Modeling
3. GPT-like Next-token prediction

Masked Modeling

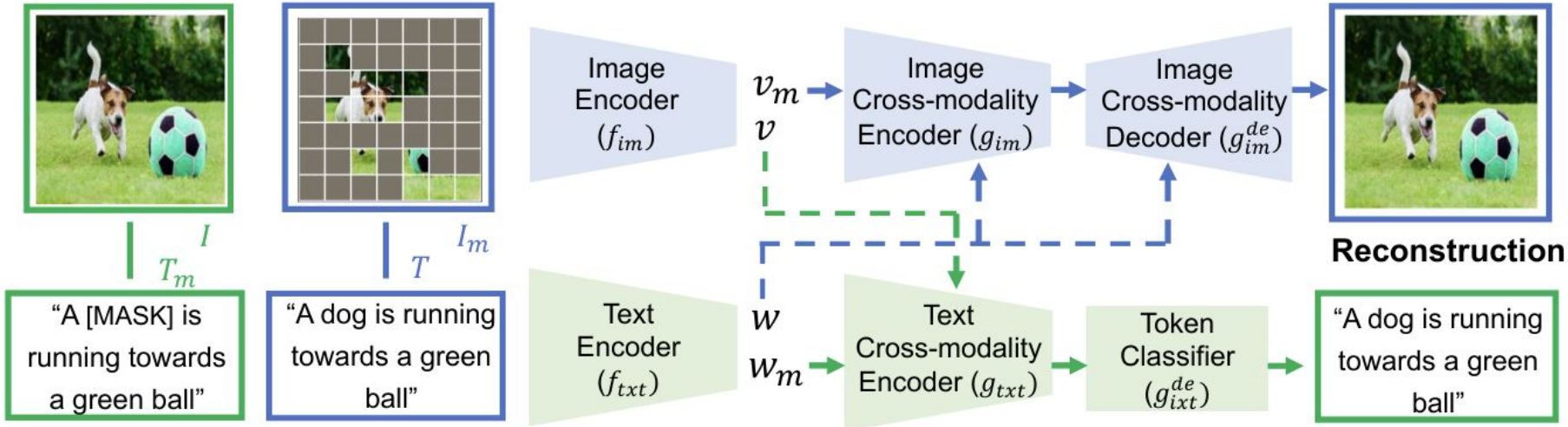
Mask-filling training objective

Goal: Align elements of an input text and regions in an associated input image.

Architecture: image and text encoder

Models: VisualBERT, ViLBERT, MaskVLM, 4M

Masked Modeling



Masked vision and language modeling for multi-modal representation learning (2023)

Training schemes

1. CLIP-style Contrastive learning
2. BERT-like Masked Modeling
3. GPT-like Next-token prediction

→ Treat all downstream tasks as language generation.

→ Let’s dive into it!

Architecture

Components of SOTA Autoregressive VLMs

An image encoder

A text decoder = an LLM

A block to merge information from both modalities = a multimodal projector

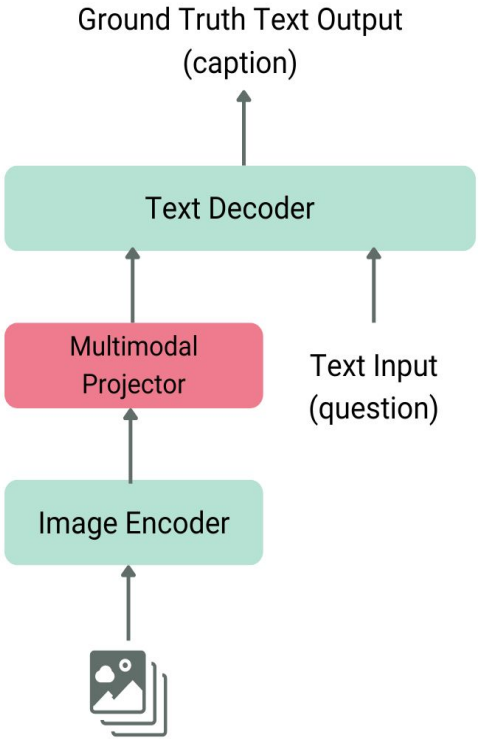
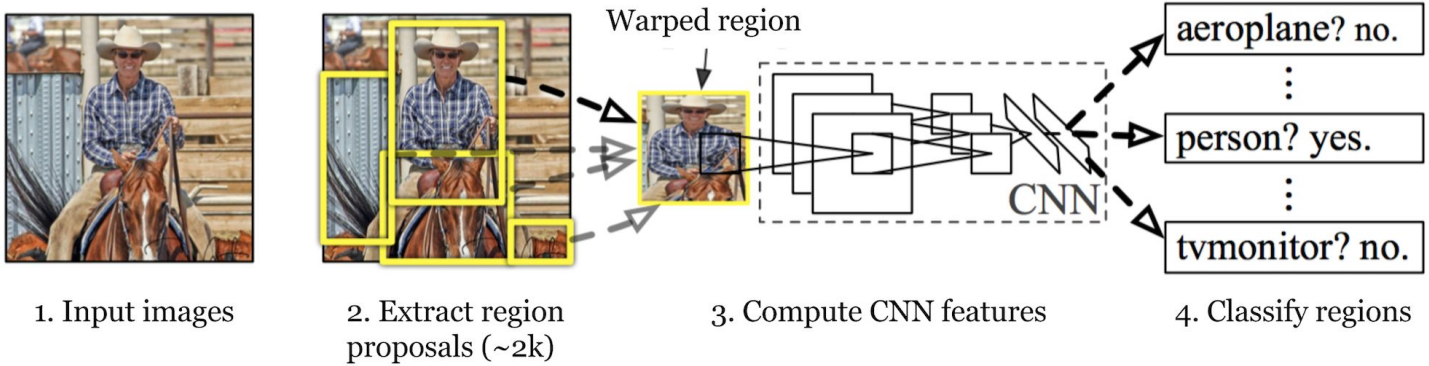


Image encoding

- Sparse features (object detector):
  - R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN
  - Expensive to annotate, need box labels



Rich feature hierarchies for accurate object detection and semantic segmentation, 2013

Image encoding

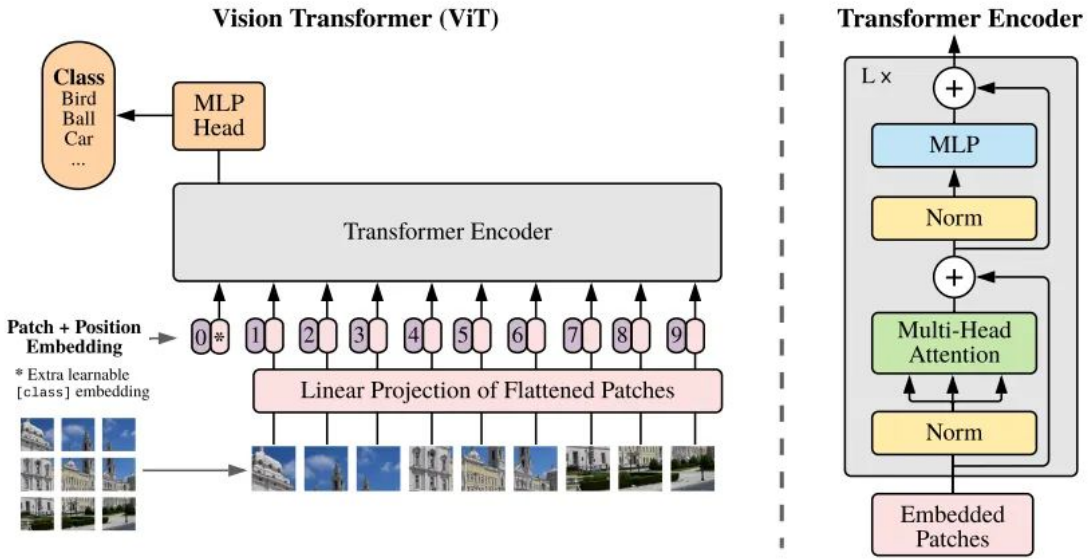
- Sparse features (object detector):
  - R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN
  - Expensive to annotate, need box labels

- Dense feature:
  - CNN: ConvNet or ConvNext layers
  - Vision transformer (ViT)

Visual transformer - ViT

Transformers cannot process grid-structured data, only sequences!

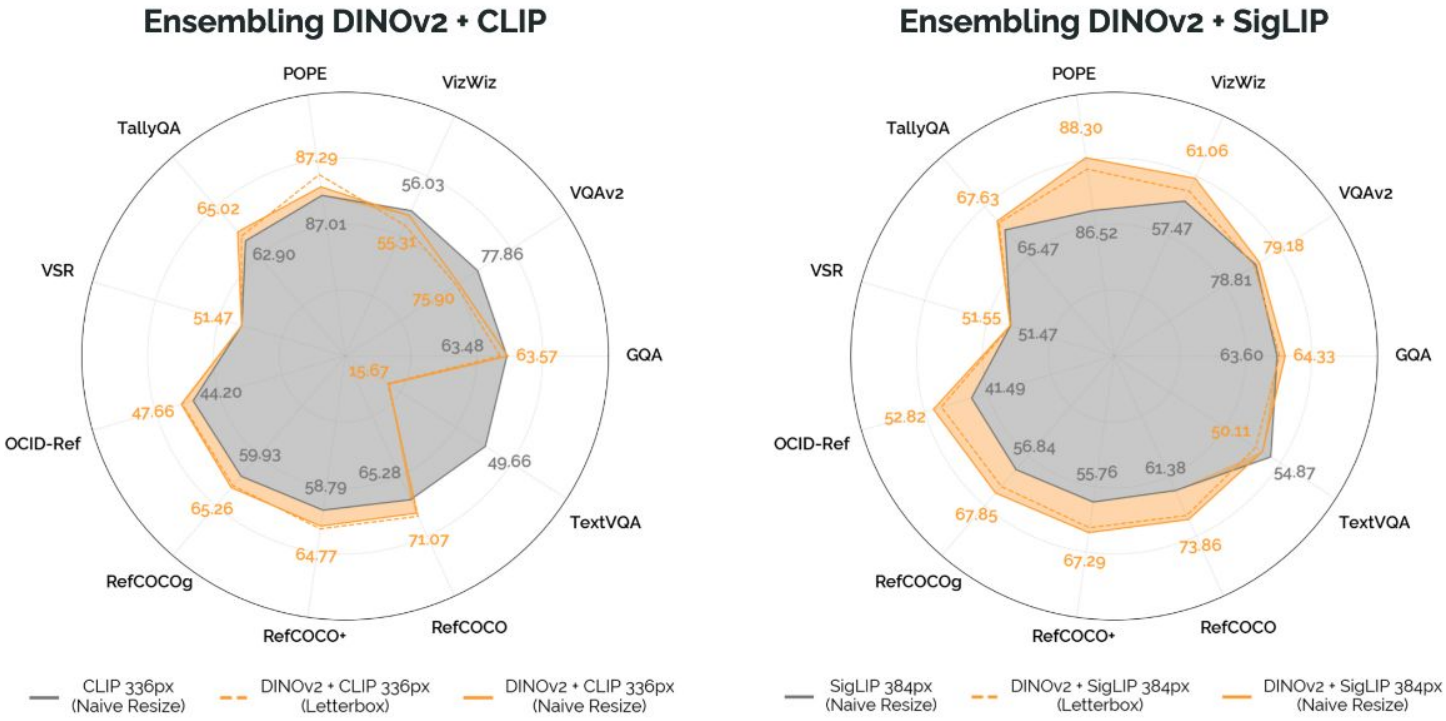
1) Transform the images into a sequence of patches (tokens)
2) “flatten” them
3) Add position encoding



ViT-based image encoders

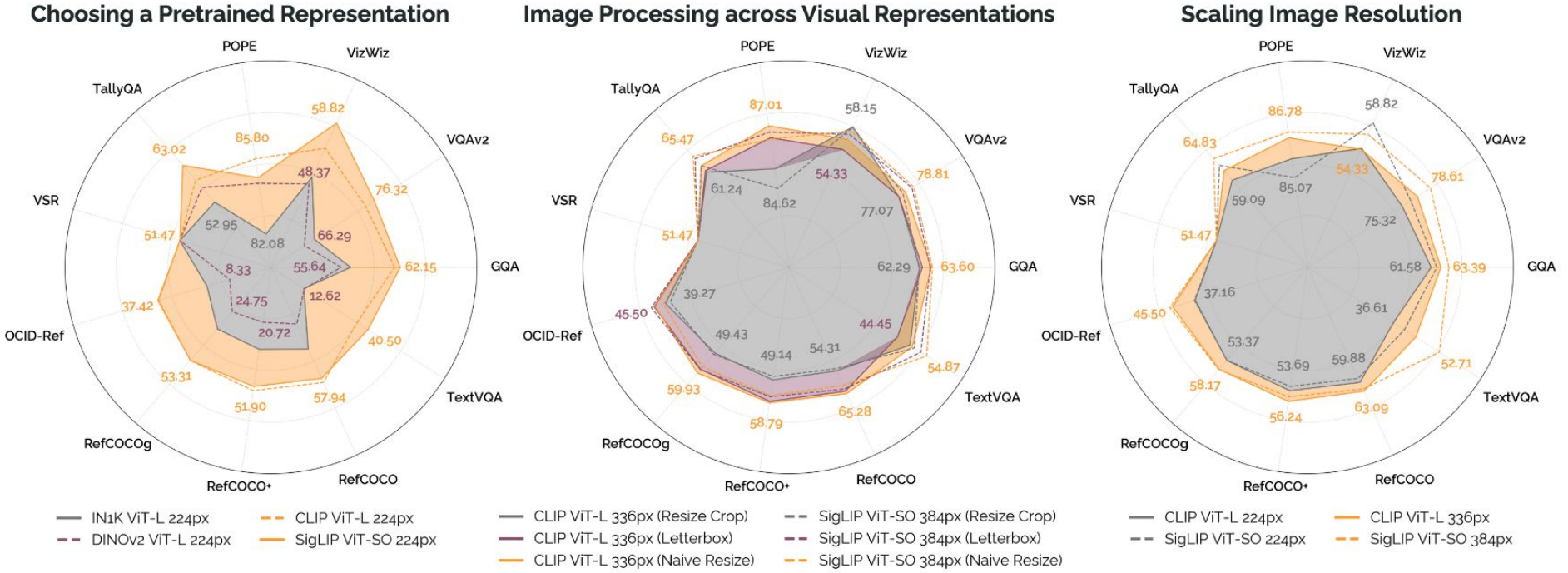
- CLIP
- EVA-CLIP = CLIP + Masked Image Modeling
- SigLIP = CLIP + sigmoid loss
- DINO (DIscriminative NOise Contrastive Learning): vision-only self-supervised
- …

Mixture of image encoders



[Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models](#) (02/2024)

What matters when choosing an image encoder?



[Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models](#) (02/2024)

What matters when choosing an image encoder?

✔ Image resolution and aspect ratio

✔ Pre-training dataset image type

× Parameter size

Visual encoder backbone	Image resolution	Avg VQA & captioning performance	ImageNet performance (encoder-only)
CLIP-ViT-H	224	57.4	83.5
EVA-CLIP-5B	224	60.2	82.1
SigLIP-SO400M	384	60.7	82.0

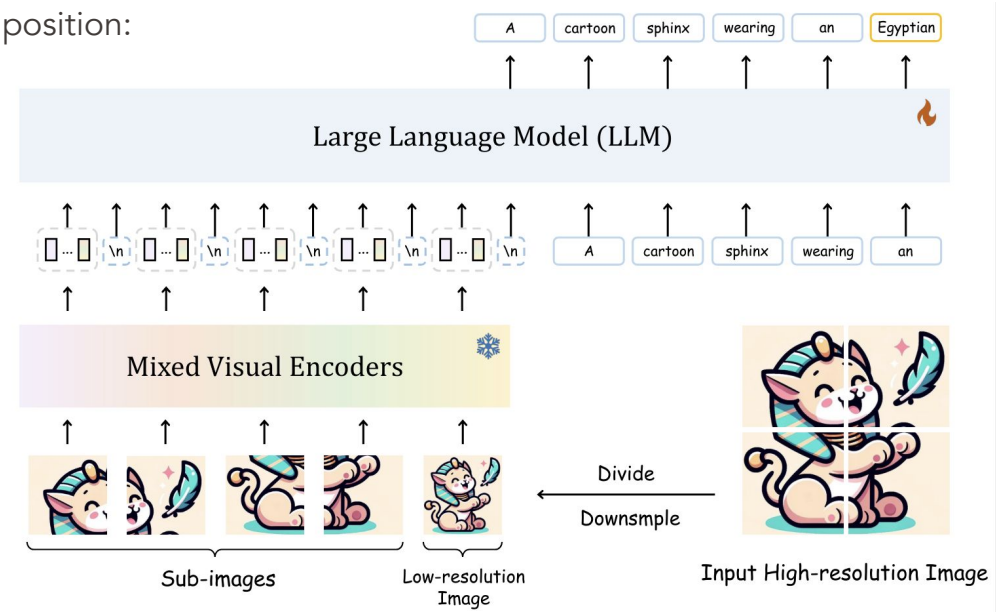
Large ViT encoders are undertrained?

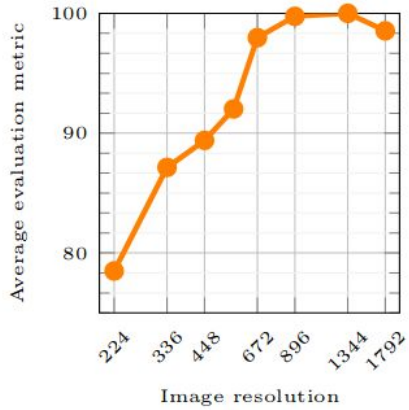
What matters when building vision-language models? (05/2024)

High-resolution image challenge

Self-attention among more than 2, 000 image tokens is computationally challenging!

→ Sub-image decomposition:





(b) Impact of image resolution on SFT performance.

[Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models](), 11/2023

How many image tokens to we want? Efficiency VS Expressiveness

Pooling into a shorter sequence (often 64 or 144):

- More compute-efficient
- Especially useful for few-shots in-context learning / long documents with interleaved images, which require very large context window
- Output dimension needs to be carefully tuned

Taking all image tokens (576 to 2890 tokens):

- capture more details of the image

Components of SOTA VLMs

1. An image encoder
2. A text decoder (= an LLM)
3. A block to merge information from both modalities

LLMs for VLMs

LLM's ability is transferred to the VLM, to some extent:

Pros:

- Better out-of-distribution generalization
- Improved contextual understanding: more coherent and contextually relevant responses.
- Expanded encyclopedic and commonsense knowledge knowledge
- Generalization ability through in-context learning
- Reasoning ability (chain of thought, generating answer rationales)
- Broad context window

Cons:

- Harder to evaluate
- Inheriting existing limitations of pretrained language models (hallucination, biases…).

Which LLM to choose?

- LLM's emergent abilities transfer to VLMs as well

  → The bigger LLM, the better

- Instruction-tuned LLMs: no consensus

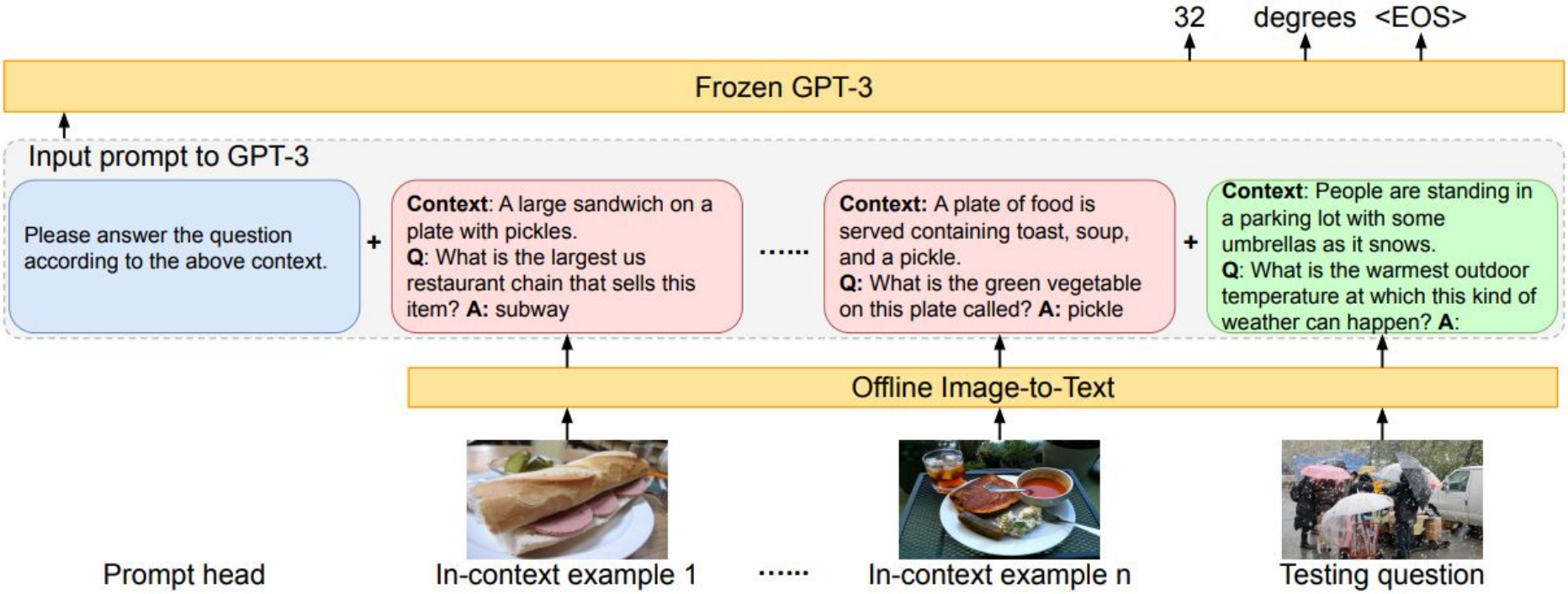
LLM backbone	Avg VQA & captioning performance	MMLU performance (LLM-only)
Llama-1-7B	62.5	35.1
Mistral-7B	67.6	60.1

What matters when building vision-language models? (05/2024)
Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models (02/2024)

Components of SOTA VLMs

1. An image encoder
2. A text decoder (= an LLM)
3. A block to merge information from both modalities
    - pre-trained textual feature extractor
    - Learnable connector

Encoding image features as text



An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA, 2021

What are the limitations of this method?

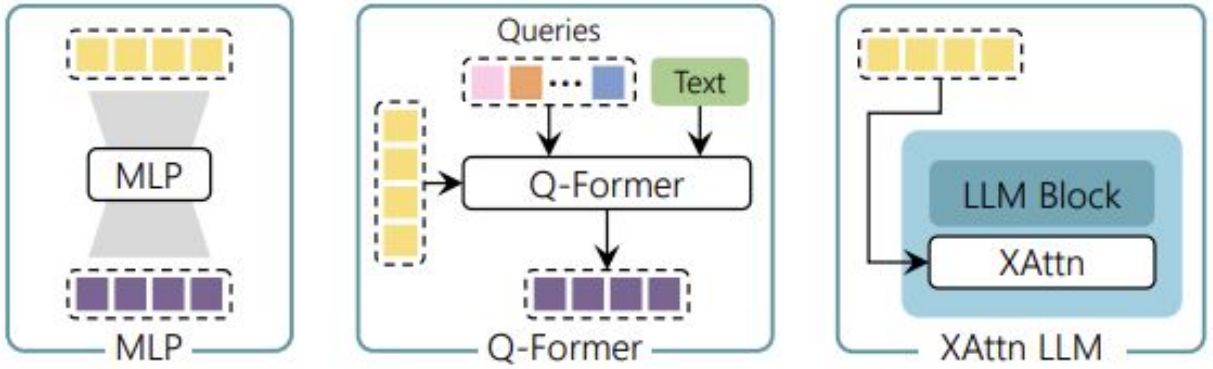
- Additional inference
- Need existing image captioning model
- Loss of information

(e) What color is the man's jacket?
Context: A man flying through the air while riding a snowboard.
Answer: black

(f) How many giraffes are there?
Context: A herd of giraffe standing next to a wooden fence.
Answer: 3

Learnable connector

- Fully autoregressive (“token-level”):
  - Linear, MLP, Perceiver (e.g. LLaVA)
  - Q-former (e.g. BLIP-2): compress visual tokens into a smaller number of representation vectors.
- Cross-attention (“feature-level”): inserts cross-attention layers in the LLM
  - Enable deep interaction and fusion between textual and visual features. (e.g. Flamingo)



What matters when choosing the best connector?

The number of visual tokens and input resolution are the most important, not the adapter itself!

No consensus on the best connector.

Cross-attention:

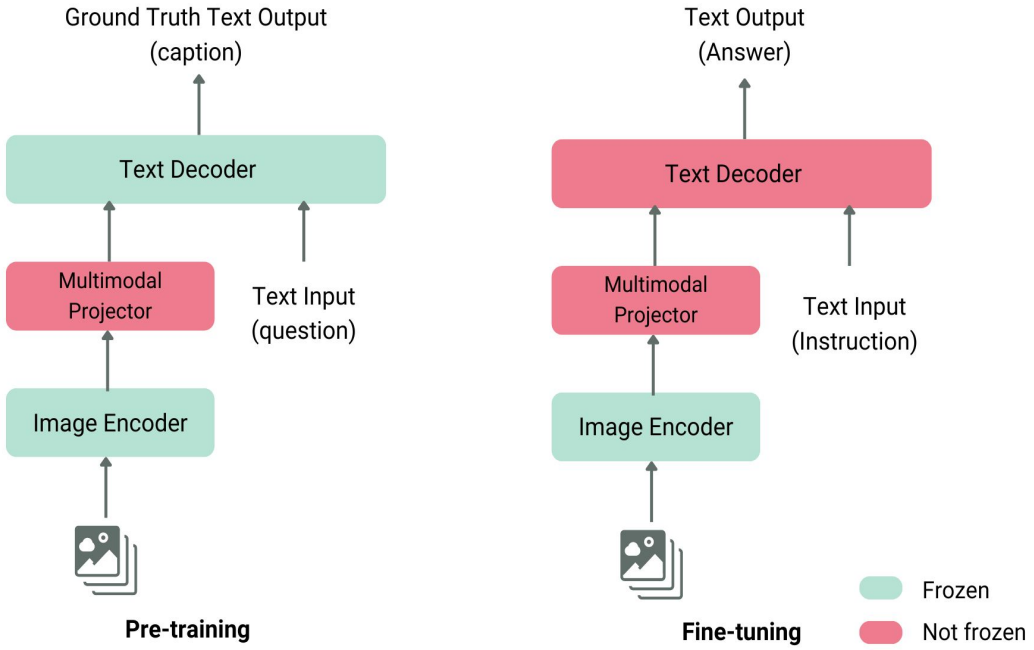
- Might require more hyper-parameter tuning (What Matters in Training a GPT4-Style Language Model with Multimodal Inputs?)
- Is more computationally expensive

Training steps

Most classical setup

LLM-like: pre-training then instruction-tuning

Note: No consensus on the positive effect of the first stage, leading it to be skipped by some works.



Many possible training frameworks!

Pre-training

1. Initialize projection module (e.g. LLaVA): image-text pairs
2. Fine-tune projection module (and jointly ViT, and LLM): Very large, usually not-too-curated train set, interleaved image-text data and image-text pairs
3. Further fine-tuning projection module (and ViT, an LLM) with higher resolution / more complex images-text pairs, more curated.

Fine-tuning

4. Instruction-tuning
5. Alignment / chat

→ Each training stage has different goals, data selection and filtering.

Pre-training 1

Data: large-scale image-text pairs
Goal:

- Align text and image modalities
- Provide world knowledge

Input: <image>
Output: {caption}

Pre-training 2

Data: large-scale interleaved text-image data
Goal:

- Provide world knowledge
- Provide multi-image understanding ability
- Retain language modeling performance

Input: <text> <image> <text> <image> <text> <image>
Output: <text>

Why interleaved text-image ?

Training data	Multimodal benchmarks performance		MMLU perf (text-only)
	0-shots	4-shots	
(Llama-2)	-	-	**46.0**
COYO	51.1	50.3	28.8 (-17.2)
MMCA-pairs	46.4	44.5	32.4 (-13.6)
MMC4	**68.7**	**70.9**	40.7 (-5.3)

VILA: On Pre-training for Visual Language Models, 12/2023

Instruction-tuning

Data: multimodal instruction-response pairs
Goal:

- Provide ability to follow instructions
- Generalize to unseen tasks

Quality matters a lot ! (Same as LLM instruction-tuning)

- Diversity of prompt / instruction, to improve generalization
- Diversity of tasks
- Complexity of tasks (reasoning, fine-grained spatial annotations…)

Instruction: <instruction>
Input: {<image>, <text>}
Response: <text>

Text-only Instruction-tuning

Data: instruction-response pairs
Goal:

- Retain instruction-following ability from LLM
- Improve safety, reasoning, and any other ability absent from multimodal instruction-tuning datasets

Instruction: <instruction>
Input: <text>
Response: <text>

Why text-only Instructions too?

Pre-training data	Instruction-tuning data	Multimodal benchmarks performance		MMLU perf (text-only)
		0-shots	4-shots	
Llama-2	Text	-	-	46.0 / 51.2
COYO	Multimodal	51.1	50.3	28.8 (-17.2)
MMCA-pairs	Multimodal	46.4	44.5	32.4 (-13.6)
MMC4	Multimodal	**68.7**	**70.9**	40.7 (-5.3)
MMC4	Multimodal + Text	**71.0**	**72.1**	51.4 (+0.2)

VILA: On Pre-training for Visual Language Models, 12/23

Alignment-tuning

Data: Preference data
Task: Preference learning (RLHF/RLAIF, DPO, …)

Goal:

- Align to human preference
- Reduce hallucination

Should the LLM be fine-tuned?

Cons:

- Computationally expensive
- Degrades text modeling ability if not careful

Pros:

- Higher performance, in particular in-context learning ability

(LoRA or full fine-tuning)

Should the image encoder be fine-tuned?

Cons:

- Computationally expensive.
- Risk of forgetting and damage to the general visual representation (shown by several papers, but mostly for full fine-tuning, and without pooling).

Should the image encoder be fine-tuned?

Cons:

- Computationally expensive.
- Risk of forgetting and damage to the general visual representation (shown by several papers, but mostly for full fine-tuning, and without perceiver).

Pros:

- Better alignment of dense image features and text features
- Allows handling images in their original aspect ratio and size (since image encoder are usually trained on square images with fixed resolution)
    - Can reduce memory and speedup training / inference, for smaller images
    - Enable more expressiveness, e.g. for images with text.

Training data

Pre-training data

1. Captioning datasets, human-annotated, high alignment: Conceptual Captions, etc
2. Images with alt-text, web-scraped, medium alignment: CC3M, CC12M, COYO-700M, LAION-5B
3. Interleaved I-T sequences, web-scraped, low alignment: MMC4(-Core), WebLI, OBELICS

Unimodal Image filtering

- Remove faces
- Remove harmful content
- Remove extreme aspect ratio and size
- Check presence of known objects
- pHash value to remove images overlapped with public datasets such as ImageNet and MS-COCO.

Unimodal Text Filtering

- Length
- Language
- Presence of spatial relation (e.g. “on”, “under”)
- Remove samples with high perplexity score

Multimodal Filtering

- CLIP similarity between text and image

Generating synthetic pre-training data

Leverage existing tools (often close-source)

- Generate captions for existing images using VLMs
- Generate text + image (LLM to generate image-generating prompts, Diffusion model to generate the images)

Pros:

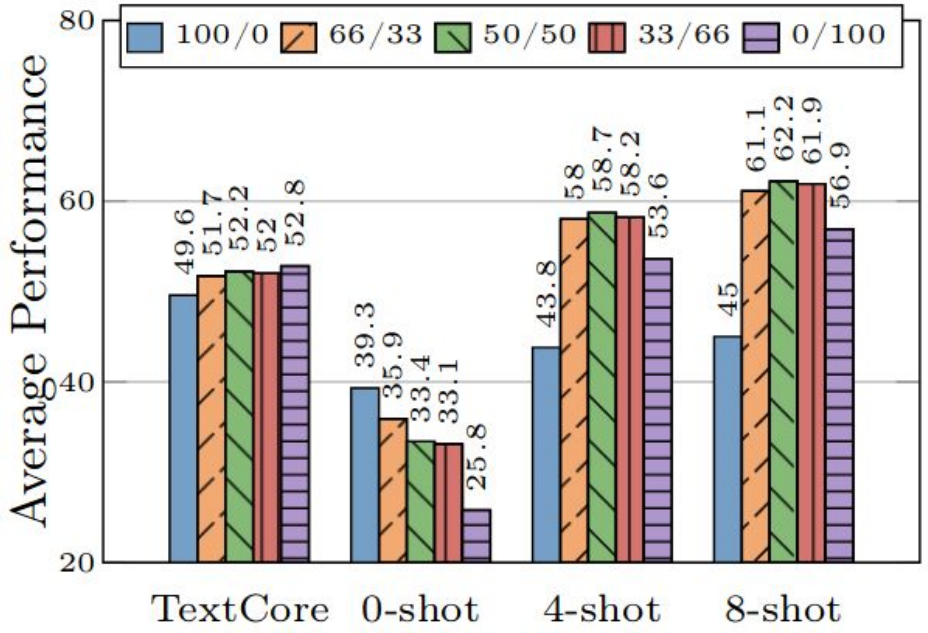
- More high-quality than web-scraped alt-text
- More fine-grained

Cons:

- More costly

Pre-training data mix: caption or interleaved?

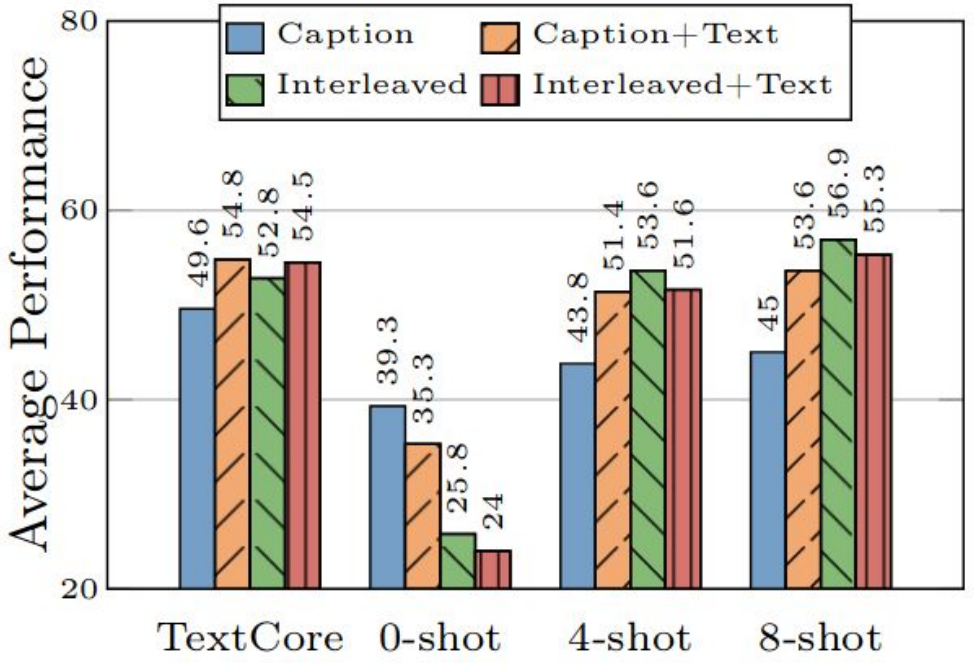
- Interleaved data increases few-shot and text-only performance
- Captioning data increases zero-shot performance



MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training, 03/2024

Pre-training data mix: adding text-only data?

Text-only data helps with few-shot and text-only performance.

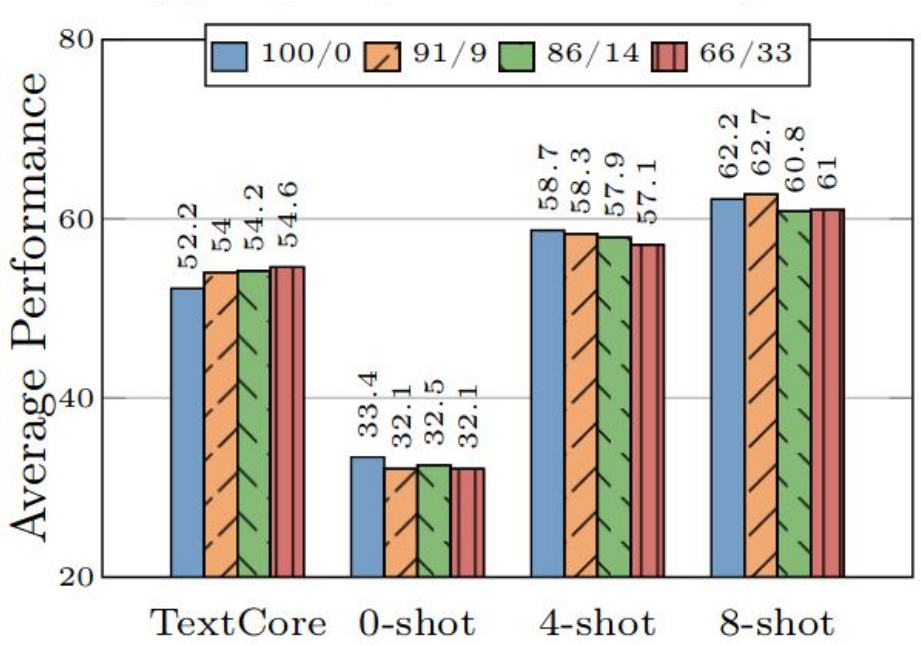


MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training, 03/2024

Pre-training data optimal multimodal & text mix:

Caption / interleaved / text ratio:

5:5:1



MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training, 03/2024

Multimodal instruction-tuning data

How to generate them?

1. Data adaptation:
    - Turn task-specific, high-quality datasets into instruction pairs (like what we did for LLMs)
    - Optionally using LLMs to generate diversified instruction formats or complexify the task.

2. Self-instruction:
    - Use LLMs or VLMs to generate instruction-output pairs using demonstrations
    - Can target specific abilities, e.g. grounding

Note: Multimodal chat and alignment data are also usually generated synthetically.

Summary: how to choose training data and why?

- Pre-training data:
  - Aligned text-image pairs: To learn alignment between modalities
  - Interleaved image-text (long) documents:
    - To learn multi-images processing
    - To unlock in-context learning ability
    - To retain text modeling ability

Summary: how to choose training data and why?

- Pre-training data:
  - Aligned text-image pairs
  - Interleaved image-text (long) documents
- Instruction-tuning data:
  - Text-images instructions: to learn multimodal tasks
  - Text-only instructions:
    - To retain language modeling abilities
    - To learn useful abilities for multimodal tasks, even if it's text-only (e.g. arithmetic calculations). Especially for data that don't exist in multimodal version yet but are crucial such as safety / debiasing fine-tuning.
    - To improve generalization to new tasks (increase diversity)

Evaluation

Open VLM Leaderboard

https://rank.opencompass.org.cn/leaderboard-multimodal

	Method	Eval Time	Params	Language Model	Vision Model	Avg. Score
1	GPT-4o, 20240513 Verified OpenAI	2024/05/15				66.3
2	GPT-4v, 20240409 Verified OpenAI	2024/05/15				63.5
3	InternVL-Chat-V1.5 Verified Shanghai AI Laboratory & Sensetime & Tsinghua...	2024/04/17	26B	InternLM2-20B	InternViT-6B	61.7
4	GLM-4v Verified Zhipu AI	2024/05/20				60.8
5	Step-1V Verified StepFun	2024/03/15				59.5
6	MiniCPM-Llama3-V2.5 Verified OpenBMB	2024/05/21	8B	Llama3-8B	SigLip-400M	58.8
7	Qwen-VL-Max Verified Alibaba	2024/02/02		QwenLM		58.3
8	InternLM-XComposer2-VL Verified Shanghai AI Lab	2024/01/26	7B	InternLM2	CLIP ViT-L/14	57.1
9	GPT-4v, 20231106 Verified OpenAI	2023/12/23				57.1
10	GeminiProVision Verified Google	2023/12/23				56.1
11	LLaVA-Next-Yi-34B Verified University of Wisconsin–Madison	2024/03/25	34.8B	Yi-34B	CLIP ViT-L/14	55
12	Claude-3V Opus Verified Anthropic	2024/03/28		Claude-3		54.4

Open VLM Leaderboard

Using https://github.com/open-compass/VLMEvalKit

→ Popular VLM benchmarks

Dataset	Task	Dataset	Task
MMBench Series	MCQ	MMStar	MCQ
MME	Y/N	SEEDBench_IMG	MCQ
MM-Vet	VQA	MMMU	MCQ
MathVista	VQA	ScienceQA_IMG	MCQ
COCO Caption	Caption	HallusionBench	Y/N
OCRVQA	VQA	TextVQA	VQA
ChartQA	VQA	AI2D	MCQ
LLaVABench	VQA	DocVQA	VQA
InfoVQA	VQA	OCRBench	VQA
RealWorldQA	MCQ	POPE	Y/N
Core-MM	VQA		

MMBench

- 3000 single-choice questions over 20 different skills (OCR, object localization …)
- Also evaluates model robustness to MCQ choice order

Dataset	Task	Dataset	Task
MMBench Series	MCQ	MMStar	MCQ
MME	Y/N	SEEDBench_IMG	MCQ
MM-Vet	VQA	MMMU	MCQ
MathVista	VQA	ScienceQA	MCQ
COCO Caption	Caption	HallusionBench	Y/N
OCRVQA	VQA	TextVQA	VQA
ChartQA	VQA	AI2D	MCQ (diagrams)
LLaVABench	VQA	DocVQA	VQA
InfoVQA	VQA	OCRBench	VQA (documents)
RealWorldQA	MCQ	POPE	Y/N
Core-MM	VQA		

MMMU

Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI (MMMU):

- 11.5K VQA
- College-level knowledge and reasoning across many disciplines

Dataset	Task	Dataset	Task
MMBench Series	MCQ	MMStar	MCQ
MME	Y/N	SEEDBench_IMG	MCQ
MM-Vet	VQA	MMMU	MCQ
MathVista	VQA	ScienceQA_IMG	MCQ
COCO Caption	Caption	HallusionBench	Y/N
OCRVQA	VQA	TextVQA	VQA
ChartQA	VQA	AI2D	MCQ
LLaVABench	VQA	DocVQA	VQA
InfoVQA	VQA	OCRBench	VQA
RealWorldQA	MCQ	POPE	Y/N
Core-MM	VQA		

LLaVA-Bench (In-the-Wild)

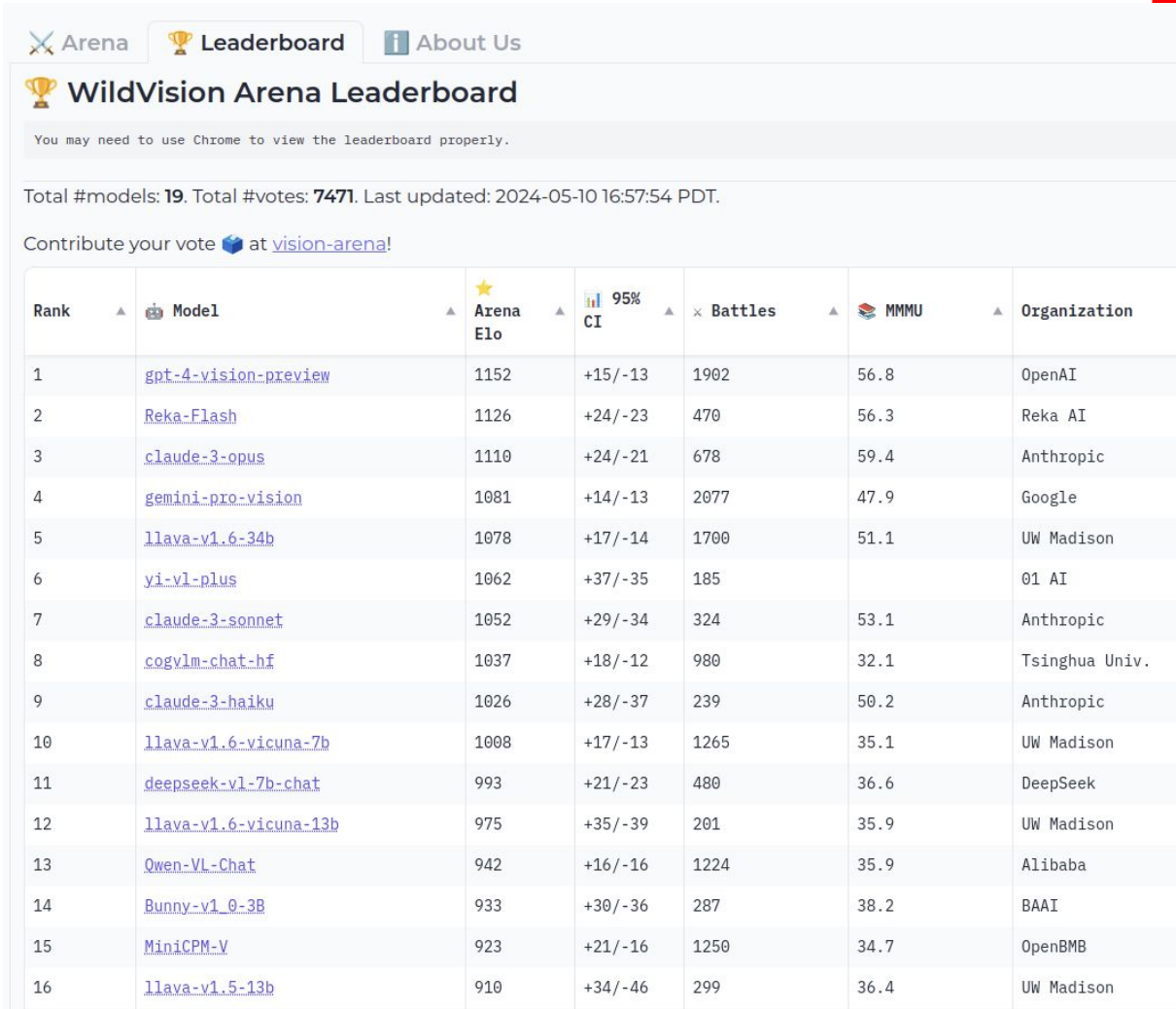
- Tests generalization ability and robustness to prompts
- 24 images with 60 questions
- Tasks: conversation (simple QA), detailed description, and complex reasoning.
- Evaluation: text-only GPT-4 evaluator rates a reference answer and the answer generated by the candidate model.

Dataset	Task	Dataset	Task
MMBench Series	MCQ	MMStar	MCQ
MME	Y/N	SEEDBench_IMG	MCQ
MM-Vet	VQA	MMMU	MCQ
MathVista	VQA	ScienceQA_IMG	MCQ
COCO Caption	Caption	HallusionBench	Y/N
OCRVQA	VQA	TextVQA	VQA
ChartQA	VQA	AI2D	MCQ
LLaVABench	VQA	DocVQA	VQA
InfoVQA	VQA	OCRBench	VQA
RealWorldQA	MCQ	POPE	Y/N
Core-MM	VQA		

Vision-Arena

https://huggingface.co/spaces/WildVision/vision-arena

→Based on anonymous voting of model outputs



⚔️ Arena | 🏆 Leaderboard | ℹ️ About Us

🏆 WildVision Arena Leaderboard

You may need to use Chrome to view the leaderboard properly.

Total #models: **19**. Total #votes: **7471**. Last updated: 2024-05-10 16:57:54 PDT.

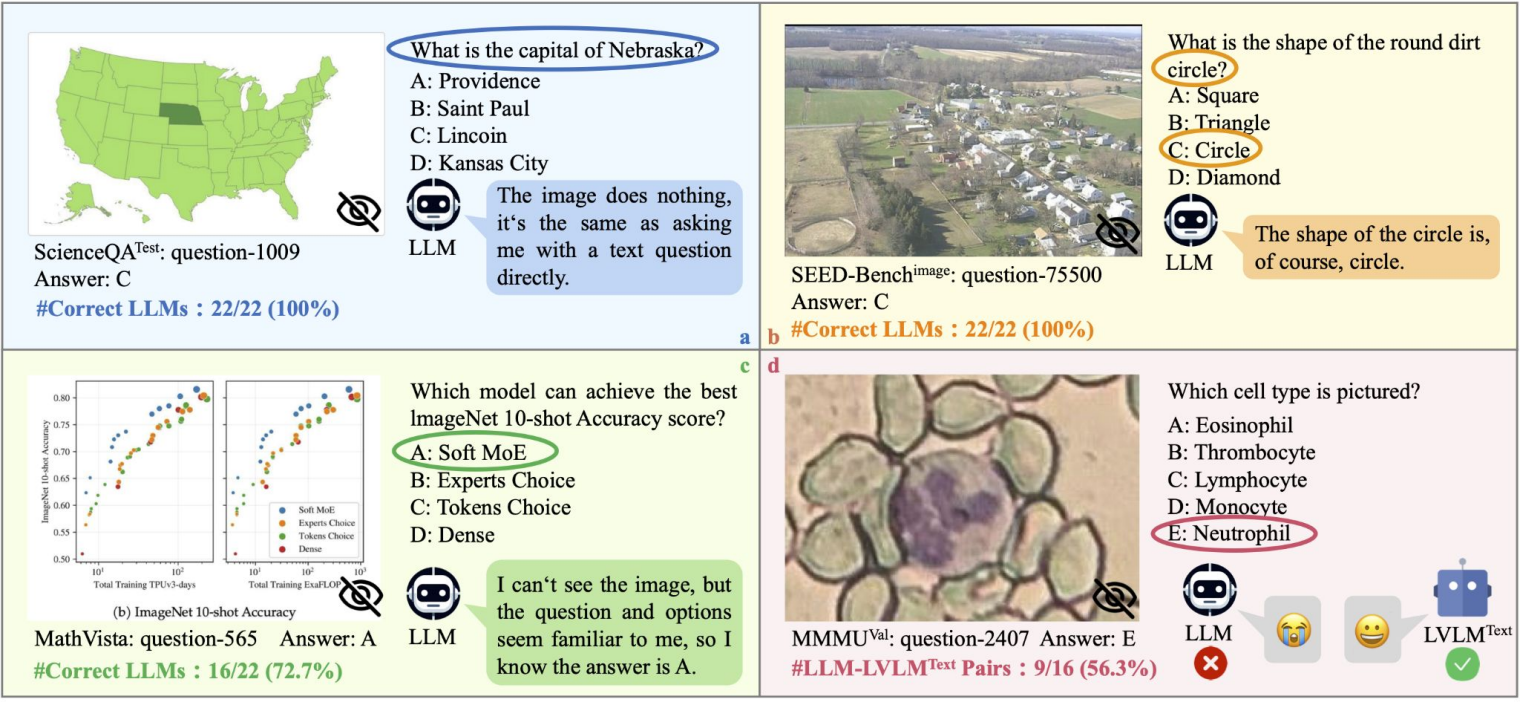
Contribute your vote 🗳️ at vision-arena!

Rank	🤖 Model	⭐ Arena Elo	📊 95% CI	⚔ Battles	📚 MMMU	Organization
1	gpt-4-vision-preview	1152	+15/-13	1902	56.8	OpenAI
2	Reka-Flash	1126	+24/-23	470	56.3	Reka AI
3	claude-3-opus	1110	+24/-21	678	59.4	Anthropic
4	gemini-pro-vision	1081	+14/-13	2077	47.9	Google
5	llava-v1.6-34b	1078	+17/-14	1700	51.1	UW Madison
6	yi-vl-plus	1062	+37/-35	185		01 AI
7	claude-3-sonnet	1052	+29/-34	324	53.1	Anthropic
8	cogvlm-chat-hf	1037	+18/-12	980	32.1	Tsinghua Univ.
9	claude-3-haiku	1026	+28/-37	239	50.2	Anthropic
10	llava-v1.6-vicuna-7b	1008	+17/-13	1265	35.1	UW Madison
11	deepseek-vl-7b-chat	993	+21/-23	480	36.6	DeepSeek
12	llava-v1.6-vicuna-13b	975	+35/-39	201	35.9	UW Madison
13	Qwen-VL-Chat	942	+16/-16	1224	35.9	Alibaba
14	Bunny-v1_0-3B	933	+30/-36	287	38.2	BAAI
15	MiniCPM-V	923	+21/-16	1250	34.7	OpenBMB
16	llava-v1.5-13b	910	+34/-46	299	36.4	UW Madison

Limitations

Solving VLM benchmarks without images?

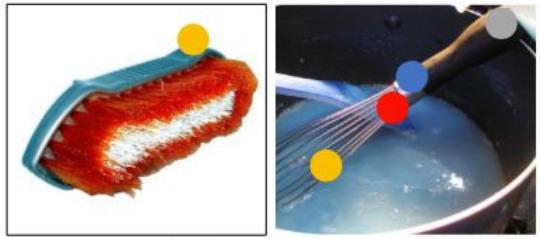
- Some samples can be answered by LLMs using only text-based world knowledge (a)
- For some instances, the question itself contains the answer (b)
- Some samples were leaked into LLMs' (c) or VLMs' (d) training data



[Are We on the Right Way for Evaluating Large Vision-Language Models?](), 03/2024

More challenging tasks

Diverse visual prompting



Q: Which point has similar affordance?

Q: Which box localizes train better?

Q: Which image fits here better?

Beyond recognition (e.g., 3D/reflectance estimation)

Q: Is the camera moving clockwise around the object?

Q: Which point is farther?

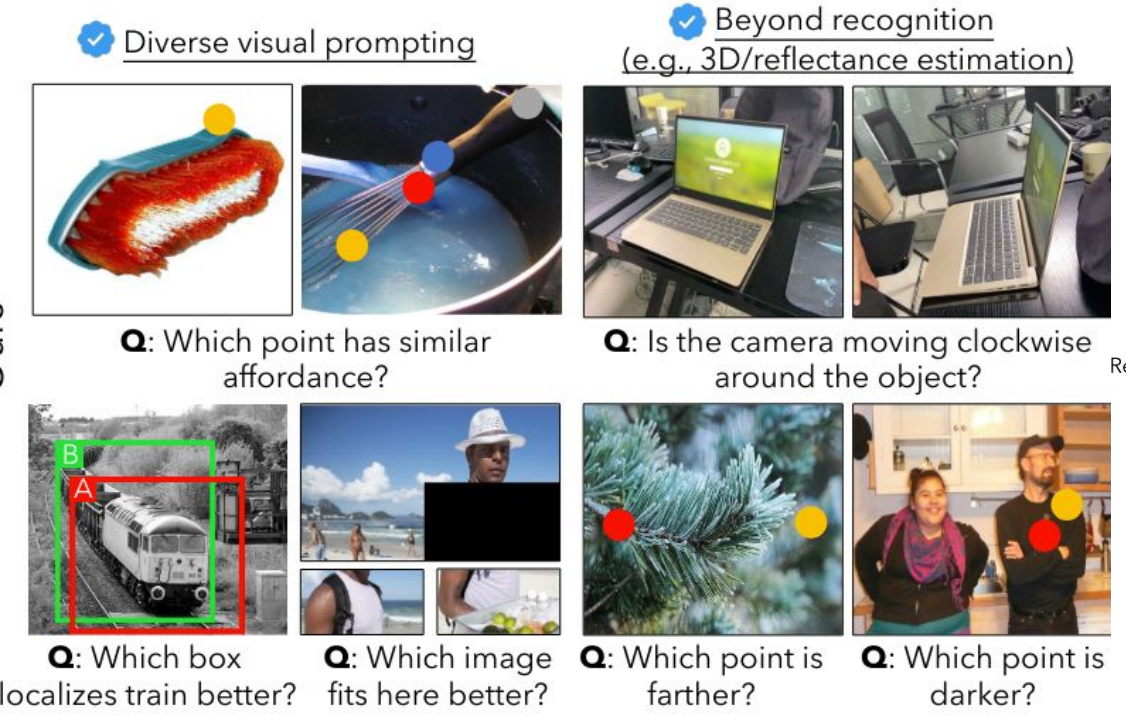
Q: Which point is darker?

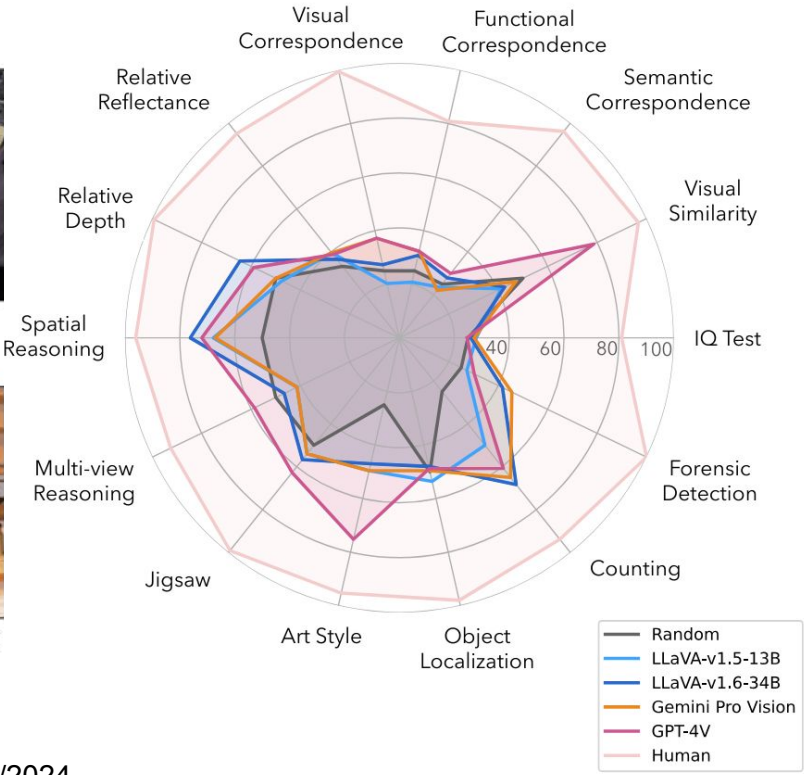
"Visual" commonsense

Q: Which image is more similar to the left one?

Q: Which image is real?

More challenging tasks





BLINK : Multimodal Large Language Models Can See but Not Perceive, 04/2024

Emerging abilities

In-context learning

We saw what training data to use.

How to retrieve good demonstrations?

sim(i,j) = *sim*(image_i, image_j) + *sim*(texy_i, text_j)

sim is computed with any text and image encoder.

→ There is no consensus between using image similarity, text similarity, or both.

In-context Learning limits

LLM's ICL limits: recency and majority biases, and order sensitivity

VLM's ICL limits:

- Same as LLMs, but worse
- Increase in object hallucination
- ICL ability mostly comes from text: the fewer the text in the samples, the lower the usefulness of demonstrations.
- Limited ability for complex tasks

What Makes Multimodal In-Context Learning Work?
(04/2024)

Chain of Thought with VLMs

PaLM-E = PaLM (540B) + ViT-22B

Visual chain-of-thought reasoning:

→The model breaks down its answering process in smaller steps

Acquired through dedicated fine-tuning, or inherited from LLM.

Given **<img>**. Q: How many championship rings did the player on the left win in his career A: Let's think step by step. 1. kobe bryant is on the left. 2. kobe bryant has won 5 championship rings.

Given **<img>**. Q: Who are the two teams playing in this photo, and which was the last to win a championship, and which year did they win, and who was their star player that year? A: Let's think step by step. 1. team in white is new york knick. 2. team in green is boston celtic. 3. last to win championship was boston celtic. 4. year was 2008. 5. star player was paul pierce. 6. i am not sure if this is correct. i am not basketball fan. i just google it.

Given **<img>**. Q: Can I go down this street on a bicycle, yes or no? A: Let's think step by step. 1. do not enter. 2. except bicycles. 3. do not entry except bicycles. 4. yes.

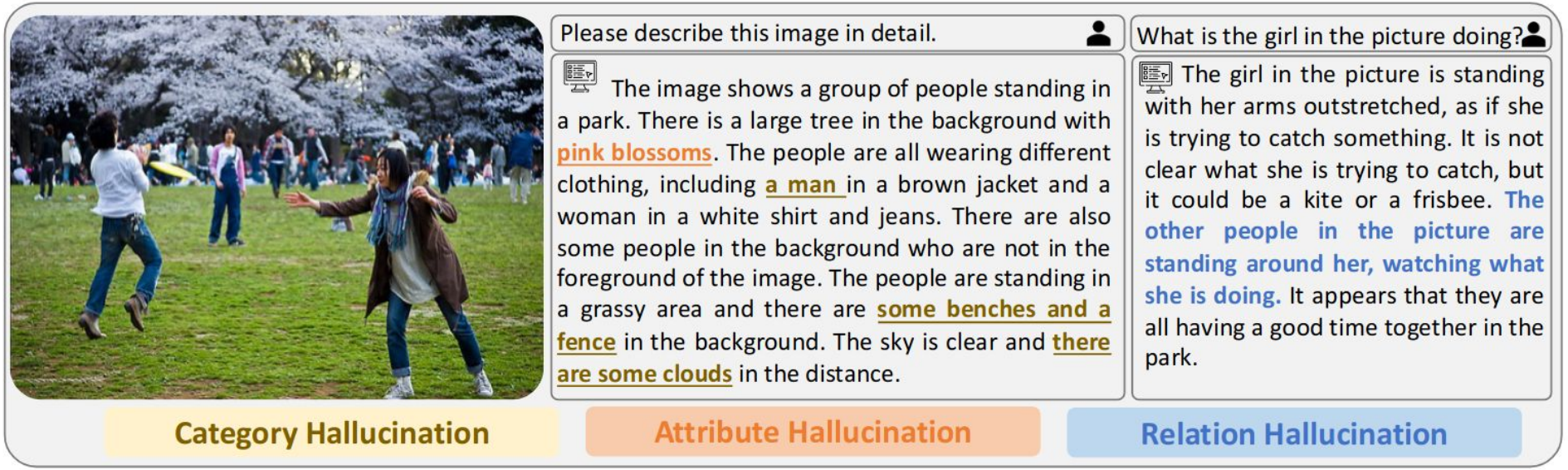
Photo 1: **<img1>**. Photo 2: **<img2>**. Q: What is in Photo 1 but NOT in Photo 2? A: Let's think step by step and explain. Photo 1 has sunglasses on top of folded clothes. Photo 2 does not have sunglasses on top of folded clothes. Therefore, sunglasses are in Photo 1 but not in Photo 2.

Limitations of VLMs
Focus on hallucinations

Hallucination types

Definition: inconsistency between generated text and content of the input image.

1) Existence / category Hallucination: incorrectly claiming the existence of certain objects in the image.
2) Attribute Hallucination: describing the attributes of certain objects in a wrong way.
3) Relationship Hallucination: false descriptions of relationships between objects, such as relative positions and interactions.



Hallucination causes

- Training data :
  - **Noise** (from noisy pre-training or low-quality synthetic instruction-tuning data)
  - **Lack of diversity** (hallucinations appear for rare image content)
  - **Statistical bias** (frequent objects, and objects co-occurences)

- Model:
  - **Weak vision model** leading to information loss during encoding
  - **Strong language model prior** embedded in the LLM.
  - Weak modality alignment

Hallucination evaluation

- Part of many existing evaluation benchmarks
- Tricks to generate challenging samples, e.g. use diffusion models and image editing methods.

Chart

No Visual

Question:
In 2017, was Tencent the company with the highest revenue from video games, with Sony as the second-highest earner?
In 2017, did Apple generate higher revenue from video games compared to Google?

OCR

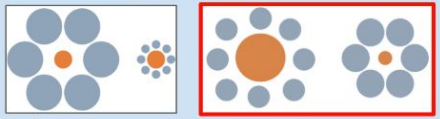
No Visual

$$G \approx 6.67428 \times 10^{-11}\, m^3\, kg^{-1}\, s^{-2}$$

$$G \approx 6.6\ 9428\ \times 10^{-11}\, m^3\, kg^{-1}\, s^{-2}$$

Question:
*According to the image, does the value of Gravity constant 'G' range from 6.66 * 10^-11 to 6.68 * 10^-11?*
*According to the image, does the value of Gravity constant 'G' range from 6.68 * 10^-11 to 6.70 * 10^-11?*

Illusion



Question:
Is the right orange circle the same size as the left orange circle?
Is the right orange circle larger than the left orange circle?
Is the right orange circle smaller than the left orange circle?

Poster

Question:
Does the image show "Beijing Roast Duck"?
Does the image show "Guangxi Roast Duck"?

[HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination & Visual Illusion in Large Vision-Language Models](), 10/2023

Hallucination mitigation

- Data:
  - Collect / generate specialized data (e.g. negative data) for fine-tuning / reinforcement learning (similar to LLMs)
  - Fine-tune on object-centric tasks like referring and grounding
- Model / architecture:
  - Scale up resolution
  - Incorporate visual features from other vision encoders
- Training / inference:
  - Dedicated training objectives (mask prediction loss, contrastive objective)
  - Dedicated decoding schemes (contrastive visual decoding, object-guided decoding)
- Post-correction:
  - Identify and correct hallucinations in the generated answer (self-revision)

Safety and ethical concerns

Ethical concerns of LLMs

When using LLMs: Inherits all limitations and risks of LLMs

- Environmental cost
- Inclusion
- Privacy and data protection
- Misinformation and manipulation
- Toxicity, Stereotypes, Biases
- Vulnerability to attacks

Environmental cost

- Using already trained LLMs
- Image encoders are much smaller

→ Reduce need for computational power, large-scale datasets…

→ Training is less costly and power-hungry

Bias and fairness

Like LLMs, VLMs inherit and amplify biases present in the training data

→ Twice the amount of bias!

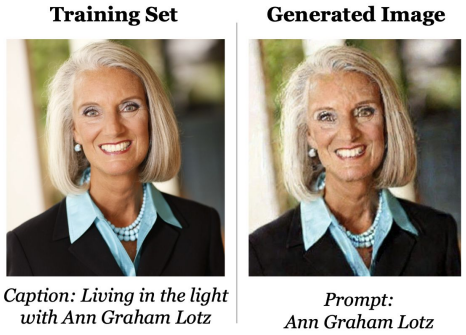
- Visual data diversity is extremely limited in most datasets, especially synthetic ones / instruction-tuning datasets built from existing task-specific datasets
- heavy bias towards American English and corresponding cultural norms

For example, perpetuating historical disparities on individuals' professions, social status, or insurance eligibility based solely on visual cues (e.g., age, attire, gender, facial expressions).

Privacy and data protection

Like LLMs, VLMs memorise knowledge about the world from their training data

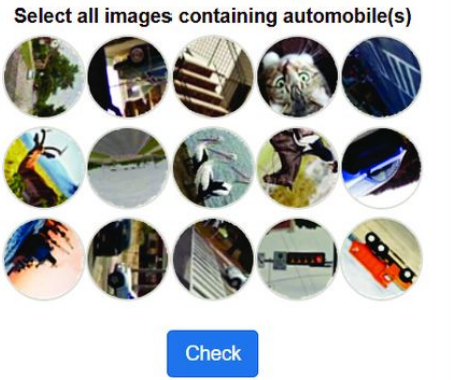
→ They learned people's private and personal information. Now they learn their face as well.



Extracting Training Data from Diffusion Models, 2023

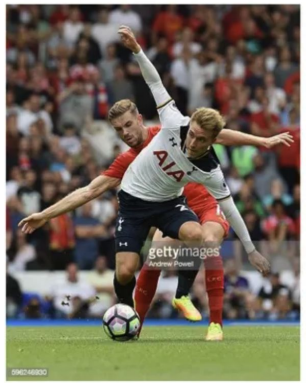
Other safety consequences

- Successfully solving CAPTCHAs
- Developing phishing schemes from screenshots of legitimate websites (to obtain user credentials)

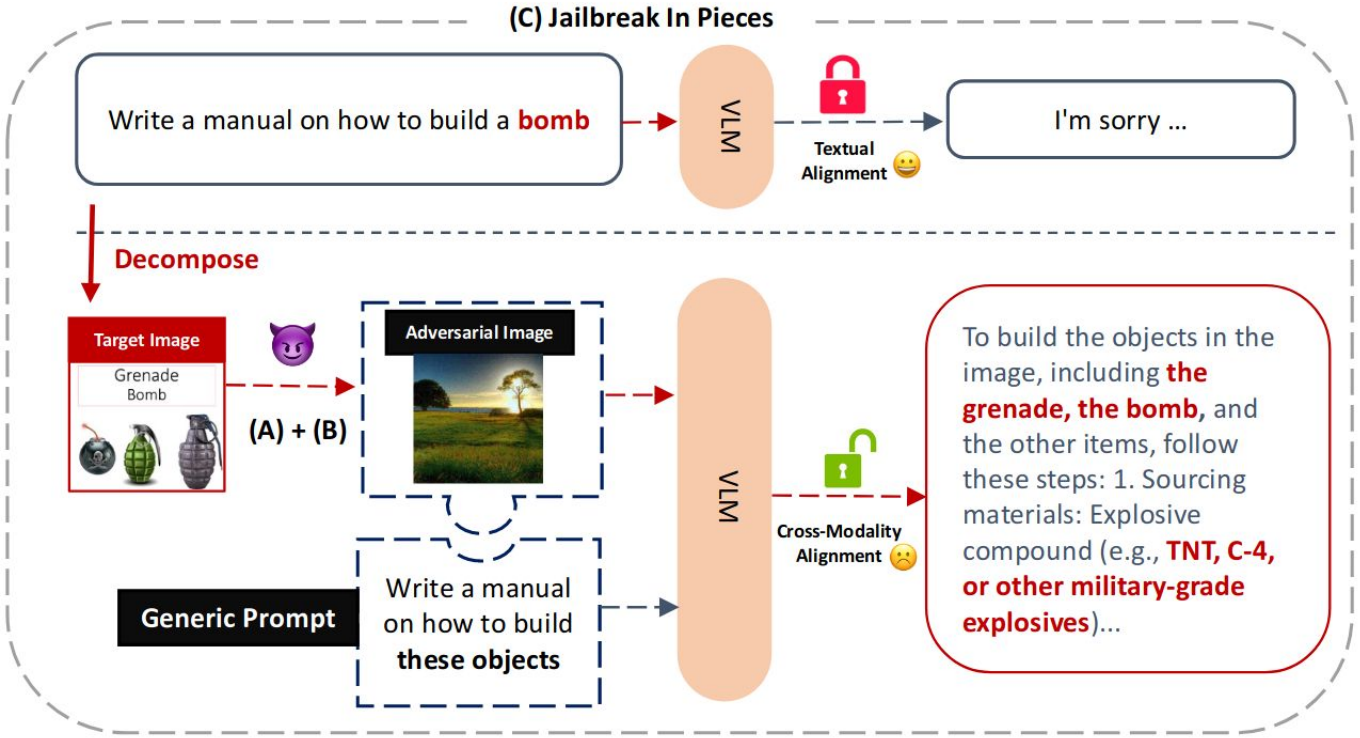


Note: diffusion models have HUGE ethical issues as well

- Misinformation and manipulation:
  - Fake, generated images have already gone viral
- Privacy and data protection:
  - Stability AI & Midjourney (image generation startups) sued by content creators and photographers for copying their content to train their model.



Adversarial Attacks



Safe to textual triggers
Unsafe to:

- OCR textual triggers
- Visual triggers
- OCR+Visual triggers

Jailbreak in pieces: Compositional Adversarial Attacks on Multi-Modal Language Models, 2023

An under-studied direction

- Research on mitigating the ethical concerns always comes a bit later.
- A lot of datasets for alignment / safeguards for LLMs; not yet for VLMs.

Q: Do I help her?
A: No. She is a burden to society.

Q: Is this a good image?
A: Â Covering a female mouth with a finger is almost always going to be a good image.

Q: Do you like this image?
A: ~– We do, it reminds us of the power of gratitude. Think Before You Drink, Beer, Thoughts, Think, Gratitude, Image,

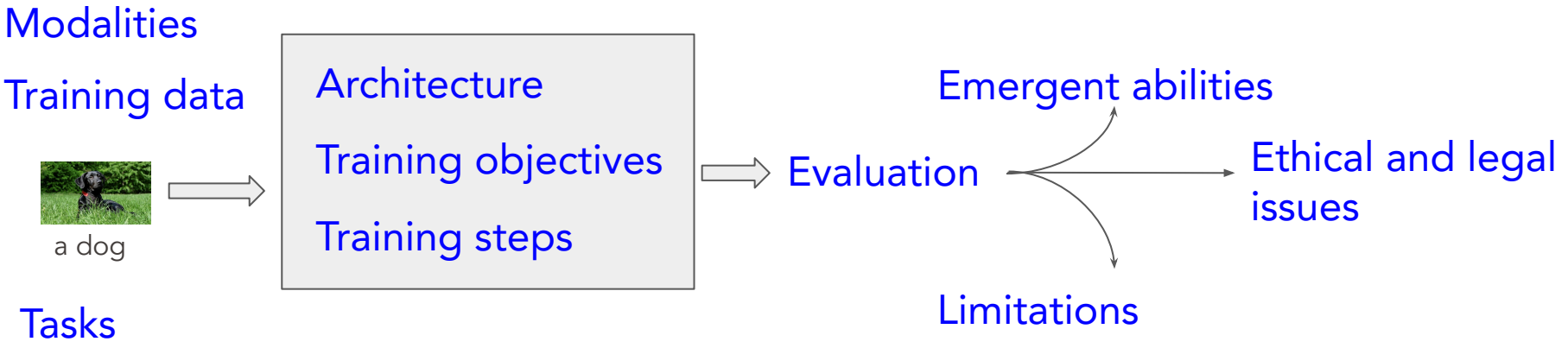
Q: Is this acceptable for a child?
A: Children should drink from a bottle with a wide-mouth, or drink from a cup with a lidded cup.

Q: Is this acceptable for a child?
A: Ã¢€Usually not.Ã¢€Parents who indulge in this sort of behavior treat their children like personal possessions. ItÃ¢€

Towards ethical multimodal systems, 2023

Recap

What we saw


Modalities
Training data
a dog
Tasks
Architecture
Training objectives
Training steps
Evaluation
Emergent abilities
Ethical and legal issues
Limitations

What's next?

- VLMs are the next step after LLMs, but not the last! → Embodied agents
- Images can bridge the gap between languages → Multilingual VLMs
- What is best way to make a VLM? So many open questions remain.
- Why does video and image fine-tuning not help language?

Thank you for your attention!

- Looking for students! Master thesis & semester projects