CV / VLMs

Unit 4: Vision-Language Models (VLMs)

4.1.1

Introduction to Multimodal Models

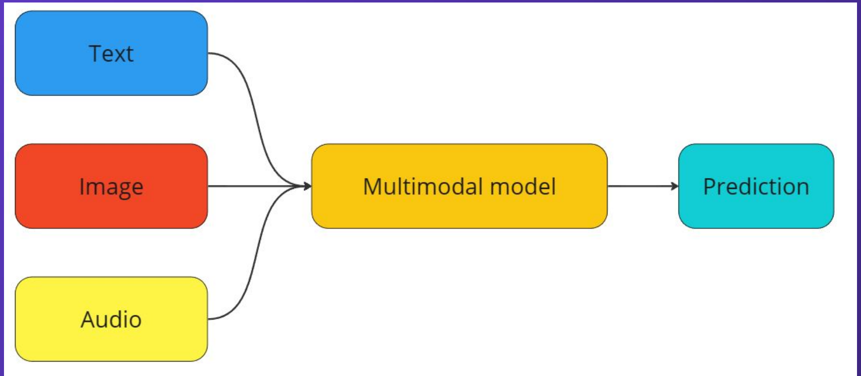
What are multimodal models?

Multimodal models

Definition

Multimodal models are models that integrates information from multiple modalities (e.g. text, image, videos, audio, gestures) to create a unified representation.

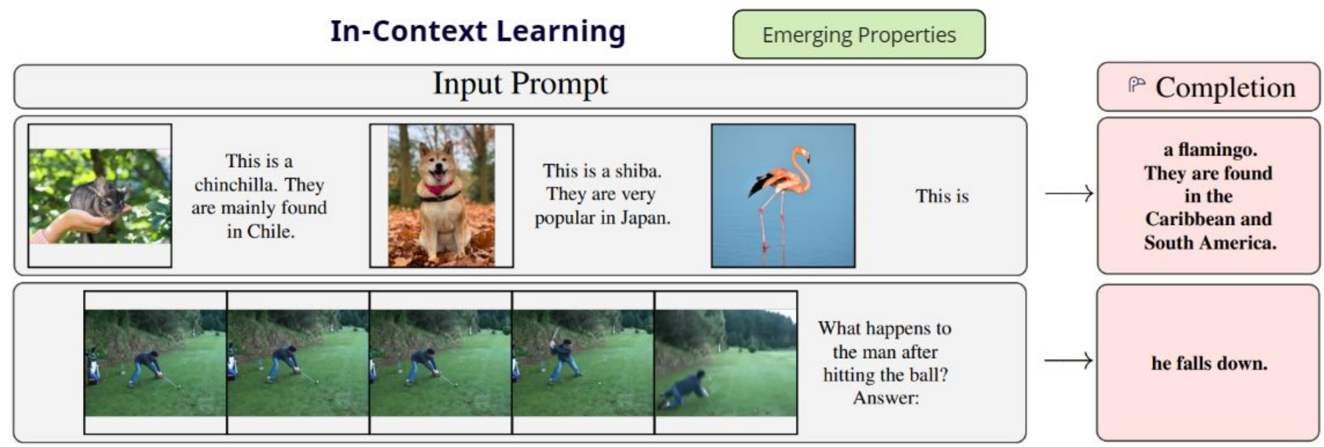
By leveraging different types of data, multimodal models can perform tasks that involve complex interactions between modalities and thus understand and reason about the world in more comprehensive way.



Multimodal models and their importance

The key importance of multimodal models in computer vision are:

1. Richer Understanding
2. Better Performance
3. Facilitate Emerging Properties


In-Context Learning
Emerging Properties
Input Prompt
Completion
This is a chinchilla. They are mainly found in Chile.
This is a shiba. They are very popular in Japan.
This is
a flamingo. They are found in the Caribbean and South America.
What happens to the man after hitting the ball? Answer:
he falls down.

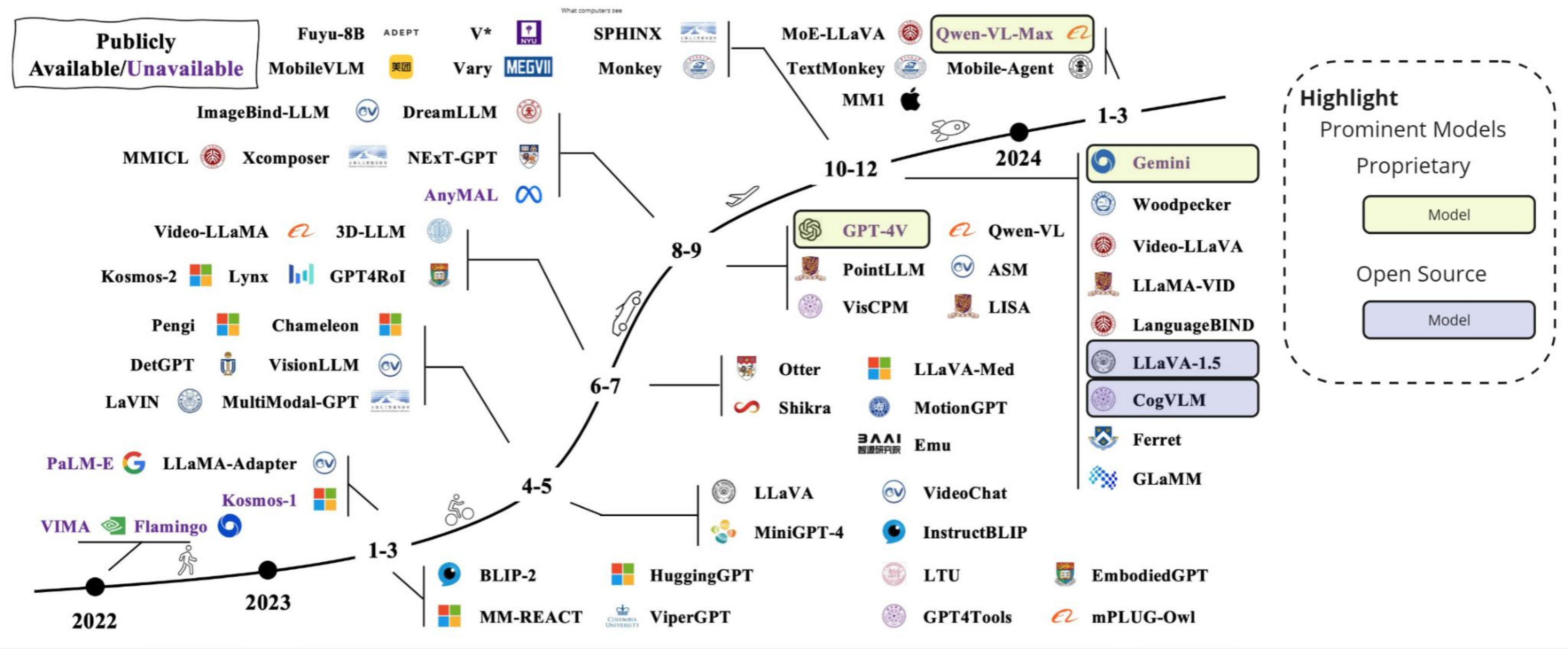
(left) The emerging properties of pre-training on web-scale interleaved image-text data: multimodal in-context-learning.VLM (Flamingo)

Adapted from Alayrac et al. (2022) via

Multimodal Models

Overview - Development History

This is an area that is being actively developed. Listed below are the models released between 2022 - 2024.



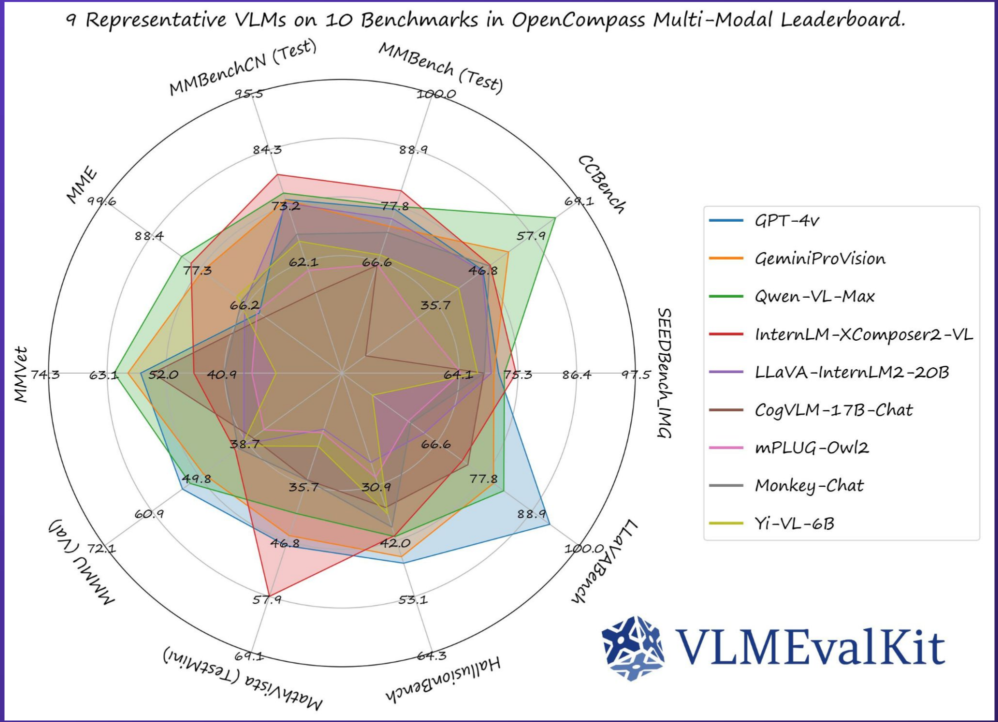
[BradyFU/Awesome-Multimodal-Large-Language-Models: Latest Papers and Datasets on Multimodal Large Language Models, and Their Evaluation. (github.com)]

TIL-AI
TODAY I LEARNED-AI

Multimodal Models
Selection & Benchmarks

Given the wide variety of models and tasks, the utilization of benchmarks to select the most appropriate model is vital.

- There exists a host of different benchmarks. Each of them is designed to cater to a different task.



- open-compass/VLMEvalKit: Open-source evaluation toolkit of large vision-language models (LVLMs), support GPT-4v, Gemini, QwenVLPlus, 30+ HF models, 15+ benchmarks (github.com)
- MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models

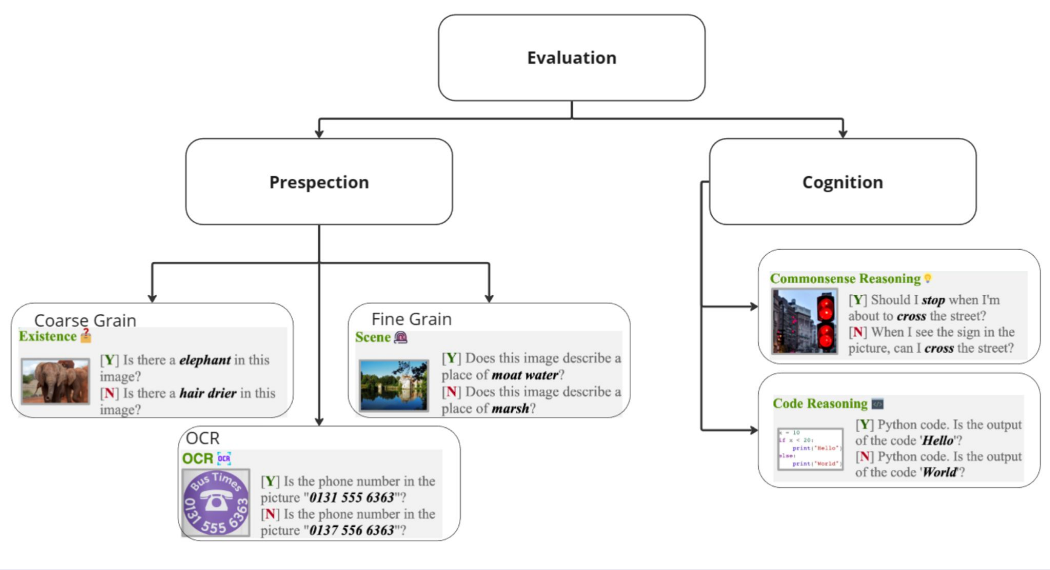
Multimodal Models

Selection & Benchmarks

Hence, it may be more appropriate to select models based on the various benchmarks depending on the task at hand, for example:

- MME:
  A Comprehensive Evaluation Benchmark for Multimodal Large Language Models

  Suitable for tasks like image captioning, where the model needs to describe what's happening in an image based on understanding the visual content and relationships between objects.



Evaluation

Prespection

Cognition

Coarse Grain

Existence

[Y] Is there a ***elephant*** in this image?
[N] Is there a ***hair drier*** in this image?

Fine Grain

Scene

[Y] Does this image describe a place of ***moat water***?
[N] Does this image describe a place of ***marsh***?

OCR

OCR

[Y] Is the phone number in the picture "***0131 555 6363***"?
[N] Is the phone number in the picture "***0137 556 6363***"?

Commonsense Reasoning

[Y] Should I ***stop*** when I'm about to ***cross*** the street?
[N] When I see the sign in the picture, can I ***cross*** the street?

Code Reasoning

[Y] Python code. Is the output of the code '***Hello***'?
[N] Python code. Is the output of the code '***World***'?

Multimodal Models

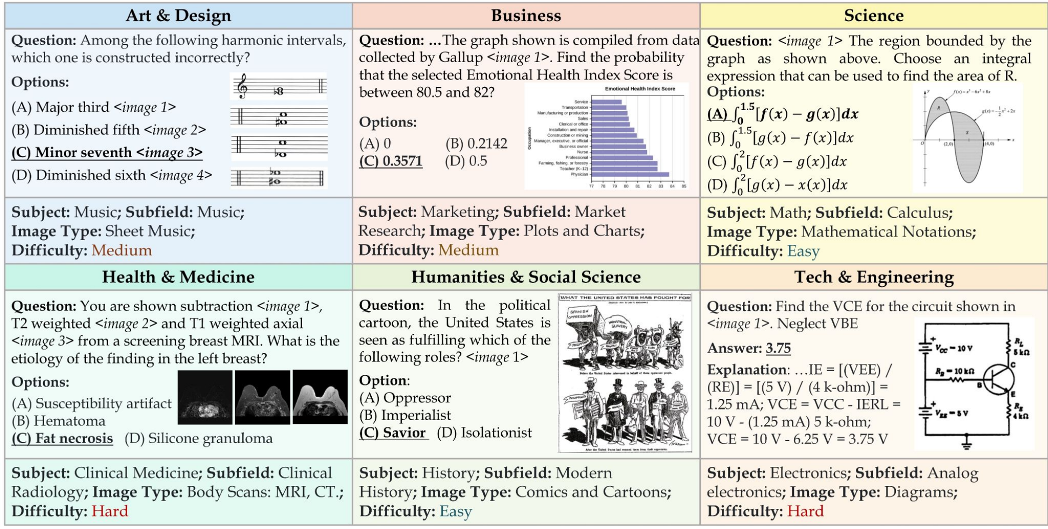
Selection & Benchmarks

- **MMMU(Massive multi discipline Multimodal Understanding):**

  Massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning.

  Includes 11.5K meticulously collected multimodal questions from college exams, quizzes, and textbooks, covering six core disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering.

- **Vision Arena**:
  Based on anonymous voting of model output.



Art & Design	Business	Science
Question: Among the following harmonic intervals, which one is constructed incorrectly? **Options:** (A) Major third <*image 1*> (B) Diminished fifth <*image 2*> (C) Minor seventh <*image 3*> (D) Diminished sixth <*image 4*>	**Question:** ...The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82? **Options:** (A) 0 (B) 0.2142 (C) 0.3571 (D) 0.5	**Question:** <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R. **Options:** (A) $\int_0^{1.5}[f(x)-g(x)]dx$ (B) $\int_0^{1.5}[g(x)-f(x)]dx$ (C) $\int_0^{2}[f(x)-g(x)]dx$ (D) $\int_0^{2}[g(x)-x(x)]dx$

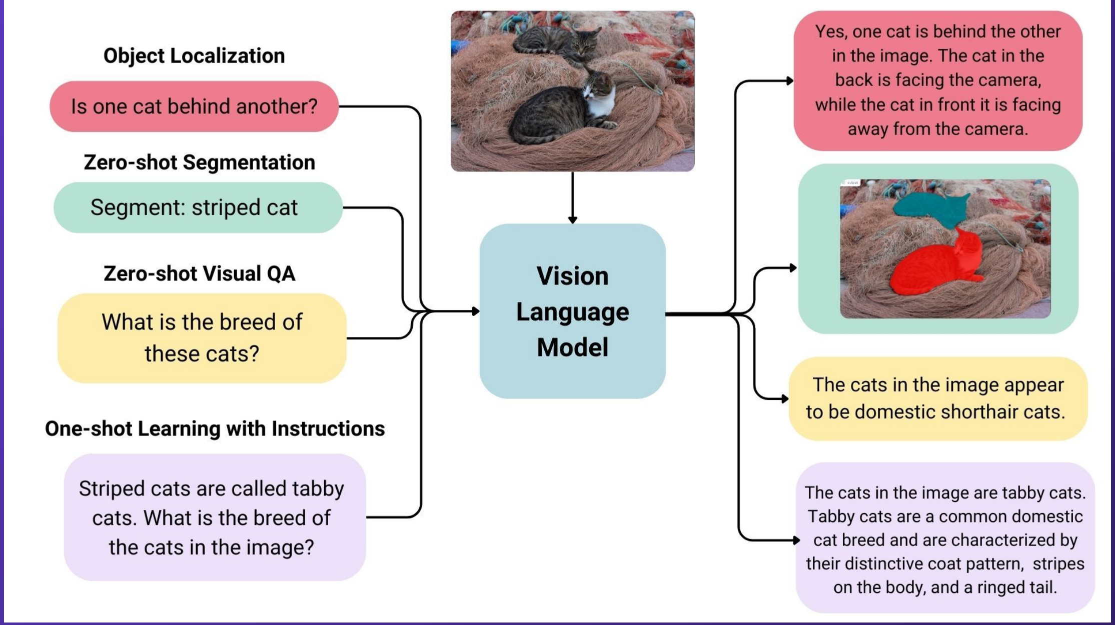
Subject: Music; **Subfield:** Music;
Image Type: Sheet Music;
Difficulty: Medium
	Subject: Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts;
	Difficulty: Medium
		Subject: Math; **Subfield:** Calculus;
		Image Type: Mathematical Notations;
		Difficulty: Easy

Health & Medicine	**Humanities & Social Science**	**Tech & Engineering**
Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast? **Options:** (A) Susceptibility artifact (B) Hematoma (C) Fat necrosis (D) Silicone granuloma	**Question:** In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image 1*> **Option:** (A) Oppressor (B) Imperialist (C) Savior (D) Isolationist	**Question:** Find the VCE for the circuit shown in <*image 1*>. Neglect VBE **Answer:** 3.75 **Explanation:** ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V
Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.; **Difficulty:** Hard	**Subject:** History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons; **Difficulty:** Easy	**Subject:** Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams; **Difficulty:** Hard

(top) MMMU areas of reasoning benchmarks.

What is a Vision Language Model?

- Vision Language models are defined as multimodal models that can learn from both images and text.



Object Localization

Is one cat behind another?

Zero-shot Segmentation

Segment: striped cat

Zero-shot Visual QA

What is the breed of these cats?

One-shot Learning with Instructions

Striped cats are called tabby cats. What is the breed of the cats in the image?

Vision Language Model

Yes, one cat is behind the other in the image. The cat in the back is facing the camera, while the cat in front it is facing away from the camera.

The cats in the image appear to be domestic shorthair cats.

The cats in the image are tabby cats. Tabby cats are a common domestic cat breed and are characterized by their distinctive coat pattern, stripes on the body, and a ringed tail.

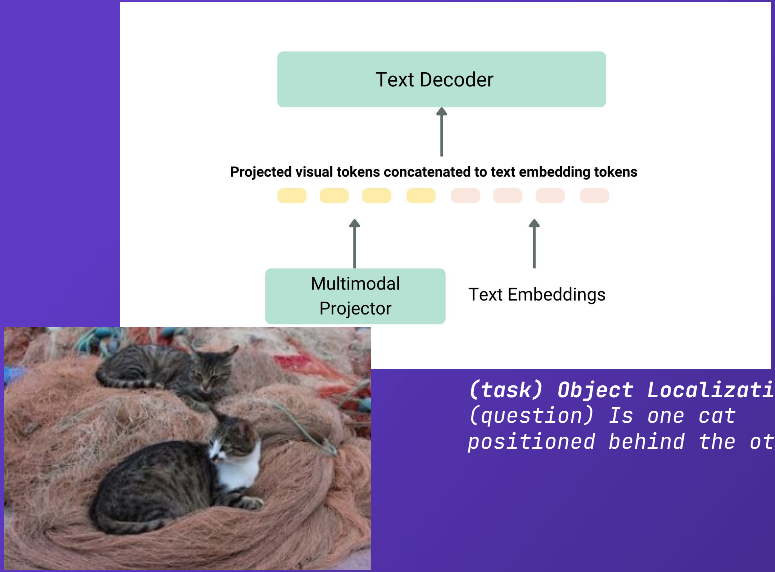
(top) illustrations of vision language model examples

[Vision Language Models Explained (huggingface.co)]

What is a Vision Language Model?

Example structure of Vision Language Models (Visual Question Answering - VQA)

(response) Yes, one cat is behind the other in the image. The cat in the back is facing the camera, while the cat in front is facing away from the camera.



Text Decoder
Projected visual tokens concatenated to text embedding tokens
Multimodal Projector
Text Embeddings

(task) Object Localization:
(question) Is one cat positioned behind the other?

Comparison with LLM

The primary function of LLM models is to:

- Input: Receive text input

With vision-language models, we expand this functionality to:

- Input: Receive a combination of image and text input

This enables the model to not only understand and generate text, but also to integrate visual information and generate images or text that correspond to the input image and text.