CSC6203/CIE6021: Large Language Model

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Lecture 8: Multimodal LLMs

Winter 2023
Benyou Wang
School of Data Science

Recap

Outline

1. Introduce multimodality
    a. What is Multimodality?
    b. Why is Multimodality?
    c. Multimodal Challenges
2. Large Multimodal Models
    a. CLIP: Contrastive Language-Image Pre-training
    b. Flamingo: the dawns of LMMs
3. Leveraging LLMs for Multimodal Purposes
    a. Multimodal in LLM era
    b. Instruction Tuning in MultiModal LLM
    c. Research Directions for Multimodal LMM
4. A case study to consider speech as the additional modality

What is Multimodality

— Let's start from the simplest concept.

What is Multimodality?

Multimodal Behaviors and Signals

Language
- **Lexicon**
  - Words
- **Syntax**
  - Part-of-speech
  - Dependencies
- **Pragmatics**
  - Discourse acts

Acoustic
- **Prosody**
  - Intonation
  - Voice quality
- **Vocal expressions**
  - Laughter, moans

Visual
- **Gestures**
  - Head gestures
  - Eye gestures
  - Arm gestures
- **Body language**
  - Body posture
  - Proxemics
- **Eye contact**
  - Head gaze
  - Eye gaze
- **Facial expressions**
  - FACS action units
  - Smile, frowning

Touch
- **Haptics**
- **Motion**

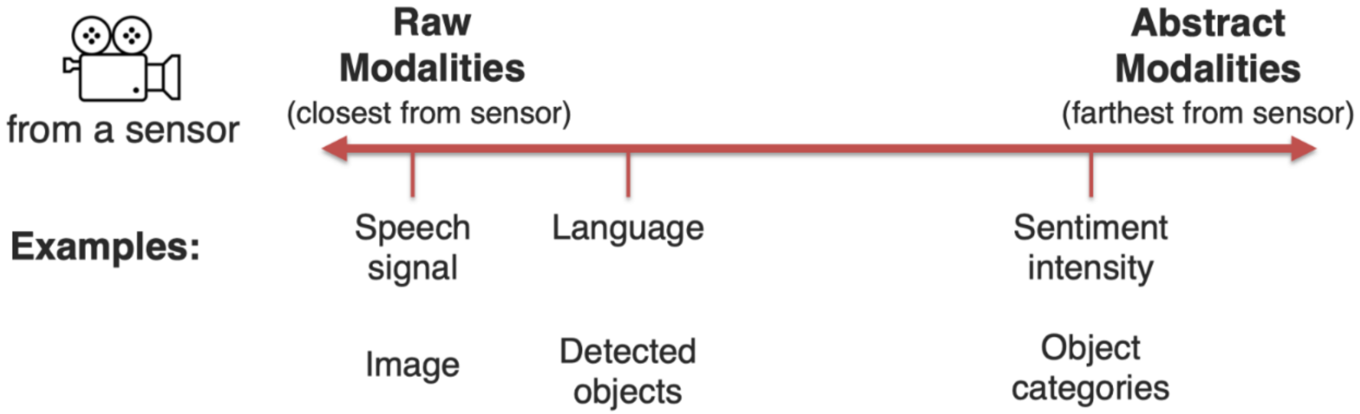
Physiological
- **Skin conductance**
- **Electrocardiogram**

Mobile
- **GPS location**
- **Accelerometer**
- **Light sensors**

What is Multimodality?

Definition

Modality refers to the way in which something expressed or perceived.



What is Multimodality?

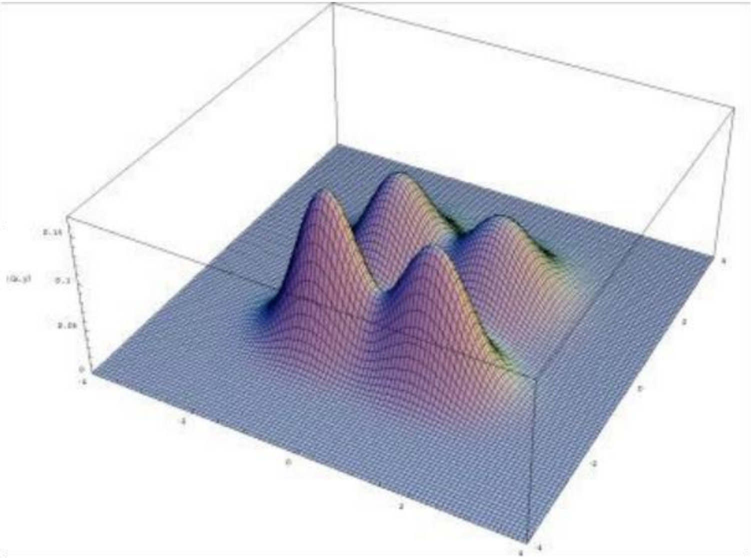
multimodal adjective

mul·ti·mod·al ˌməl-tē-ˈmō-dᵊl -ˌtī-

: having or involving several modes, modalities, or maxima

multimodal distributions

multimodal therapy



In our case, focusing on NLP: text + one or more other *modality* (images, speech, audio, olfaction, others). We'll mostly focus on images as the other modality.

What is Multimodality?

A dictionary definition…

Multimodal: with multiple modalities

A research-oriented definition…

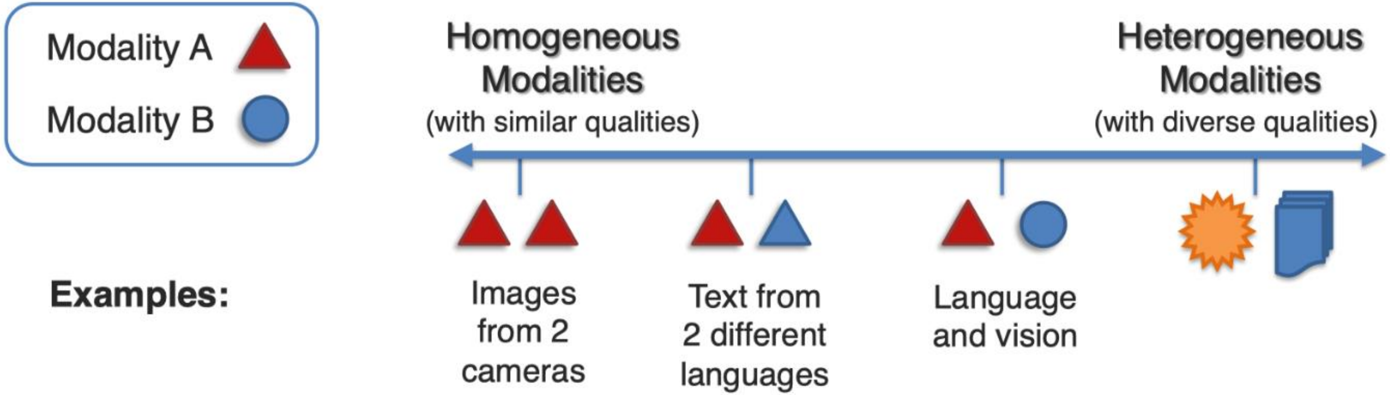
Multimodal **is the scientific study of**

heterogeneous and interconnected data

Connected + Interacting

Heterogeneous Modalities

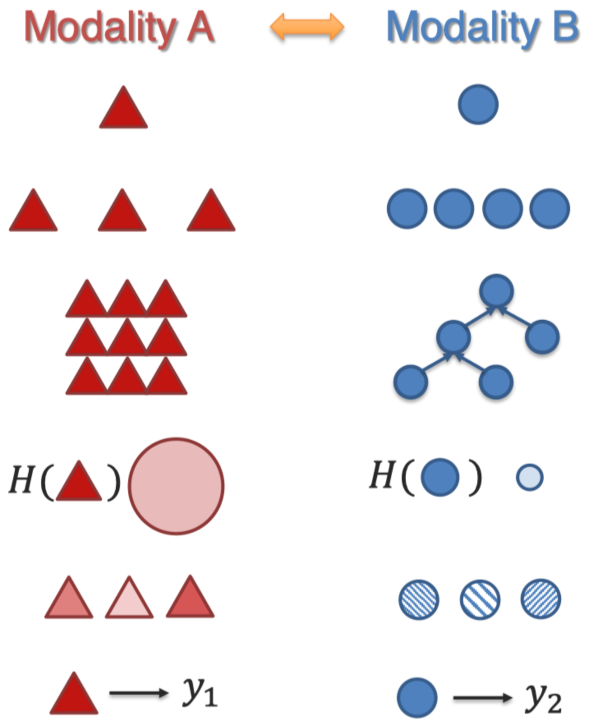
Heterogeneous: Diverse qualities, structures and representations.



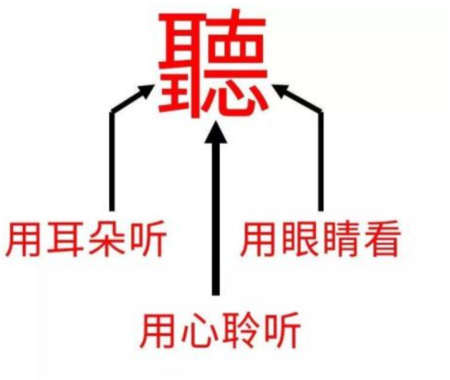
Abstract modalities are more likely to be homogeneous

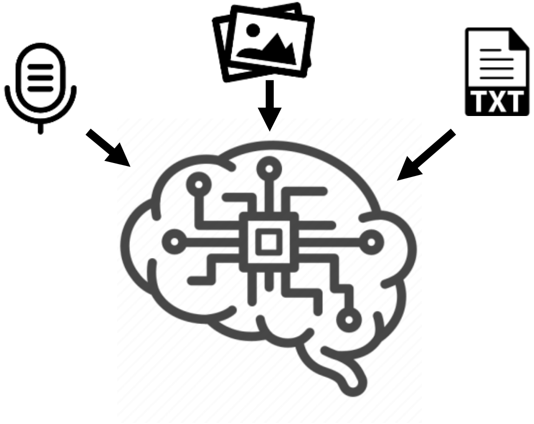
Dimensions of Heterogeneity

1. **Element representations:** Discrete, continuous, granularity
2. **Element distributions:** Density, frequency
3. **Structure:** Temporal, spatial, latent, explicit
4. **Information:** Abstraction, entropy
5. **Noise:** Uncertainty, noise, missing data
6. **Relevance:** Task, context dependence



Why is Multimodality





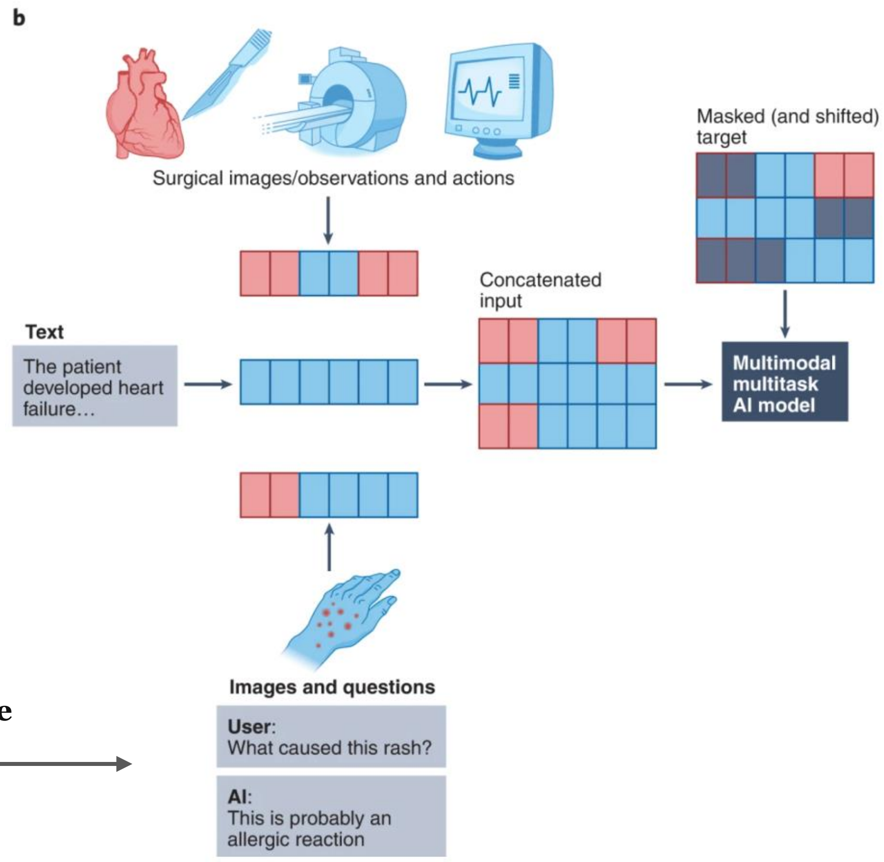
Human processes **multimodal** infos simultaneously

[1] Bergen, Benjamin K. Louder than words: The new science of how the mind makes meaning. Basic Books, 2012

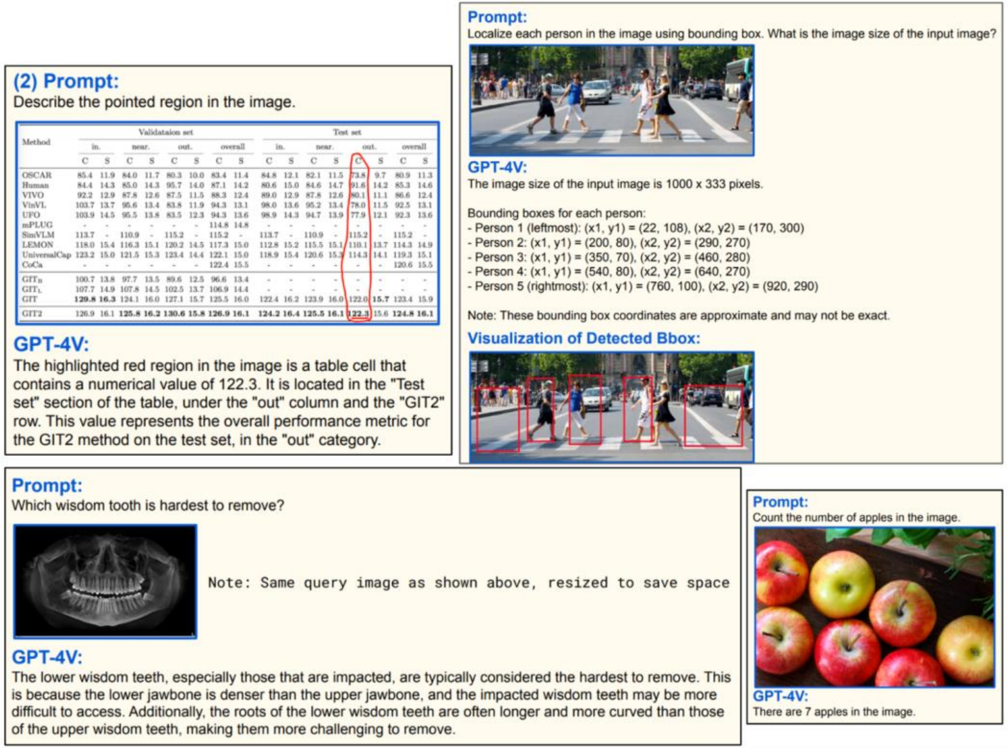
Why is Multimodality?

Many use cases are impossible without multimodality, especially those in industries that deal with a mixture of data modalities such as healthcare, robotics, e-commerce, retail, gaming, etc.

An example of how multimodality can be used in healthcare. Image from Multimodal biomedical AI



Why is Multimodality?



Some cool multimodal use cases from GPT-4V

Why is Multimodality?

A range of very good reasons:

- Faithfulness: Human experience is multimodal
- Practical: The internet & many applications are multimodal
- Data efficiency and availability:
    - Efficiency: Multimodal data is rich and “high bandwidth” (compared to language; quoting LeCun, “an imperfect, incomplete, and low-bandwidth serialization protocol for the internal data structures we call thoughts”), so better for learning?
    - Scaling: More data is better, and we’re running out of high quality text data.

Multimodality is one of the main frontiers of the new foundation model revolution.

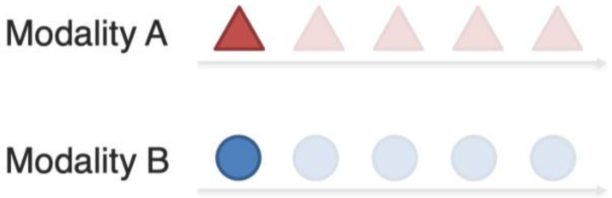
Multimodal Challenges

Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

⇨ This is a core building block for most multimodal modeling problems!

Individual elements:



It can be seen as a "local" representation

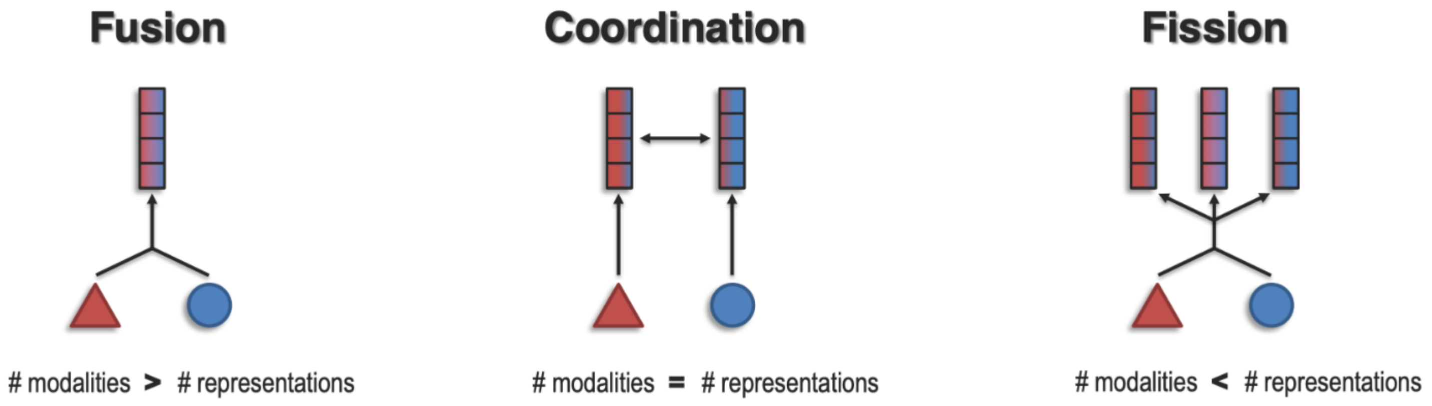
or

representation using holistic features

Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

Sub-challenges:

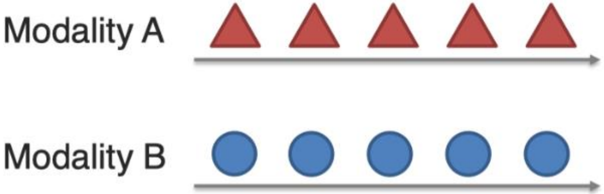


Challenge 2: Alignment

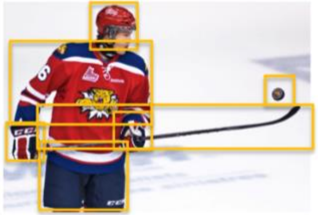
Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure

→ Most modalities have internal structure with multiple elements

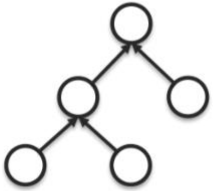
Elements with temporal structure:



Other structured examples:



Spatial

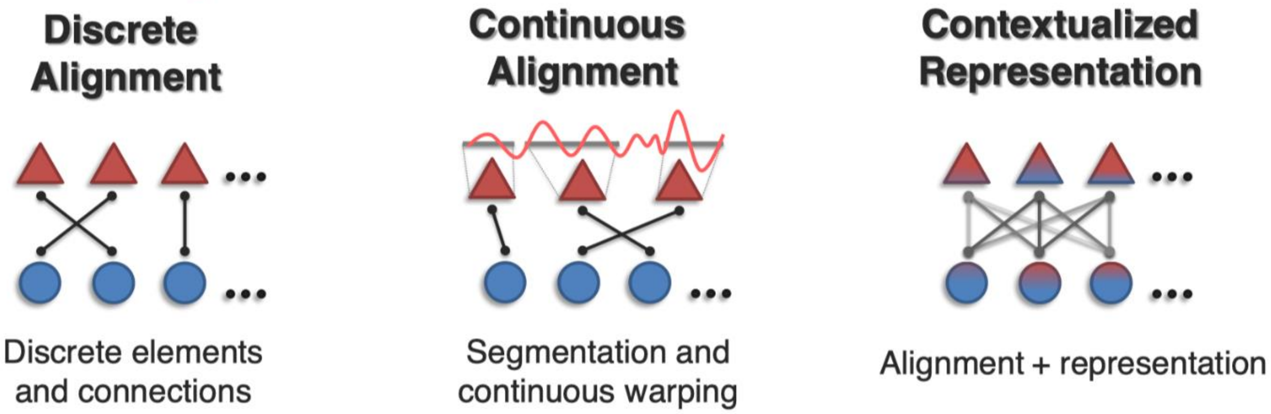


Hierarchical

Challenge 2: Alignment

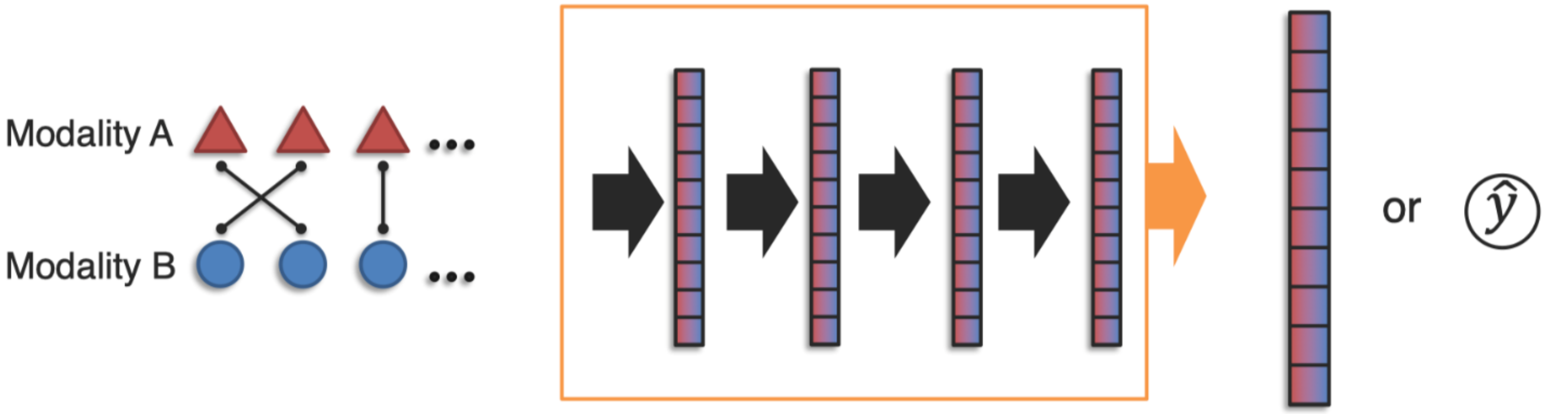
Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure

Sub-challenges:



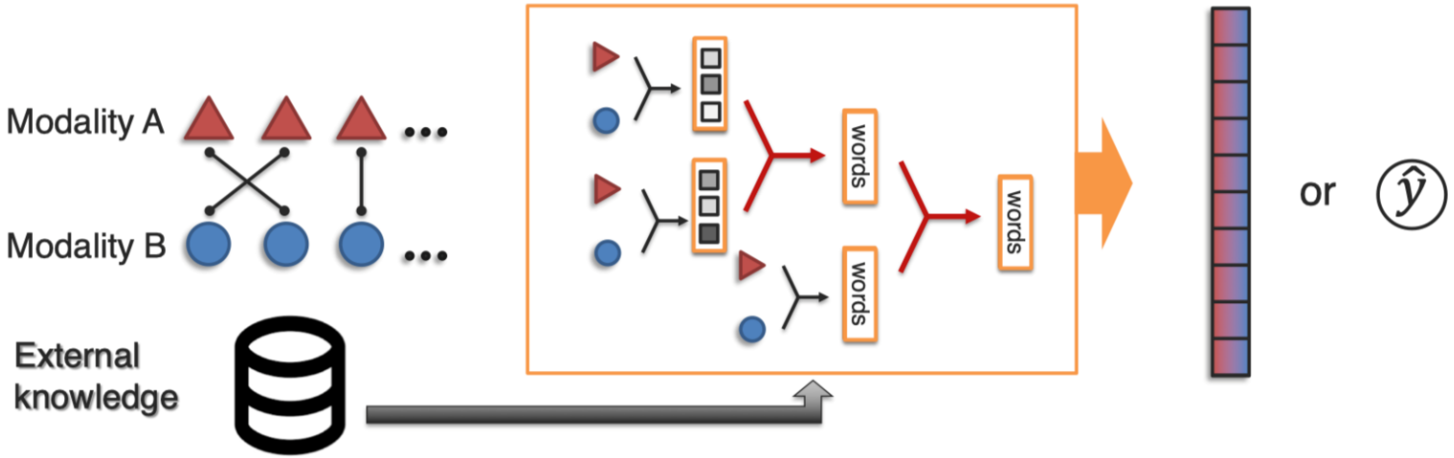
Challenge 3: Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure



Challenge 3: Reasoning

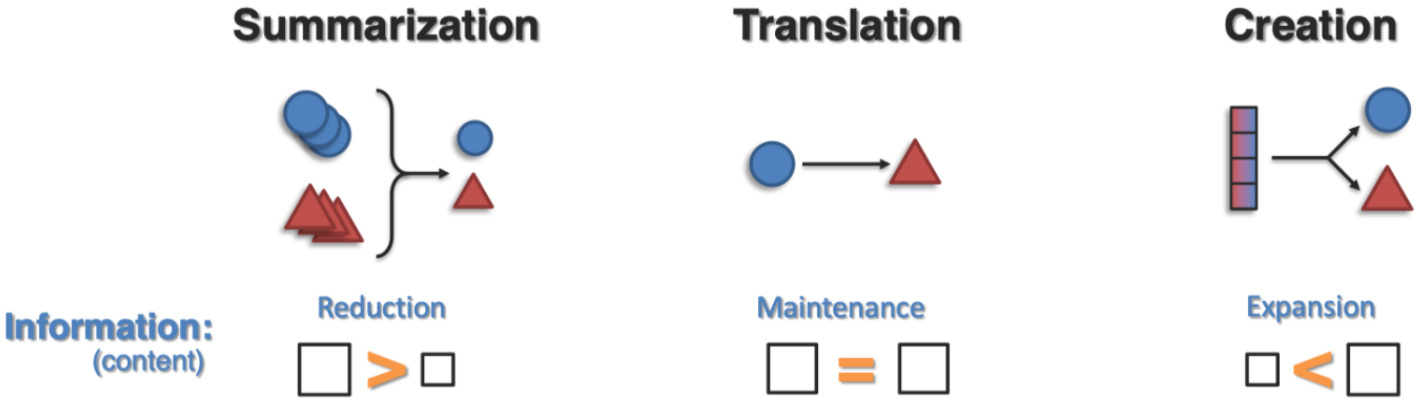
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure



Challenge 4: Generation

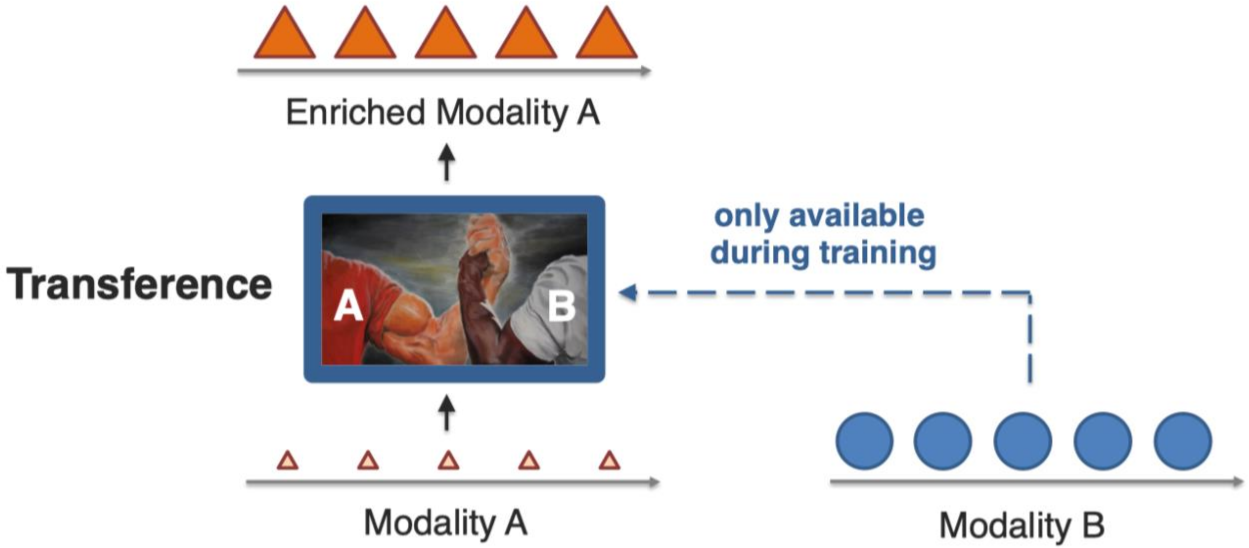
Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure and coherence

Sub-challenges:



Challenge 5: Transference

Definition: Transfer knowledge between modalities, usually to help the target modality which may be noisy or with limited resources



Large Multimodal Models

— The age of the large model is upon us, so let's just skip the small model and leap directly into the large model era.

Milestones

Given the existence of so many amazing multimodal systems, a challenge of writing this PPT is choosing which systems to focus on. Here, we will focus on two models: **CLIP (2021)** and **Flamingo (2022)** both for their significance as well as availability and clarity of public details.

- **CLIP** was the first model that could generalize to multiple image classification tasks with zero- and few-shot learning.
- **Flamingo** wasn't the first large multimodal model that could generate open-ended responses (Salesforce's BLIP came out 3 months prior). However, Flamingo's strong performance prompted some to consider it the GPT-3 moment in the multimodal domain.

- [CLIP] Learning Transferable Visual Models From Natural Language Supervision (OpenAI, 2021)
- Flamingo: a Visual Language Model for Few-Shot Learning (DeepMind, April 29, 2022)

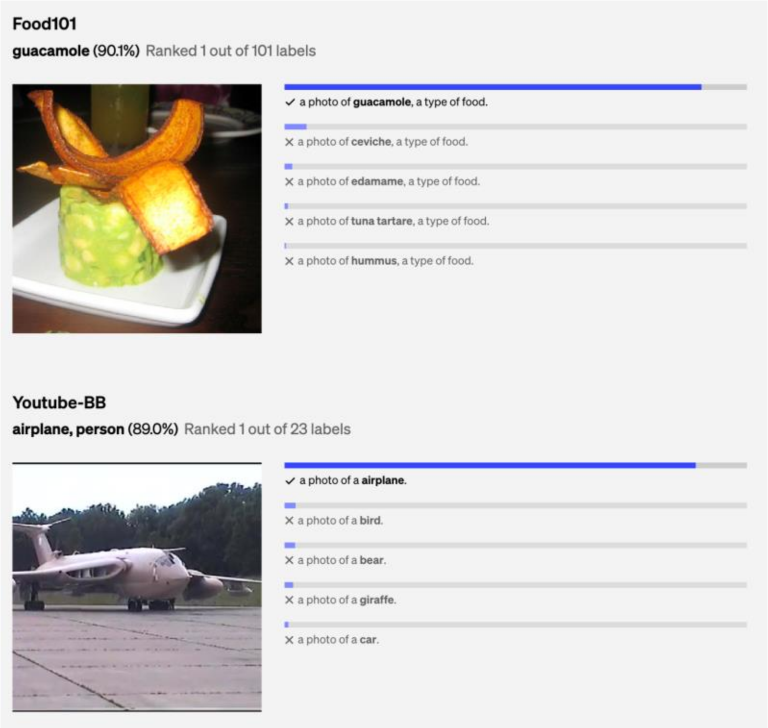
CLIP

Contrastive Language-Image Pre-training

a good practice for alignment

CLIP: Contrastive Language-Image Pre-training

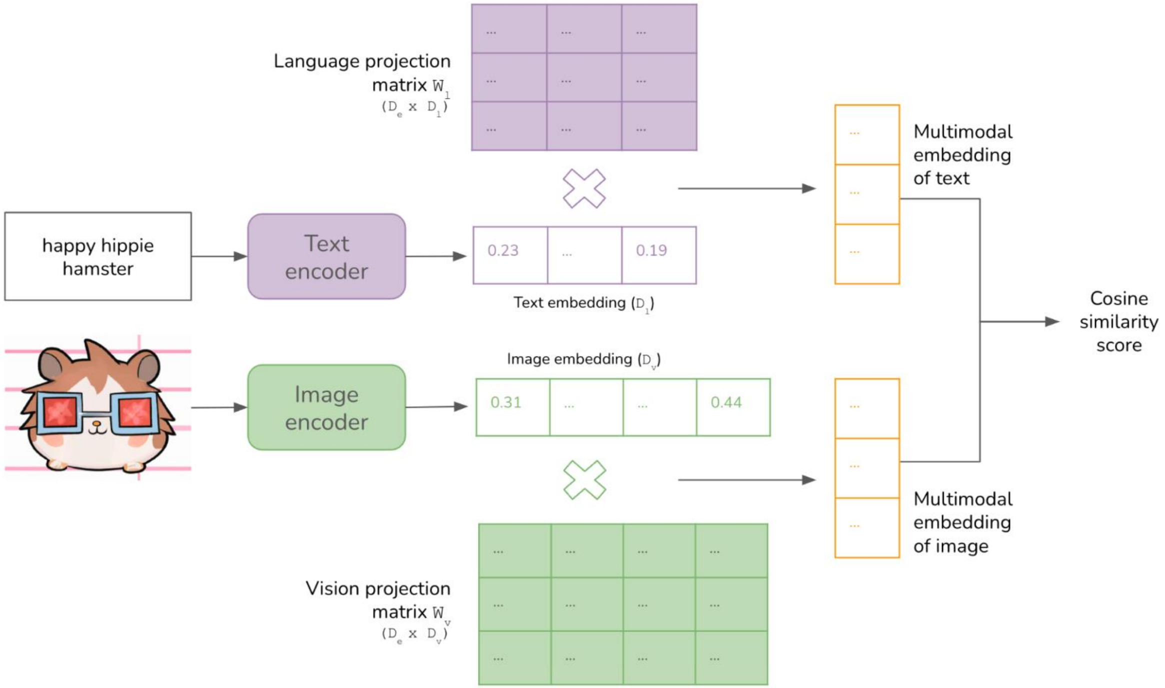
➔ CLIP leveraged **natural language supervision** and **contrastive learning**, which allowed CLIP to both scale up their data and make training more efficient. We'll go over why/how these two techniques work.



Zero-shot image classification with CLIP

- [CLIP] Learning Transferable Visual Models From Natural Language Supervision (OpenAI, 2021)

CLIP's high-level architecture



CLIP's architecture. Both encoders and projection matrices are jointly trained together from scratch. The training goal is to maximize the similarity scores of the right (image, text) pairings while minimizing the similarity scores of the wrong pairings (contrastive learning).

- [CLIP] Learning Transferable Visual Models From Natural Language Supervision (OpenAI, 2021)

Scalable Data

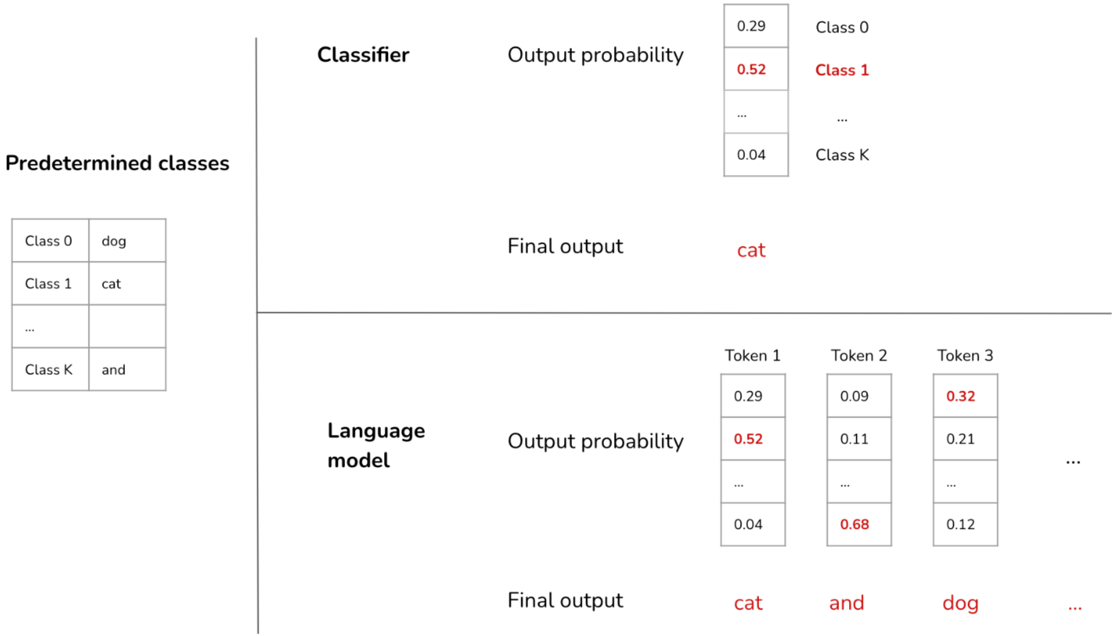
For many years, image models were trained with manually annotated (image, text) datasets (e.g. ImageNet, MS COCO). This isn't scalable. Manual annotation is time-consuming and expensive.

The CLIP paper noted that none of the then-available (image, text) datasets was big and high quality enough. They created their own dataset – 400M (image, text) pairs – as follows.

1. Construct a list of 500,000 queries. Queries are common words, bigrams, and titles of popular Wikipedia articles.
2. Find images matching these queries (string and substring match). The paper mentioned this search did NOT happen on search engines but didn't specify where. My theory is that since OpenAI already scraped the entire Internet for their GPT models, they probably just queried their internal database.
3. Each image is paired with a text that co-occurs with it (e.g. captions, comments) instead of the query since queries are too short to be descriptive.

- [CLIP] Learning Transferable Visual Models From Natural Language Supervision (OpenAI, 2021)

Language model objective



If a classifier outputs only one class for each input, a language model outputs a sequence of classes. Each generated class is called a token. Each token is from a predetermined list, the vocabulary, of the language model.

- [CLIP] Learning Transferable Visual Models From Natural Language Supervision (OpenAI, 2021)

Contrastive objective (CLIP)

Several men in hard hats are operating a giant pulley system.
Workers look down from up above on a piece of equipment.
Two men working on a machine wearing hard hats.
Four men on top of a tall structure.
Three men on a large rig.

A black dog and a white dog with brown spots are staring at each other in the street.
A black dog and a tri-colored dog playing with each other on the road.
Two dogs of different breeds looking at each other on the road.
Two dogs on pavement moving toward each other.
A black dog and a spotted dog are fighting.

Texts for an image is **diverse**

While the language model objective allows for vastly more flexible outputs, CLIP authors noted this objective made the training difficult. They hypothesized that this is because the model tries to generate exactly the text accompanying each image, while many possible texts can accompany an image: alt-text, caption, comments, etc.

- [CLIP] Learning Transferable Visual Models From Natural Language Supervision (OpenAI, 2021)

Contrastive objective (CLIP)

Contrastive learning is to overcome this challenge. Instead of predicting the exact text of each image, CLIP was trained to predict whether a text is more likely to accompany an image than other texts.

For each batch of N (image, text) pairs, the model generates N text embeddings and N image embeddings.

- Let $V_1, V_2, \ldots, V_n$ be the embeddings for the N images.
- Let $L_1, L_2, \ldots, L_n$ be the embeddings for the N texts.

CLIP computes the cosine similarity scores of the N^2 possible (V_i, L_j) pairings. The model is trained to maximize the similarity scores of the N correct pairings while minimizing the scores of the $N^2 - N$ incorrect pairings. For CLIP, $N = 32,768$.

- [CLIP] Learning Transferable Visual Models From Natural Language Supervision (OpenAI, 2021)

CLIP

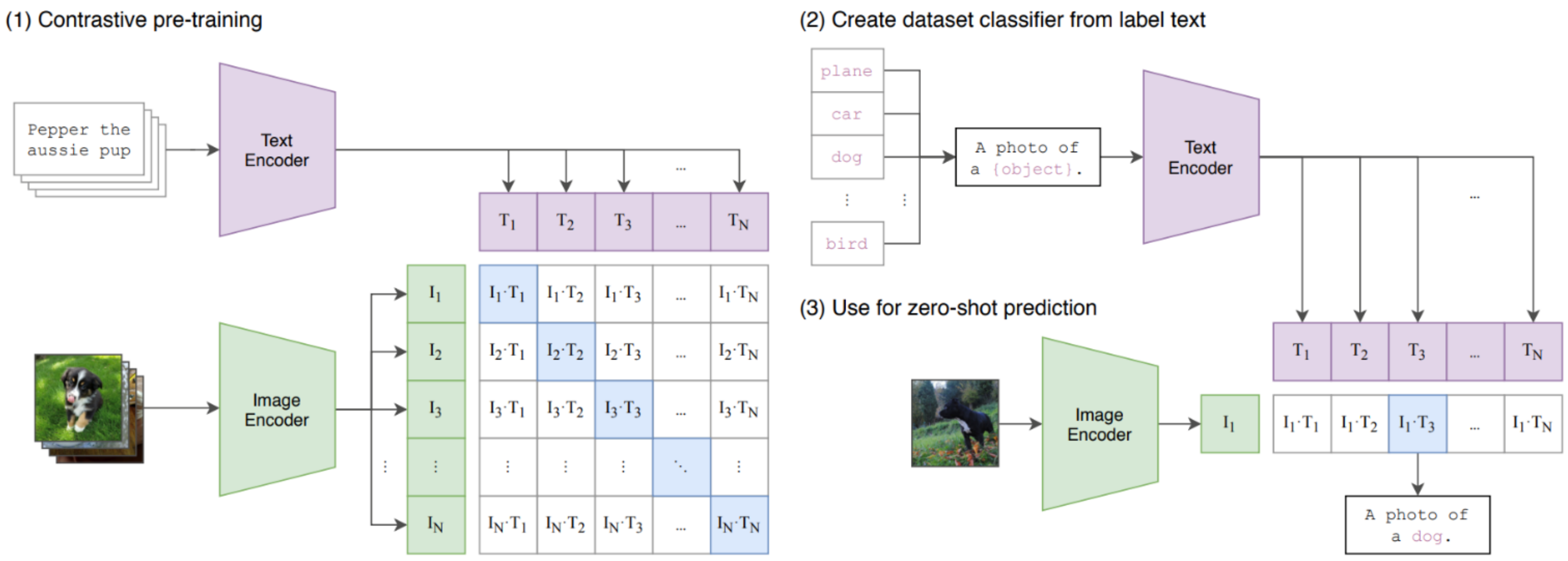
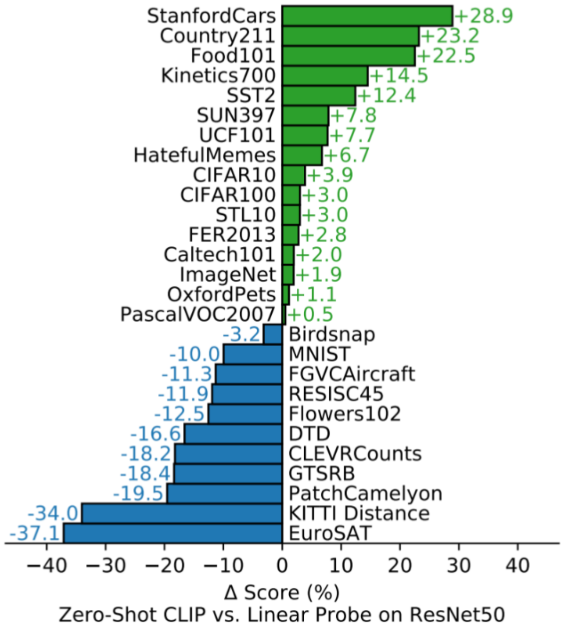


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

- [CLIP] Learning Transferable Visual Models From Natural Language Supervision (OpenAI, 2021)

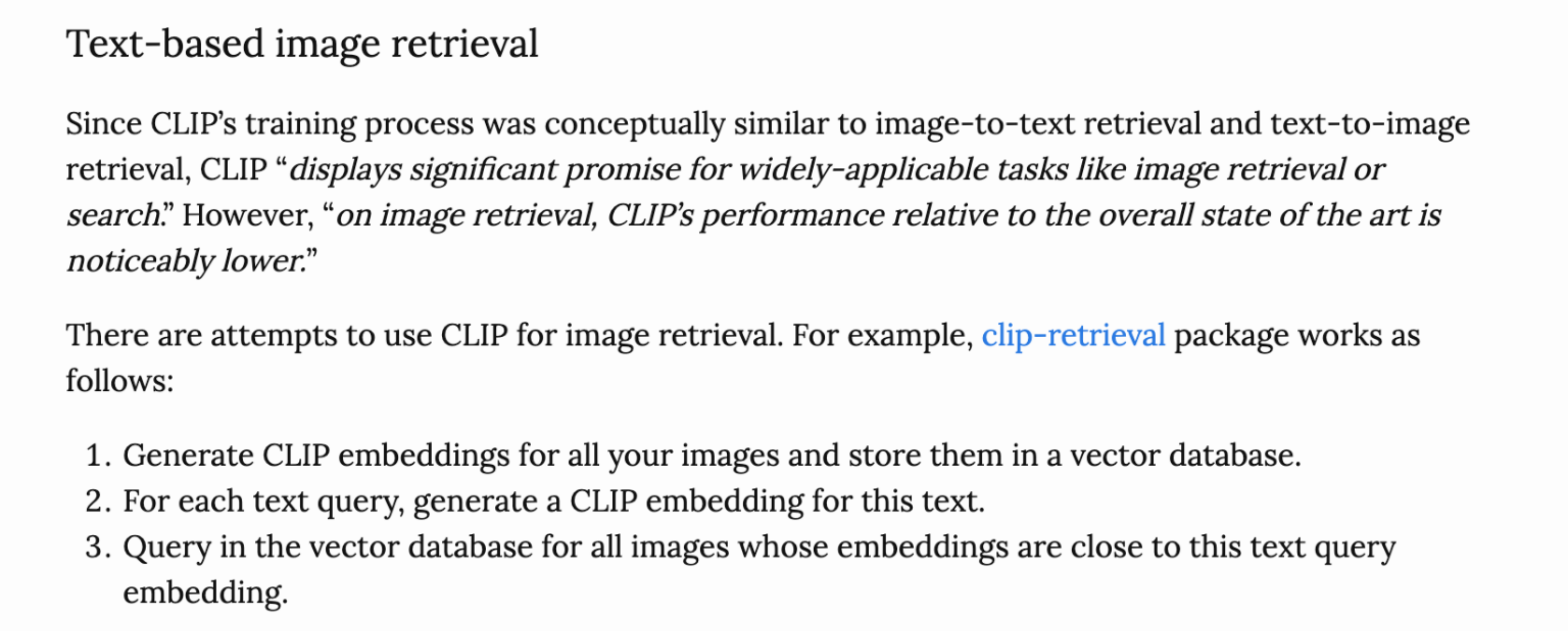
CLIP applications: classification

Today, for many image classification tasks, CLIP is still a strong out-of-the-box baseline to be used as-is or fine-tuned.



- [CLIP] Learning Transferable Visual Models From Natural Language Supervision (OpenAI, 2021)

CLIP applications: text-based image retrieval



Text-based image retrieval

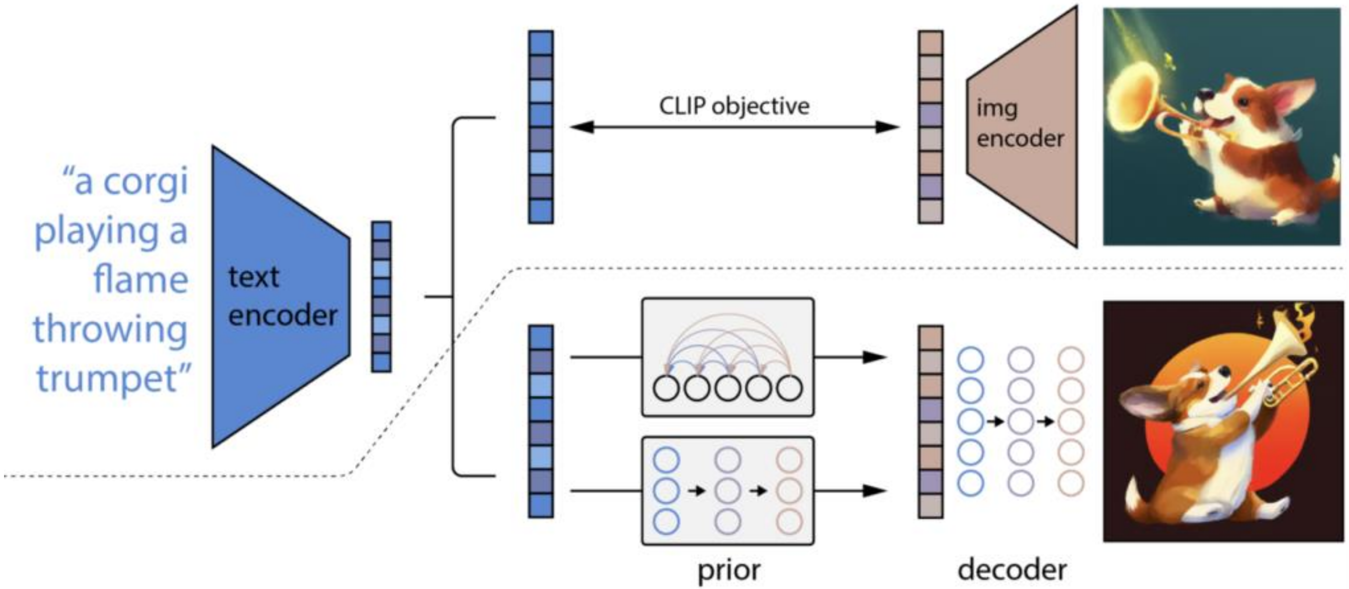
Since CLIP's training process was conceptually similar to image-to-text retrieval and text-to-image retrieval, CLIP "*displays significant promise for widely-applicable tasks like image retrieval or search.*" However, "*on image retrieval, CLIP's performance relative to the overall state of the art is noticeably lower.*"

There are attempts to use CLIP for image retrieval. For example, clip-retrieval package works as follows:

1. Generate CLIP embeddings for all your images and store them in a vector database.
2. For each text query, generate a CLIP embedding for this text.
3. Query in the vector database for all images whose embeddings are close to this text query embedding.

- [CLIP] Learning Transferable Visual Models From Natural Language Supervision (OpenAI, 2021)

CLIP applications: image generation



- [CLIP] Learning Transferable Visual Models From Natural Language Supervision (OpenAI, 2021)
- https://tryolabs.com/blog/2022/08/31/from-dalle-to-stable-diffusion?utm_source=blog&utm_medium=edgeAIalliance&utm_campaign=edgeAIalliance&utm_id=edgeAIalliance

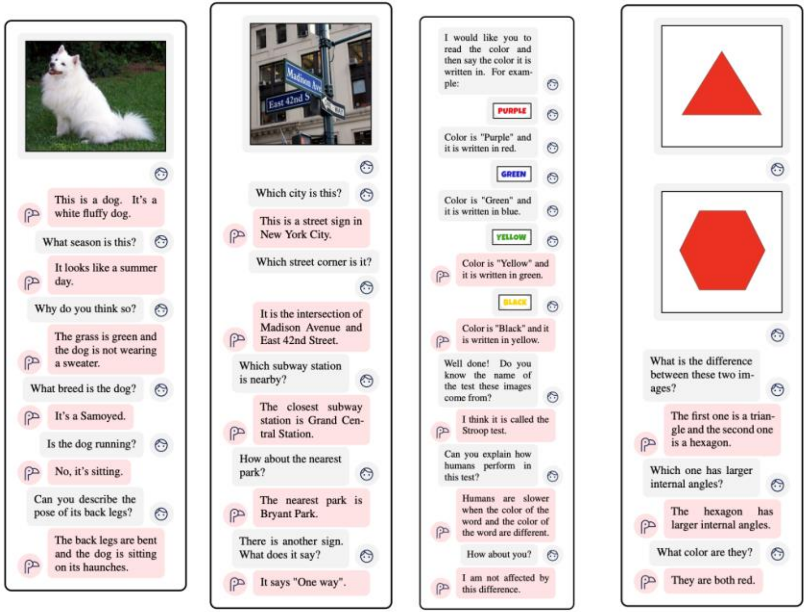
Flamingo, the dawns of LMMs

a good practice for fine-grained alignment

Flamingo: the dawns of LMMs (多模态大模型的曙光)

In a reductive view, **Flamingo = CLIP + An autoregressive language model**

- Adding a decoder LM. Unlike CLIP, Flamingo can generate text responses.
- Attention interaction between image encoder and LM



Flamingo can generate text responses conditioned on both text and images

- Flamingo: a Visual Language Model for Few-Shot Learning (DeepMind, April 29, 2022)

Flamingo's high-level architecture

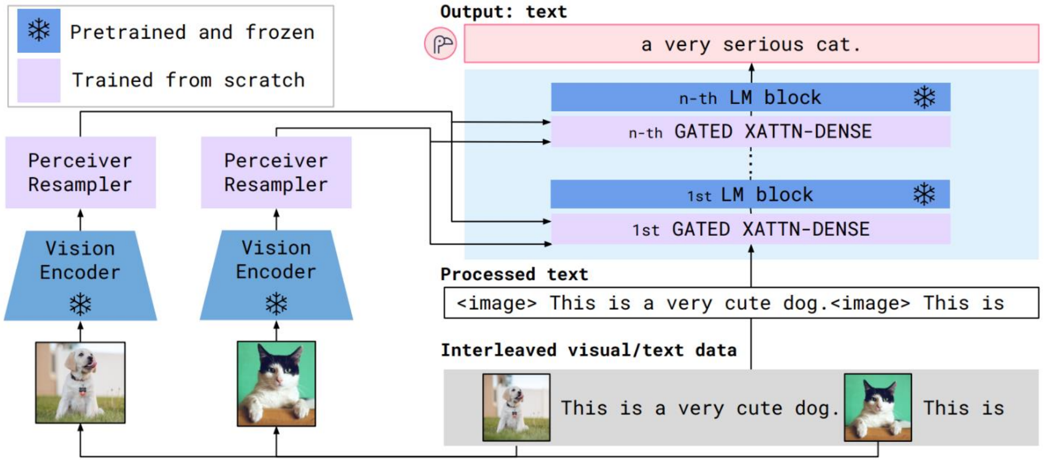


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

At a high level, Flamingo consists of **2 parts**:

- Vision encoder: a CLIP-like model is trained using contrastive learning. The text encoder of this model is then discarded. The vision encoder is frozen to be used in the main model.
- Language model: Flamingo finetunes Chinchilla to generate text tokens, conditioned on visuals and text, using language model loss, with two additional components Perceiver Resampler and GATED XATTN-DENSE layers.

- Flamingo: a Visual Language Model for Few-Shot Learning (DeepMind, April 29, 2022)

Data

Flamingo used 4 datasets: 2 (image, text) pair datasets, 1 (video, text) pair dataset, and 1 interleaved image and text dataset.

Image-Text Pairs dataset — Video-Text Pairs dataset — Multi-Modal Massive Web (M3W) dataset

Dataset	Type	Size	How	Training weight
M3W	Interleaved image and text dataset	43M webpages	For each webpage, they sample a random subsequence of 256 tokens and take up to the first 5 images included in the sampled sequence.	1.0
ALIGN	(Image, text) pairs	1.8B pairs	Texts are alt-texts, averaging 12 tokens/text.	0.2
LTIP	(Image, text) pairs	312M pairs	Texts are long descriptions, averaging 20.5 tokens/text.	0.2
VTP	(Video, text) pairs	27M short videos	~22 seconds/video on average	0.03

- Flamingo: a Visual Language Model for Few-Shot Learning (DeepMind, April 29, 2022)

Flamingo's vision encoder

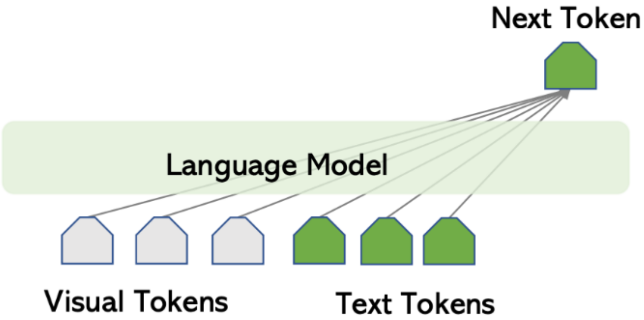
Flamingo first trains a CLIP-like model from scratch using contrastive learning. This component only uses the 2 (image, text) pair datasets, ALIGN and LTIP, totaling 2.1M (image, text) pairs. This is 5x larger than the dataset CLIP was trained on.

- For the text encoder, Flamingo uses BERT instead of GPT-2.
- For the vision encoder, Flamingo uses a NormalizerFree ResNet (NFNet) F6 model.
- Text and vision embeddings are meanpooled before being projected to the joint embedding space.

- Flamingo: a Visual Language Model for Few-Shot Learning (DeepMind, April 29, 2022)

Flamingo's language model

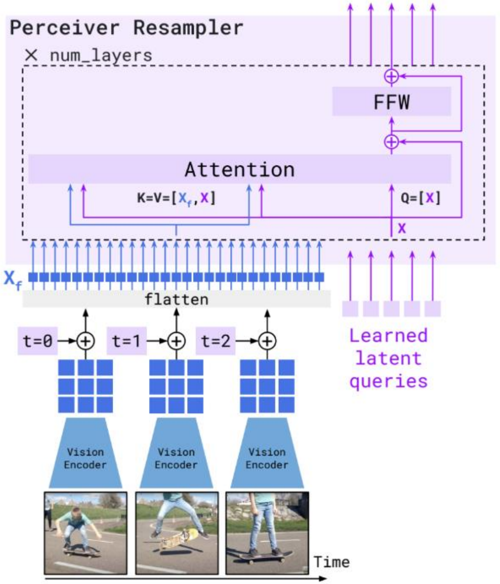
Flamingo uses Chinchilla as their language model. More specifically, they freeze the 9 pretrained Chinchilla LM layers. A traditional language model predicts the next text token based on the preceding text tokens. Flamingo predicts the next text token based on both the preceding text and visual tokens.



- Flamingo: a Visual Language Model for Few-Shot Learning (DeepMind, April 29, 2022)

Perceiver Resampler

As the visual inputs can be both images and videos, the vision encoder can produce a variable number of image or video features. Perceiver Resampler converts these variable features into a consistent 64 visual outputs.




```python
def perceiver_resampler(
    x_f,  # The [T, S, d] visual features (T=time, S=space)
    time_embeddings,  # The [T, 1, d] time pos embeddings.
    x,  # R learned latents of shape [R, d]
    num_layers,  # Number of layers
):
  """The Perceiver Resampler model."""

  # Add the time position embeddings and flatten.
  x_f = x_f + time_embeddings
  x_f = flatten(x_f)  # [T, S, d] -> [T * S, d]
  # Apply the Perceiver Resampler layers.
  for i in range(num_layers):
    # Attention.
    x = x + attention_i(q=x, kv=concat([x_f, x]))
    # Feed forward.
    x = x + ffw_i(x)
  return x
```

- Flamingo: a Visual Language Model for Few-Shot Learning (DeepMind, April 29, 2022)

GATED XATTN-DENSE layers

GATED XATTN-DENSE layers are inserted between existing and frozen LM layers to allow the language model to attend more efficiently to the visual tokens when generating text tokens. Without these layers, Flamingo authors noted a drop of 4.2% in the overall score.

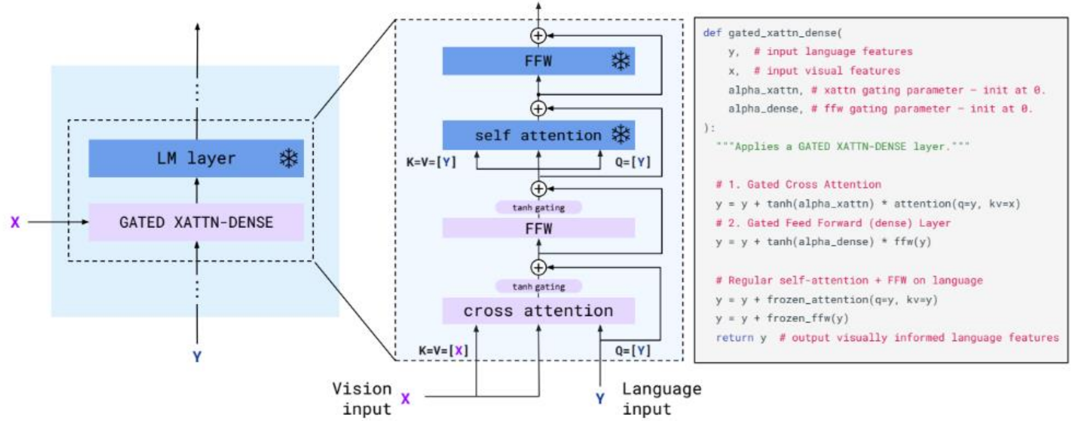


Figure 4: **GATED XATTN-DENSE layers.** To condition the LM on visual inputs, we insert new cross-attention layers between existing pretrained and frozen LM layers. The keys and values in these layers are obtained from the vision features while the queries are derived from the language inputs. They are followed by dense feed-forward layers. These layers are *gated* so that the LM is kept intact at initialization for improved stability and performance.

- Flamingo: a Visual Language Model for Few-Shot Learning (DeepMind, April 29, 2022)

Loss function

Flamingo computes the likelihood of text y conditioned on the interleaved images and videos x.

$$p(y|x) = \prod_{l=1}^{N} p(y_l \,|y_{<l}\,, x_{\leq l})$$

The training loss function was a weighted sum of expected negative log-likelihoods of generated text across all 4 datasets, with λ_m being the training weight of dataset m.

$$\sum_{m=1}^{M} \lambda_m E_{(x,y)\sim D_m}[-\sum_{l=1}^{L} \log p(y|x)]$$

- Flamingo: a Visual Language Model for Few-Shot Learning (DeepMind, April 29, 2022)

Flamingo Training

While Flamingo isn't open-sourced, there are many open-source replications of Flamingo.

- IDEFICS (HuggingFace)
- mlfoundations/open_flamingo

- Flamingo: a Visual Language Model for Few-Shot Learning (DeepMind, April 29, 2022)

CLIP vs. Flamingo

		CLIP (2021)	Flamingo (2022)
Data		- WIT (WebImageText): 400M (image, text) pairs scraped from the Internet. 500K queries Up to 20K (image, text) per query	4 datasets: - ALIGN: 1.8B (image, text) pairs. Avg. 12 tokens/text. - LTIP (Long Text & Image Pairs): 312M pairs. Avg. 20.5 tokens/text. - VTP (Video & Text Pairs): 27M (short video, text) pairs. Avg. 22 seconds/video. - M3W (MultiModal MassiveWeb): 43M webpages with interleaved images and text. Up to 256 tokens and 5 images per page.
Natural language supervision for vision encoder	Text encoder	CBOW or **text transformer** (a smaller version of GPT-2)	BERT. It's only used to train the vision encoder and discarded after.
	Vision encoder	ResNet or **Vision Transformer (ViT)**	NormalizerFree ResNet (NFNet). Once trained, it's frozen before being used in Flamingo's main model.
	Training	- Text and image embeddings are projected into a common multimodal embedding space using **linear projections** - Text and image encoders are jointly trained from scratch using **contrastive objectives**.	Same contrastive objectives as CLIP
Language model		X	- Chinchilla (9 layers) - **Perceiver Resampler** maps from varying-sized visual features to a fixed number of visual tokens (64). - **GATED XATTN-DENSE layers** are inserted between Chinchilla LM layers to allow the LM to attend to visual tokens when generating text tokens. - Chinchilla is pretrained and frozen. Perceiver Resampler and GATED XATTN-DENSE layers are trained from scratch.

- Flamingo: a Visual Language Model for Few-Shot Learning (DeepMind, April 29, 2022)

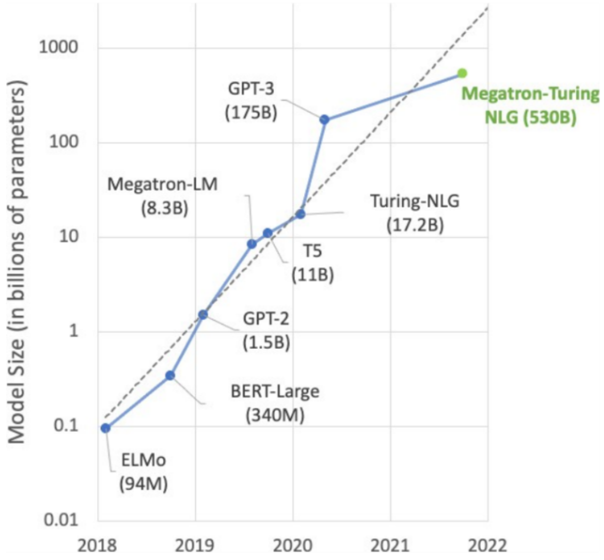
Leveraging LLMs for Multimodal Purposes

— Welcome to the Era of Large Language Models.

Multimodality in LLM era

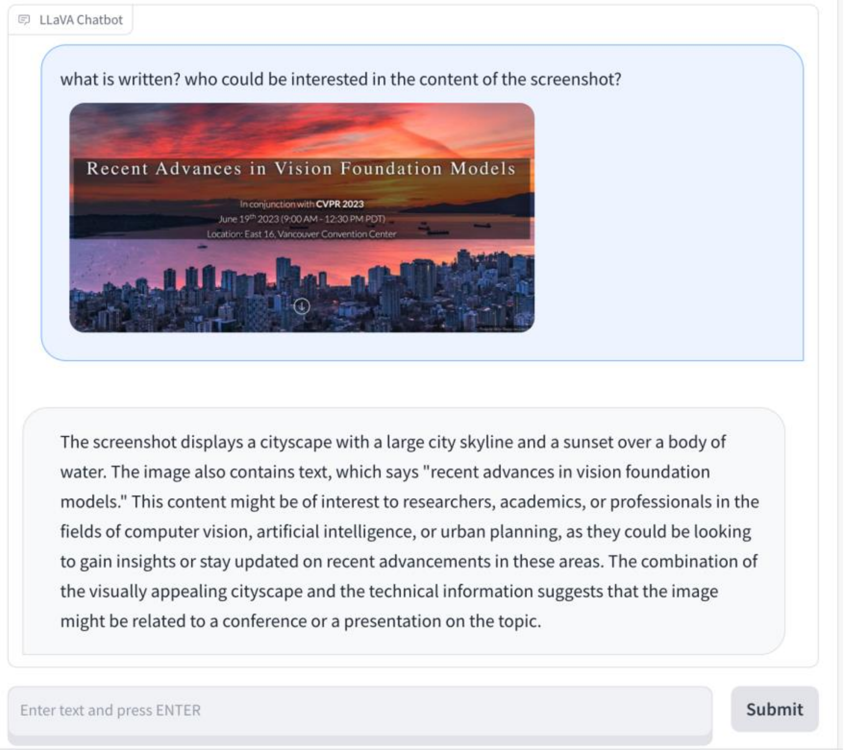
(Large) language models really improve the **cognition**-level abilities,
but it also needs **perceive** data from many other modalities.

The Era of Large Language Models (LLMs)



Model Name	Year	# Parameters
T0	2021	11B
LaMDA	2021	137B
InstructGPT	2022	175B
GPT-NeoX	2022	20B
OPT	2022	175B
PaLM	2022	**540B**
GLM-130B	2022	130B
BLOOM	2022	176B
Galactica	2022	120B
ChatGPT	2022	**1760B**

Multimodal LLMs



Visual chat with LMM. Generated by LLaVA: https://llava-vl.github.io/

Multimodal LLMs is hot right now

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

Shaohan Huang,* Li Dong,* Wenhui Wang,* Yaru Hao,* Saksham Singhal,* Shuming Ma*
Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi
Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, Furu Wei†
Microsoft

GPT-4V

OpenAI

- Model Details: Unknown
- Capability: Strong zero-shot visual understanding & reasoning on many user-oriented tasks in the wild

- How can we build Multimodal GPT-4 like models?

GPT-4 visual input example, Extreme Ironing:

User What is unusual about this image?

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

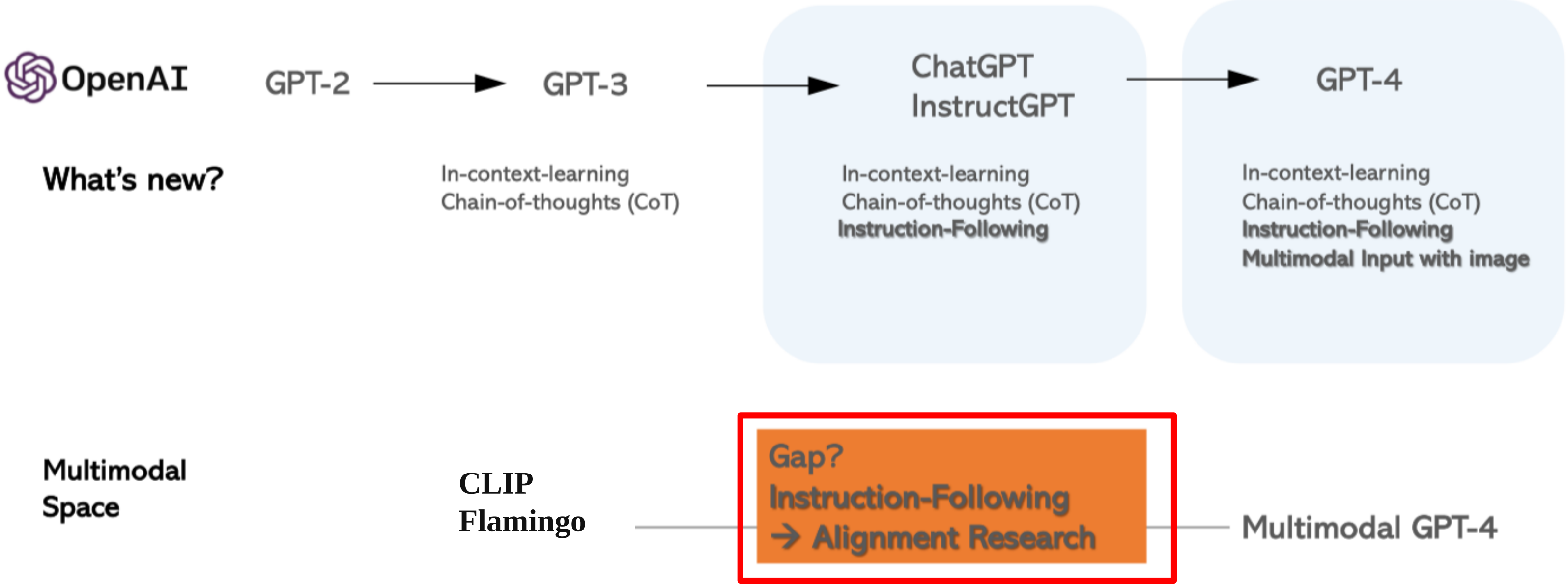
GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

GPT-4 visual input example, Chicken Nugget Map:

User Can you explain this meme?

GPT-4 This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets.
The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world.
The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

Recap on Language Modeling



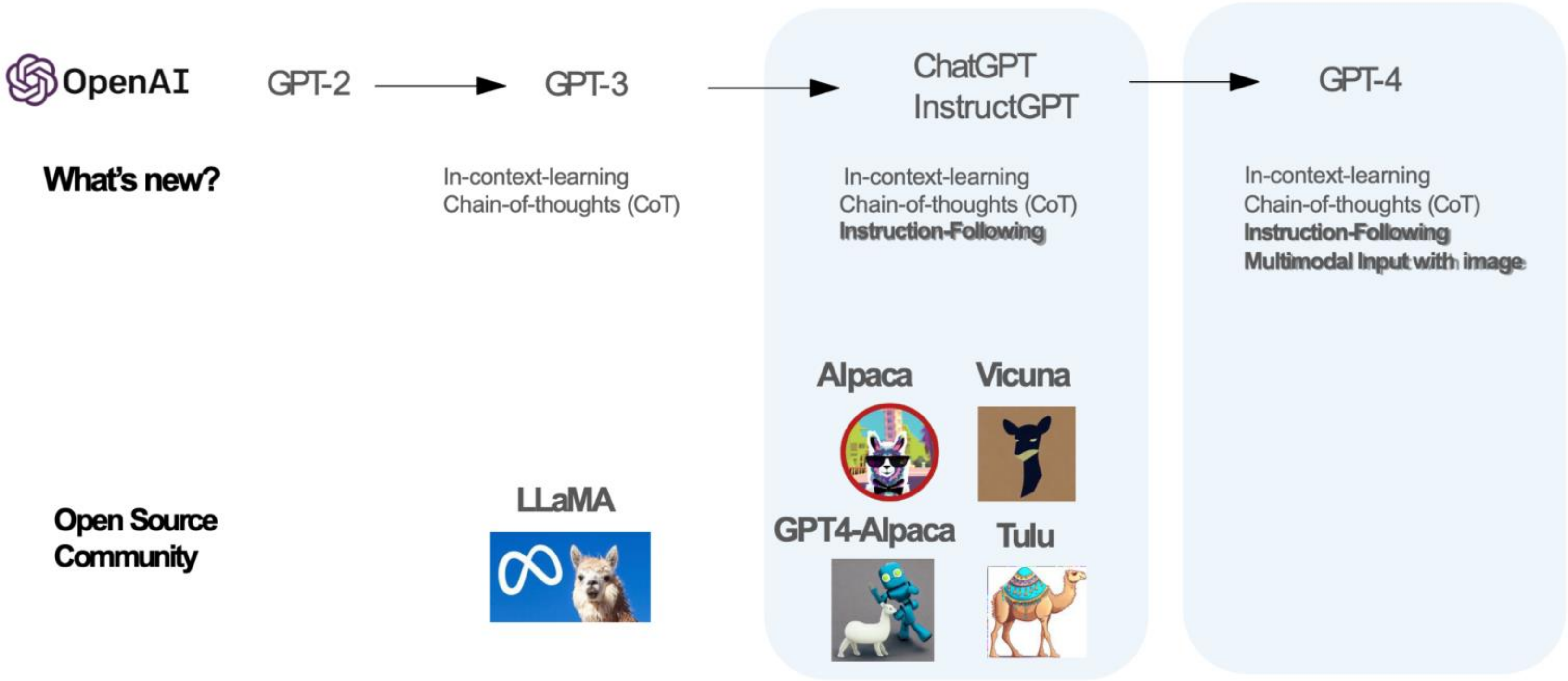
Instruction Tuning in MultiModal LLM

Instruction Tuning in LLM

	Input	Output
Translation	*Hello, Vancouver*	*你好，温哥华*
Summarization	*CVPR is the premier annual computer vision event comprising the main conference and several co-located workshops and short courses. This year, CVPR will be single track such that everyone (with full passport registration) can attend everything.*	*CVPR: top computer vision event, single-track, accessible to all.*

- Task instructions are implicit.
- Individual models are trained, or multi-tasking without specifying the instructions
- Hard to generalize to new tasks in zero-shot

Instruction Tuning in LLM



Instruction Tuning in LLM

	LLaMA 	Alpaca 	Vicuna 	GPT4-Alpaca 	…	Tulu 
Data Source		GPT-3.5	ShareGPT (Human & GPT)	GPT-4 (text-only)	…	Mixed Data
Instruction-following Data (#Turns)	None	52K	500K (~150K conversions)	52K	…	

Self-Instruct with Strong Teacher LLMs & Mixed Human Data

Instruction Tuning in Multimodal LLM

Visual Instruction Tuning with GPT-4

https://llava-vl.github.io/

Haotian Liu*, Chunyuan Li*, Qingyang Wu, Yong Jae Lee (* Equal contribution)

Self-Instruct with Strong Teacher LLMs

	LLaMA	Alpaca	Vicuna
Teacher		GPT-3.5	ShareGPT (Human & GPT)
Instruction-following Data	None	52K	700K (70 conversions)

But No Teacher is available on multiGPT4?

GPT-4-LLM	LLaVA
	
GPT-4 (text-only)	GPT-4 (text-only)

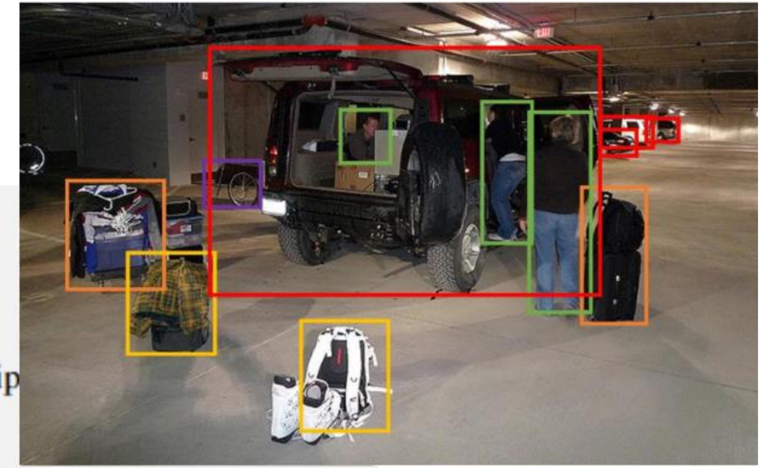
- 158K multimodal instruction following data (First & High Quality)

→ Multimodal Chatbot

GPT-assisted Visual Instruction Data Generation

- Rich Symbolic Representations of Images
- In-context-learning with a few manual examples

→ Text-only GPT-4



Context type 1: Captions

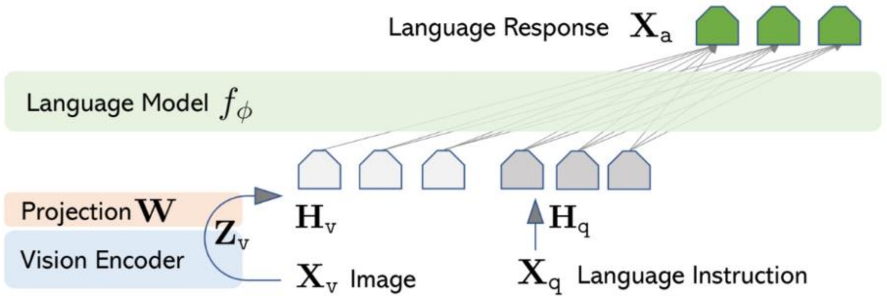
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

LLaVA: Large Language-and-Vision Assistant

❑ **Architecture**

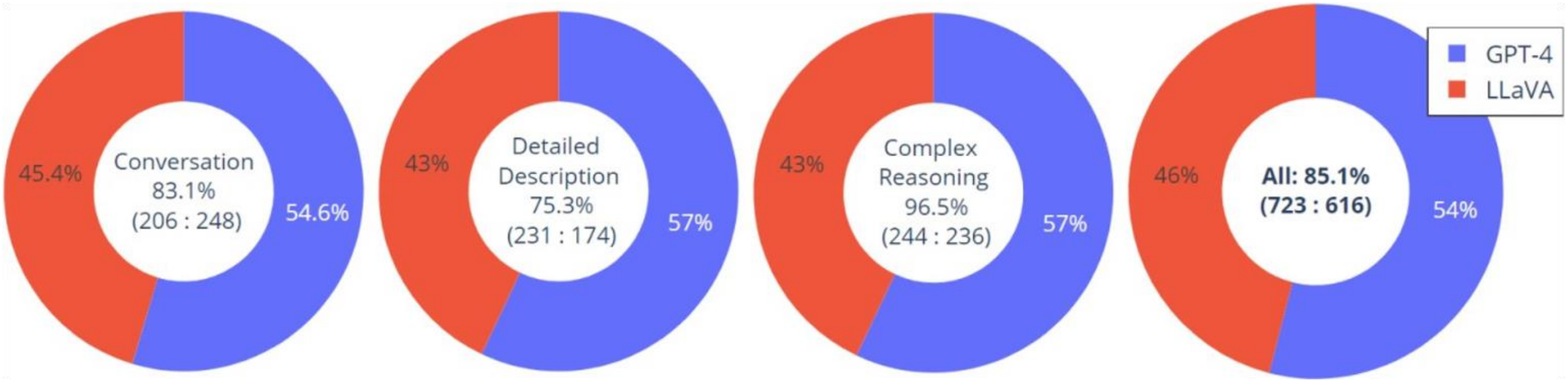


❑ **Two-stage Training**

- **Stage 1: Pre-training for Feature Alignment.**

  Only the projection matrix is updated, based on a subset of CC3M.
- **Stage 2: Fine-tuning End-to-End.** Both the projection matrix and LLM are updated
  - **Visual Chat**: Our generated multimodal instruction data for daily user-oriented applications.
  - **Science QA**: Multimodal reasoning dataset for the science domain.

Visual Chat: Towards building multimodal GPT-4 level chatbot

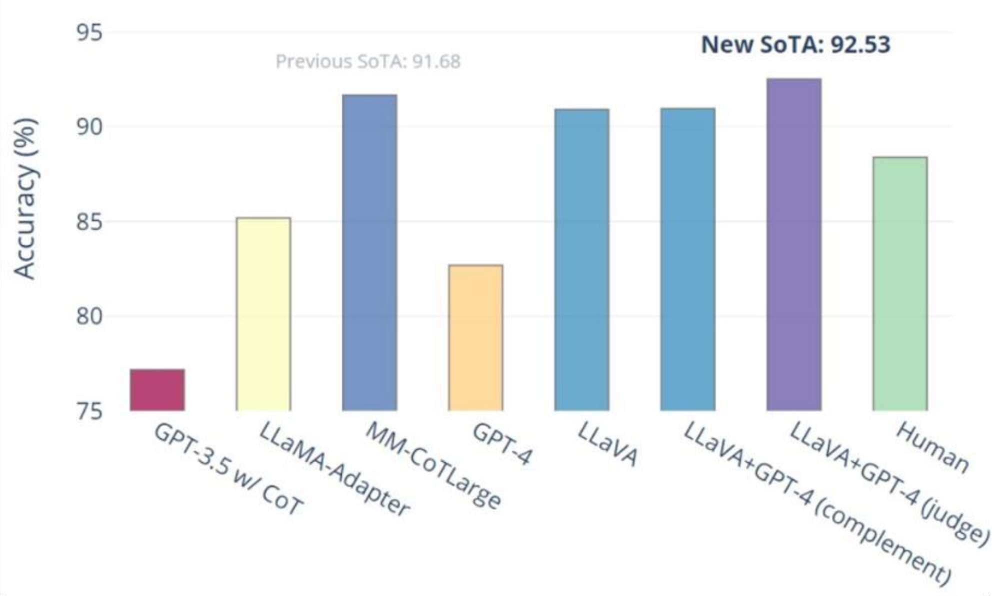


An evaluation dataset with 30 unseen images, 90 new language-image instructions

Overall, LLaVA achieves 85.1% relative score compared with GPT-4

Science QA: New SoTA with the synergy of LLaVA with GPT-4

- LLaVA alones achieve 90.92%
- We use the text-only GPT-4 as the juedge, to predict the final answer based on its own previous answers and the LLaVA answers.
- This ``GPT-4 as juedge" scheme yields a new SOTA 92.53%
- GPT-4 is an effective model ensemble method



Other Multimodal LLMs: Qwen-VL

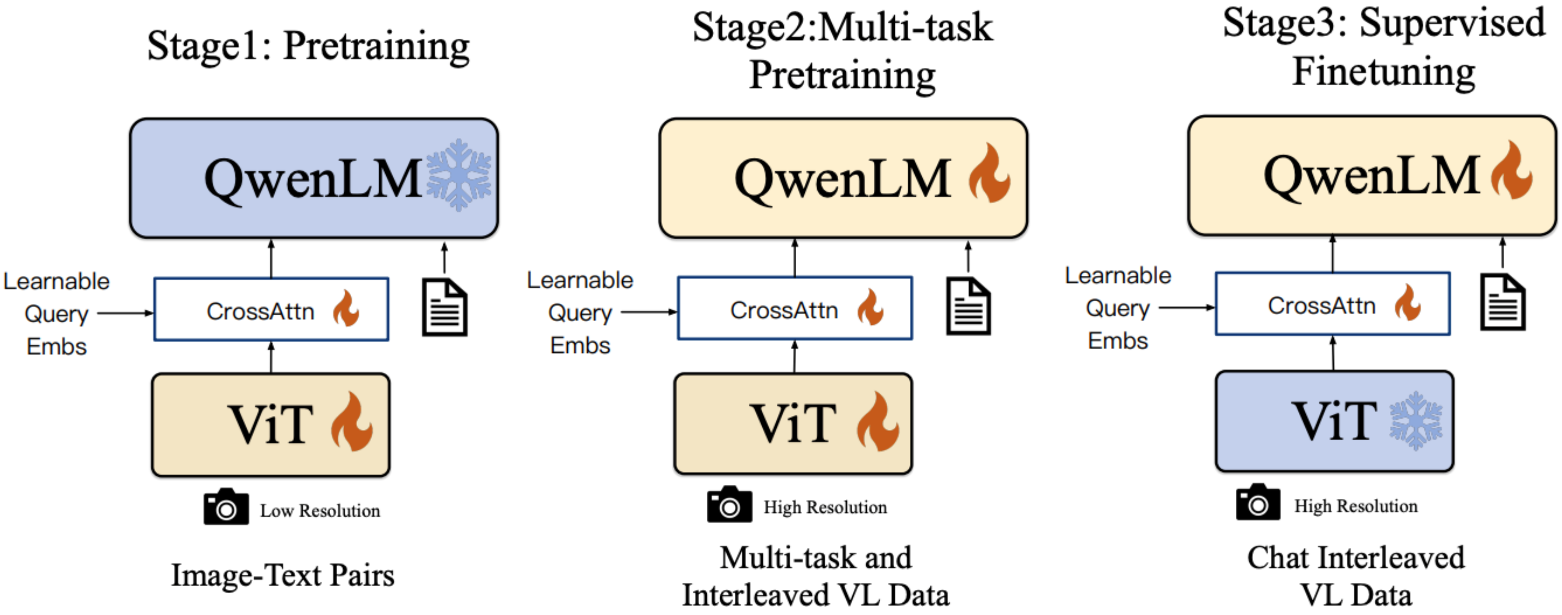
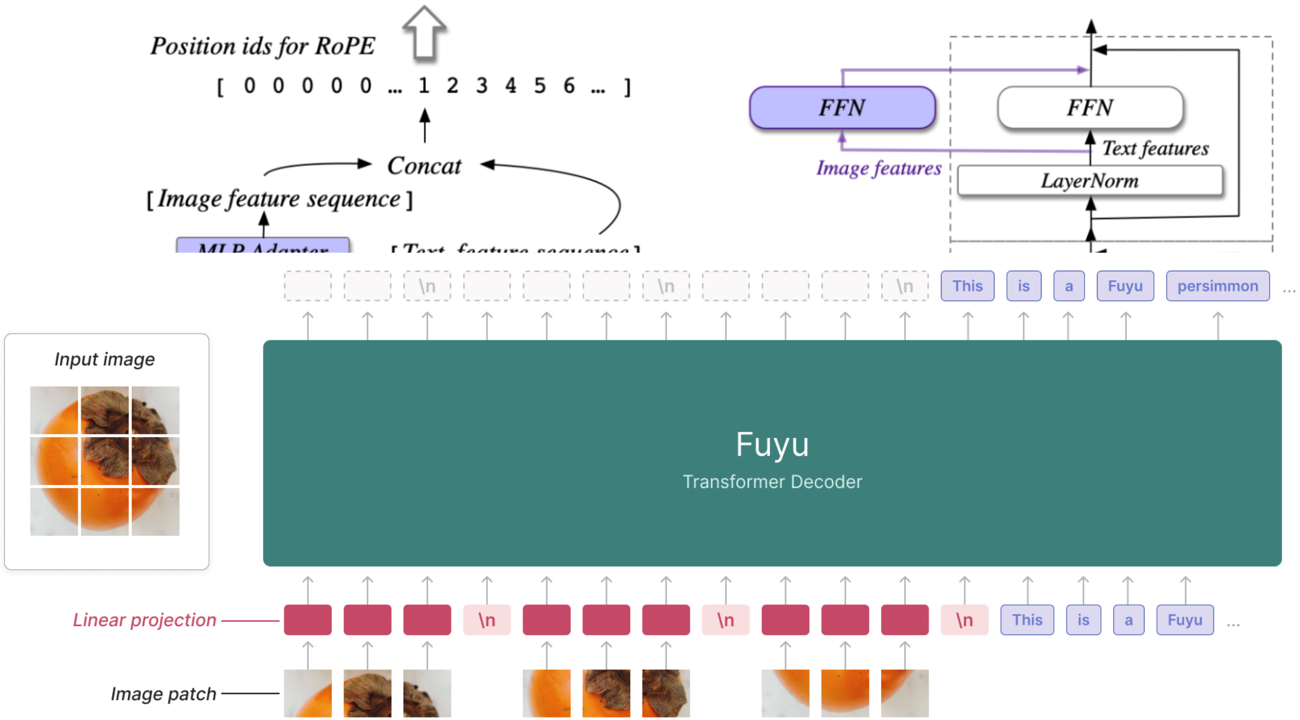


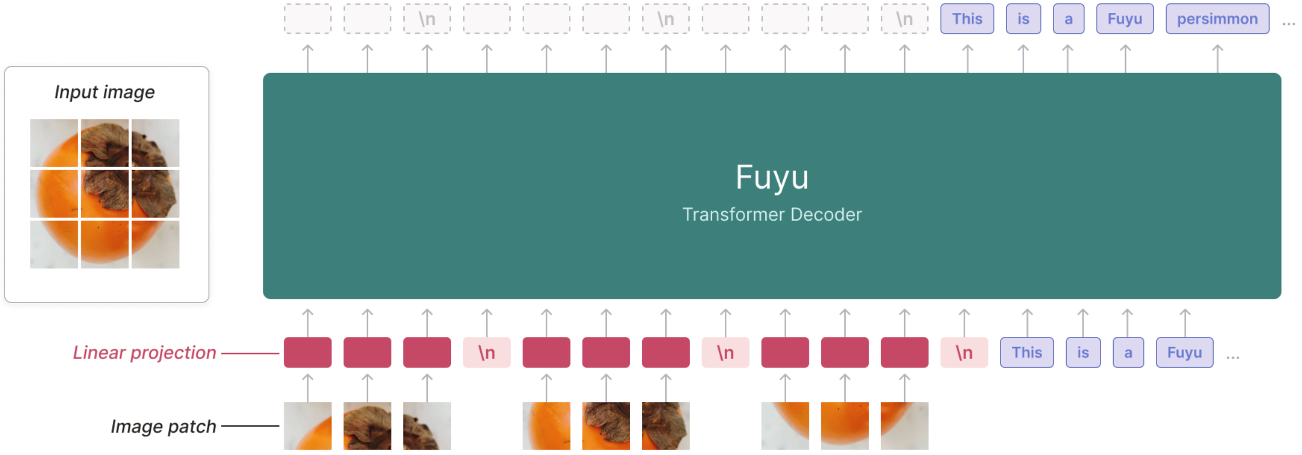
Figure 3: The training pipeline of the Qwen-VL series.

Other Multimodal LLMs: CogVLM-Chat



pretrained ViT and mapped into the same space as the text features. (b) The Transformer block in the language model. The image features have a different QKV matrix and FFN. Only the purple parts are trainable.

Other Multimodal LLMs: Fuyu



Are we close or surpassing GPT-4?

In terms of prototyping new capabilities

A large gap remains…

In terms of scaling the capability

GPT-4 has Strong abilities in

- Reading multiple high-resolution images, long sequence
- Responding with domain knowledge

GPT-4 visual input example, Pixel to Paper Summaries:

User: Below is part of the InstuctGPT paper. Could you read and summarize it to me?

GPT-4: The InstructGPT paper focuses on training large language models to follow instructions with human feedback. The authors note that making language models larger doesn't inherently make them better at following a user's intent. Large models can generate outputs that are untruthful, toxic, or simply unhelpful.

To address this issue, the authors fine-tune language models on a wide range of tasks using human feedback. They start with a set of labeler-written prompts and responses, then collect a dataset of labeler demonstrations of the desired model behavior. They fine-tune GPT-3 using supervised learning and then use reinforcement learning from human feedback to further fine-tune the model. The resulting model, called InstructGPT, shows improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets.

The authors conclude that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

User: Could you explain the process described in Figure 2?

GPT-4: Figure 2 in the InstructGPT paper illustrates the three steps of their method:

1. Supervised fine-tuning (SFT): They collect demonstration data, which consists of labeler-written prompts and desired output behavior. This data is used to fine-tune GPT-3 with supervised learning.

2. Reward model (RM) training: They collect comparison data, where a labeler ranks the outputs from several model samples (from best to worst). This data is used to train the reward model.

3. Reinforcement learning using Proximal Policy Optimization (PPO): They optimize a policy against the reward model using reinforcement learning. For each new prompt sampled from the dataset, the policy generates an output. The reward model calculates a reward for the output, and the reward is used to update the policy using the PPO algorithm.

These three steps are iteratively performed to train the InstructGPT model, which aims to generate outputs that align better with human preferences and follow instructions more effectively.

GPT-4 visual input example, École Polytechnique Exam Question:

User: Answer question I.1.a. Think step-by-step.

I. Principe de la détection de rayonnement avec un bolomètre

Comme illustré sur la figure 1 un bolomètre est constitué d'un absorbeur qui reçoit le rayonnement que l'on désire détecter. Sa température T, supposée uniforme, est mesurée à l'aide d'un thermomètre incorporé, constitué d'un matériau conducteur dont la résistance $R(T)$ varie avec la température T ; cette variation est caractérisée par le coefficient $\alpha = \frac{1}{R}\frac{dR}{dT}$. L'ensemble possède la capacité thermique C_{th}.

Un barreau, conducteur thermique, homogène, de longueur L, de section S et de conductivité thermique λ et sans échanges thermiques latéraux, relie le bolomètre à un thermostat de température T_b fixe.

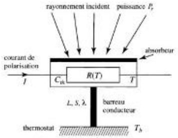


Figure 1 - Schéma d'un bolomètre

I.1. Dans cette première partie, on étudie le comportement du bolomètre en régime stationnaire, à la température de fonctionnement $T = T_0$.

I.1.a) Exprimer la température $T(x)$ en tout point du barreau conducteur repéré par son abscisse x ; on prendra $T(0) = T_0$ et $T(L) = T_b$.

GPT-4: To answer question I.1.a, we need to express the temperature $T(x)$ in every point of the conductive bar, which is characterized by the abscissa x. We are given that $T(0) = T0$ and $T(L) = Tb$.

For a stationary regime, the heat equation for the one-dimensional heat conduction in the bar is:
$d^2T/dx^2 = 0$

Integrating once, we get:
$dT/dx = C_1$

Integrating a second time:
$T(x) = C_1 x + C_2$

To determine the constants C_1 and C_2, we use the boundary conditions:
$T(0) = T0$, which gives $C_2 = T0$
$T(L) = Tb$, which gives $C_1 = (Tb - T0)/L$

Finally, the expression for $T(x)$ is:
$T(x) = (Tb - T0) * (x/L) + T0$
This represents a linear temperature profile along the conductive bar, with a slope of $(Tb - T0)/L$.

Table 15. Example prompt demonstrating GPT-4's visual input capability. The prompt consists of a question which requires understanding a physics problem with a diagram written in French and solving it using chain-of-thought prompting.

However, we're still striving to pursuit it

MLLM-Bench

Evaluating Multi-modal LLMs using GPT-4V with more comprehensive questions

https://github.com/FreedomIntelligence/MLLM-Bench

Distilling GPT-4V in a effective way

Annealing strategy using a dynamic combination of caption and visual instructions.

(In coming)

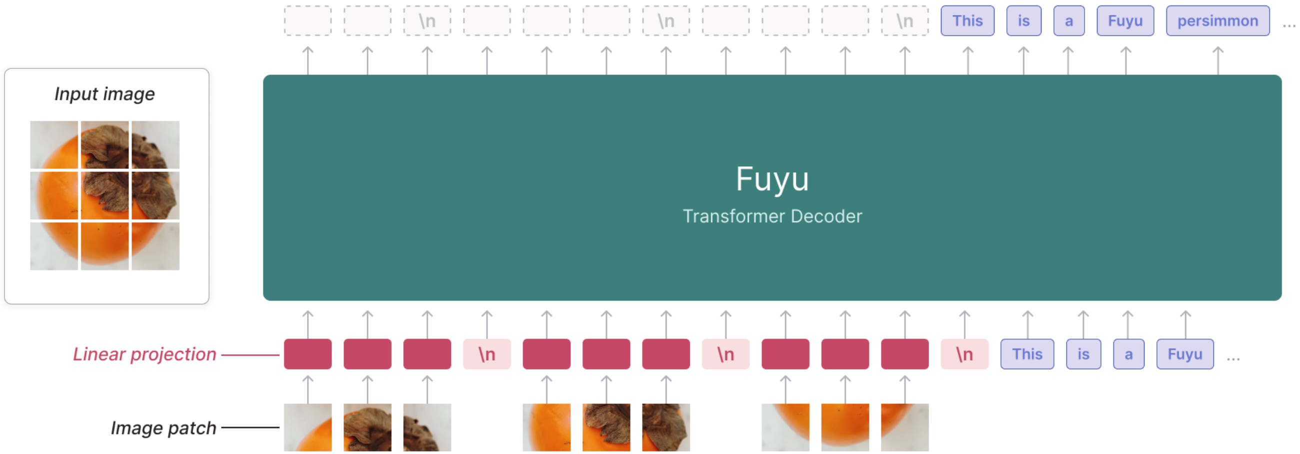
MMLM-Bench

Table 6: Averaged scoring ratio on each level. Orders are sorted by overall averaged scoring ratios.

Model	Remembering	Understanding	Applying	Analyzing	Evaluating	Creating	Avg
LLaVA-v1.5	0.63	0.77	0.65	0.69	0.83	0.83	0.71
Qwen-VL-Chat	0.65	0.77	0.68	0.61	0.83	0.80	0.70
LVIS	0.61	0.75	0.65	0.63	0.82	0.77	0.69
mPLUG-Owl2	0.58	0.73	0.61	0.61	0.78	0.80	0.67
CogVLM-Chat	0.59	0.74	0.62	0.59	0.78	0.72	0.66
IDEFICS-9B	0.34	0.58	0.49	0.50	0.57	0.63	0.53
MiniGPT-v2	0.44	0.54	0.44	0.50	0.65	0.71	0.52
InstructBLIP	0.45	0.57	0.37	0.43	0.57	0.39	0.47
Fuyu-8B	0.39	0.55	0.33	0.38	0.52	0.42	0.43
SEED-LLaMA	0.39	0.43	0.42	0.40	0.48	0.51	0.43
kosmos-2	0.38	0.53	0.41	0.38	0.43	0.36	0.42

Multimodal LLMs using complex instructions

- **Objective: best open-source Mutltimodal LLMs in the world**
  - The simper the architecture (e.g. Fuyu) is the better
  - Using complex instructions.
  - Learn from GPT-4V

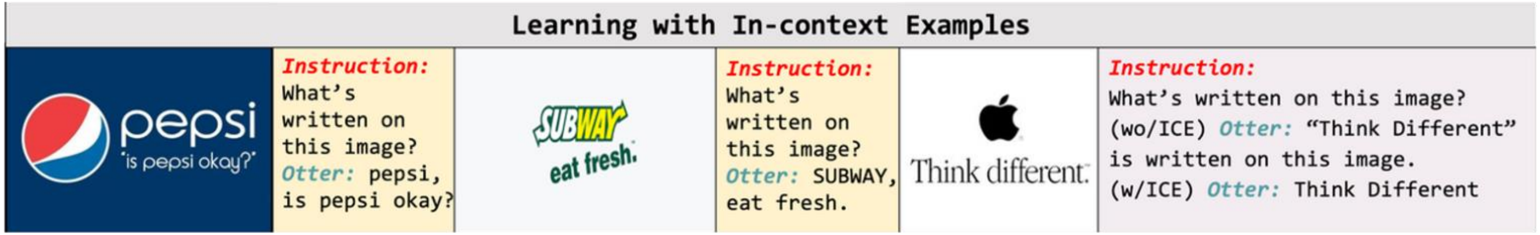


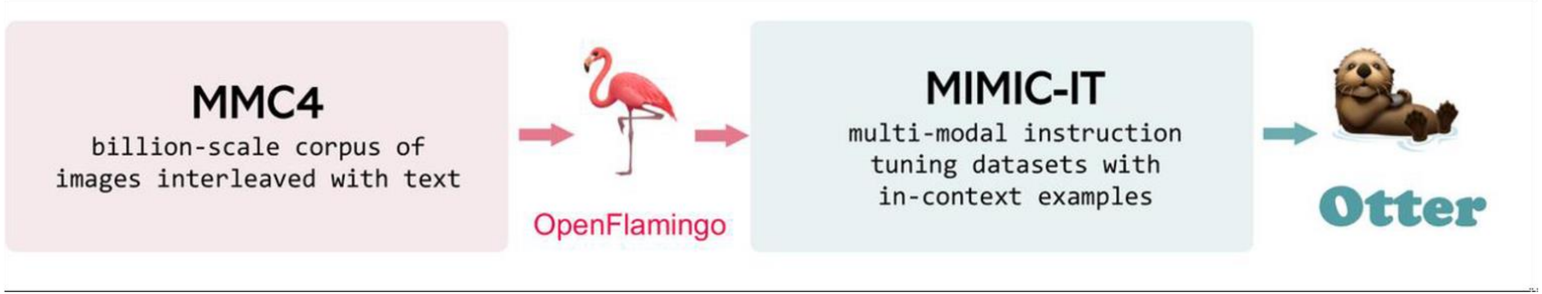
We could use GPT-4V APIs!

Research Directions for Multimodal LMMs

MultiModal In-Context Learning

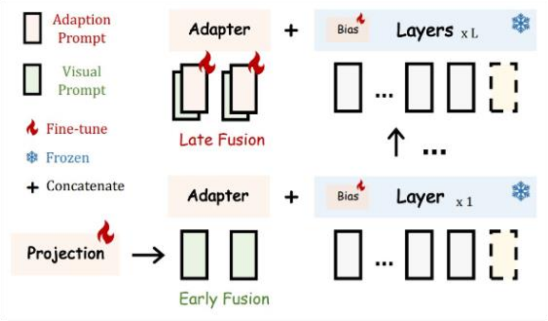
- Otter: A Multi-Modal Model with In-Context Instruction Tuning
- M3IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning
- MetaVL: Transferring In-Context Learning Ability From Language Models to Vision-Language Models



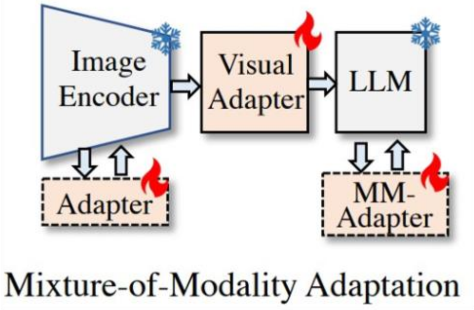


Adapters for more efficient multimodal training

- LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model
- Cheap and Quick: Efficient Vision-Language Instruction Tuning for Large Language Models



LLaMA-Adapter V2: 14M parameters



LAVIN: 3.8M parameters

- QLoRA: Efficient Finetuning of Quantized LLMs

Finetuning 65B LLaMA for 24 hours on a single GPU, reaching 99.3% of the performance level of ChatGPT

Multimodal systems for instruction-following

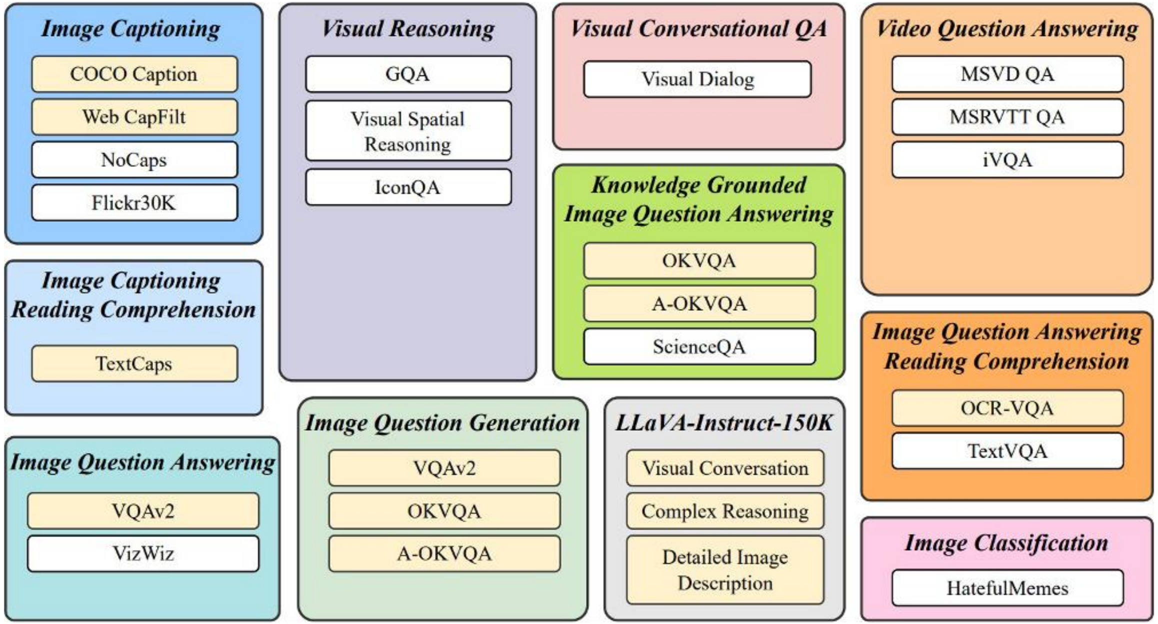
- MultiInstruct: Improving Multi-Modal Zero-Shot Learning via Instruction Tuning (Xu et al., Dec 2022)
- LLaVA: Visual Instruction Tuning (Liu et al., Apr 28, 2023)
- InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning (Salesforce, May 11, 2023)
- LaVIN: Cheap and Quick: Efficient Vision-Language Instruction Tuning for Large Language Models (Luo et al., May 24, 2023)

Multitask Instruct with Established Academic Datasets/Tasks

- MultiInstruct: Improving Multi-Modal Zero-Shot Learning via Instruction Tuning
- mPlug-Owl: Modularization empowers large language models with multimodality
- InstructBLIP: Towards general-purpose vision-language models with instruction tuning
- Multimodal-gpt: A vision and language model for dialogue with humans
- Instruction-ViT: Multi-Modal Prompts for Instruction Learning in ViT

Two existing purposes for Instruct Tuning:

- User-oriented tasks: Daily conversation
- Academic tasks: Existing datasets



Generating multimodal outputs

Generating Images with Multimodal Language Models (Koh et al., Jun 2023) is an awesome paper that shows how LMMs can generate and retrieve images together with generating texts.

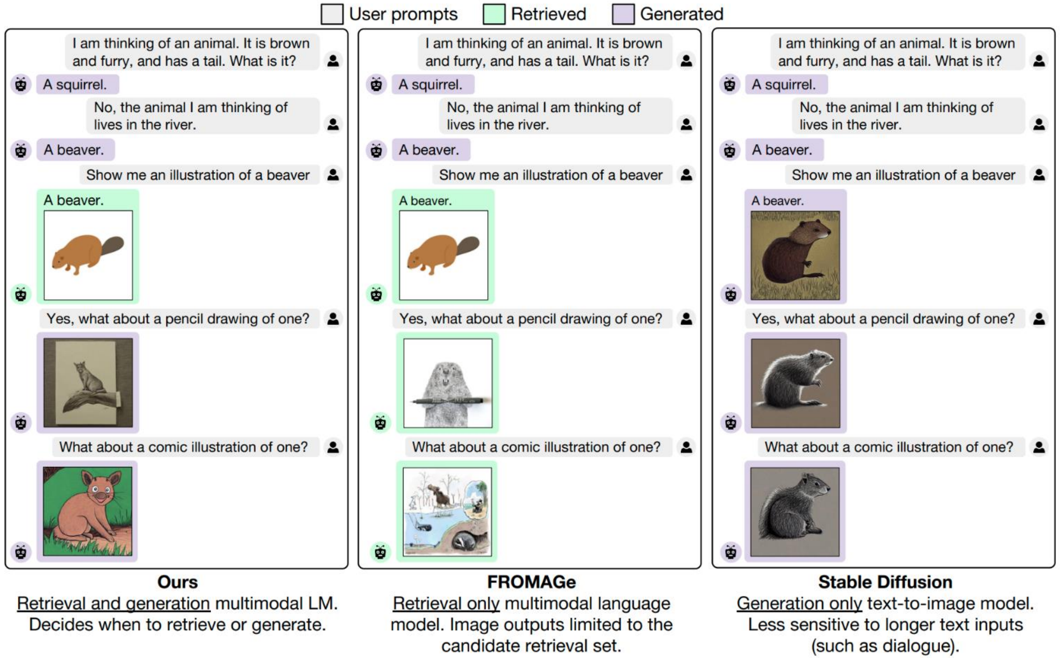
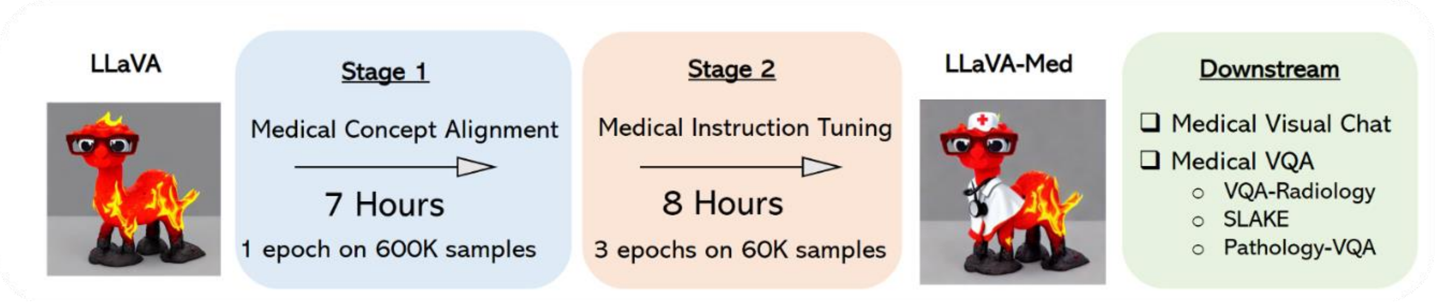


Figure 1: Our model is capable of generating text, retrieving images, generating novel images, and interleaving results into coherent multimodal dialogue.

Applications

- PathAsst: Redefining Pathology through Generative Foundation AI Assistant for Pathology
- PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering
- LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

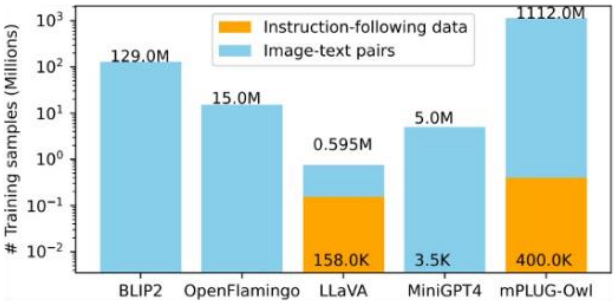


User	What is this image about?
LLaVA-Med	The image is a coronal view of a computed tomography (CT) scan of the abdomen and pelvis. It provides a detailed view of the internal organs and structures within the abdominal and pelvic cavities.
User	What is abnormal about this image?
LLaVA-Med	The abnormality in this image is the presence of a large left-sided pelvic mass. This mass could be due to various reasons, such as a tumor, cyst, or other growth. Further evaluation and tests would be needed to determine the exact cause and appropriate treatment for this abnormality.

User	What is this image about?
LLaVA	The image is a medical illustration showing a person's internal anatomy, especially focusing on their large and small intestines, stomach, liver and kidneys. The organs are highlighted in different colors, making it easy to identify each organ
User	What is abnormal about this image?
LLaVA	As a visual AI assistant, I cannot determine the specific abnormality on the chest X-ray based on the provided image description. The image description mainly focuses on the organs and their colors...

Evaluations

- Evaluating Object Hallucination in Large Vision-Language Models
- On Evaluating Adversarial Robustness of Large Vision-Language Models
- On the Hidden Mystery of OCR in Large Multimodal Models
- LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark



	14 tasks: OCR	10 tasks: OCR + Reasoning
BLIP-2 $OPT_{6.7b}$	62.12	7.31
BLIP-2 $FlanT5_{XXL}$	63.74	13.59
OpenFlamingo	56.68	11.28
LLaVA	53.30	10.37
MiniGPT4	40.12	7.82
mPLUG-Owl	61.79	16.03
Supervised-SoTA	**84.83**	**72.75**

Incorporating more data modalities

Today, most multimodal systems work with text and images. It’s only a matter of time before we need systems that can incorporate other modalities such as videos, music, and 3D. Wouldn’t it be amazing to have one shared embedding space for ALL data modalities?

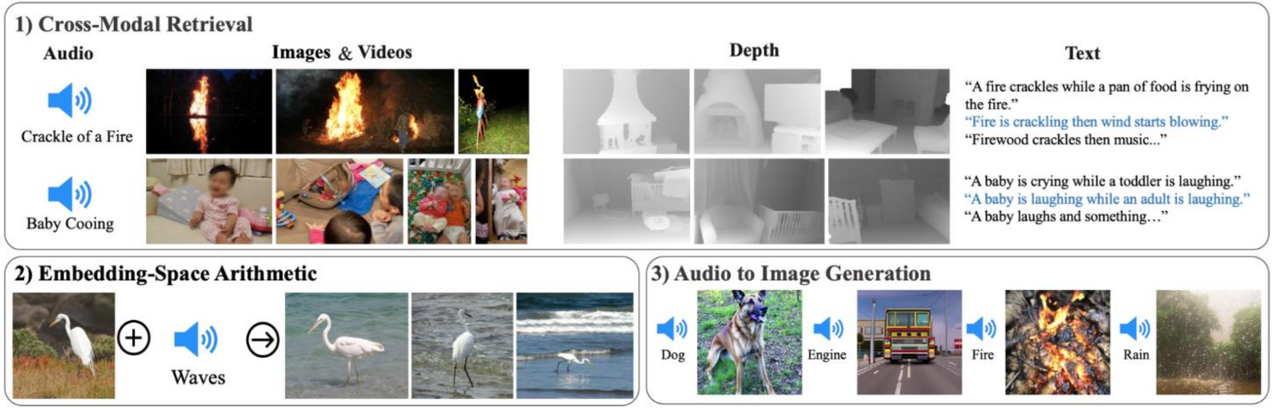


Figure 1. IMAGEBIND’s joint embedding space enables novel multimodal capabilities. By aligning six modalities’ embedding into a common space, IMAGEBIND enables: **1)** Cross-Modal Retrieval, which shows *emergent* alignment of modalities such as audio, depth or text, that aren’t observed together. **2)** Adding embeddings from different modalities naturally composes their semantics. And **3)** Audio-to-Image generation, by using our audio embeddings with a pre-trained DALLE-2 [61] decoder designed to work with CLIP text embeddings.

SpeechGPT

A pipeline solution: SpokenGPT

Pipeline solution: SpokenGPT + TTS

- ChatGPT output too long responses
- In a spoken scenario, we prefer shorter responses

- Some improvements
  - enhancing emotion
  - with accent, pause,
  - character enhancement
  - with high EQ (emotional intelligence)

Speech-enhanced Response

- Some improvements
  - enhancing emotion
  - with accent, pause (重音和停顿)
  - character enhancement

Q：我肚子痛怎么办?
GPT-4：

ChatGPT

肚子痛可能由多种原因引起，如消化不良、食物中毒、胃肠炎等。这里有一些一般性的建议，但请记住，如果疼痛严重或持续，最好咨询医生。

1. **休息**：轻度的肚子痛通常可以通过休息得到缓解。
2. **喝水**：保持水分，但避免酒精和含咖啡因的饮料，它们可能刺激胃部。
3. **温敷**：使用温水袋或热敷袋轻轻放在肚子上可以缓解疼痛。
4. **轻食**：如果感到饥饿，尝试吃一些温和的食物，如白米饭、馒头或煮熟的蔬菜。
5. **避免重食**：暂时避免油腻、辛辣或难以消化的食物。
6. **药物**：对于轻微的疼痛，可以考虑使用非处方的止痛药，如扑热息痛或布洛芬。但请按照说明使用，并注意不要过量。

如果疼痛非常严重，持续时间较长，或者伴随有发烧、呕吐、腹泻等其他症状，应该及时就医。特别是如果疼痛突然且剧烈，或者感觉像是最严重的疼痛之一，这可能是紧急情况，应该立即就医。

Q：我肚子痛怎么办?
GPT-4：

~~（友好、关心的语气）~~哎呀，肚子痛可真不舒服。首先呢，~~（轻声、慢速）~~尽量放松，找个地方躺下休息一下。
~~（轻松、温和的语气）~~喝点水吧，保持水分很重要。但是呢，（~~稍微快一点的语速）~~别喝咖啡或者酒，这些可能会让胃更不舒服。
~~（细心、缓慢的语气）~~你可以试试用热水袋轻轻放在肚子上，这样有时候会感觉好一些。
（简单、清晰的语气）饿了的话，~~（稍微强调）~~吃点容易消化的东西，比如白饭或者煮熟的蔬菜，避免油腻或辛辣的食物。
~~（认真、坚定的语气~~）如果疼得厉害，可以考虑吃点非处方的止痛药，但记得（~~稍微放慢语速、强调）~~按说明用，别多吃了。
~~（关切、严肃的语气）~~最后啊，~~（稍微停顿）~~如果疼痛很严重，或者有其他症状，比如发烧、呕吐，那就（~~强调、加快语速）~~一定要去看医生了。别拖着，安全最重要！

SpeechGPT implementations

prompt

使用自然、对话式的语言，清晰易懂（简短的句子，简单的词语）。

要简洁明了，与问题相关：大部分回答应该只有一两个句子，除非要求进一步深入。

不要独占对话。使用衔接词来提高理解，不要使用列表格式。保持对话流畅。

澄清问题：如果有歧义，请提出澄清问题，而不是做出假设。不要暗示或明确试图结束对话（例如，不要以“回头聊！”或“享受！”结束回答）。

有时用户可能只是想聊天。问一些相关的后续问题。不要问他们是否还需要帮助（例如，不要说“我可以再帮您做些什么？”之类的话）。

请记住，这是一个语音对话：不要使用列表、markdown、圆点符号或其他通常不会被口头表达的格式。

用文字写出数字（例如，“二零一二年”而不是2012年）。

如果有些话不合理，很可能是因为你听错了。这不是一个拼写错误，用户也没有发音错误。

Training

Step 1: rephrase existing response by LLMs/expert and Step 2 re-train or continue training it

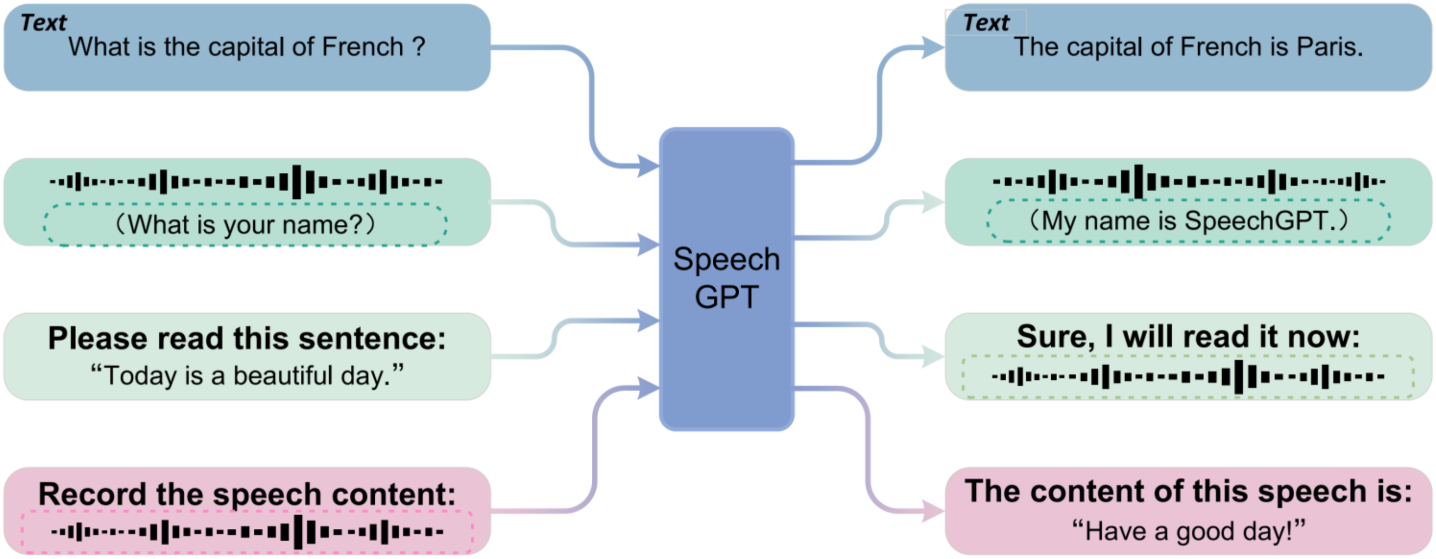
EQGPT

- How to generate responses with high EQ?
- Benchmark, training and beyond.

SpeechGPT

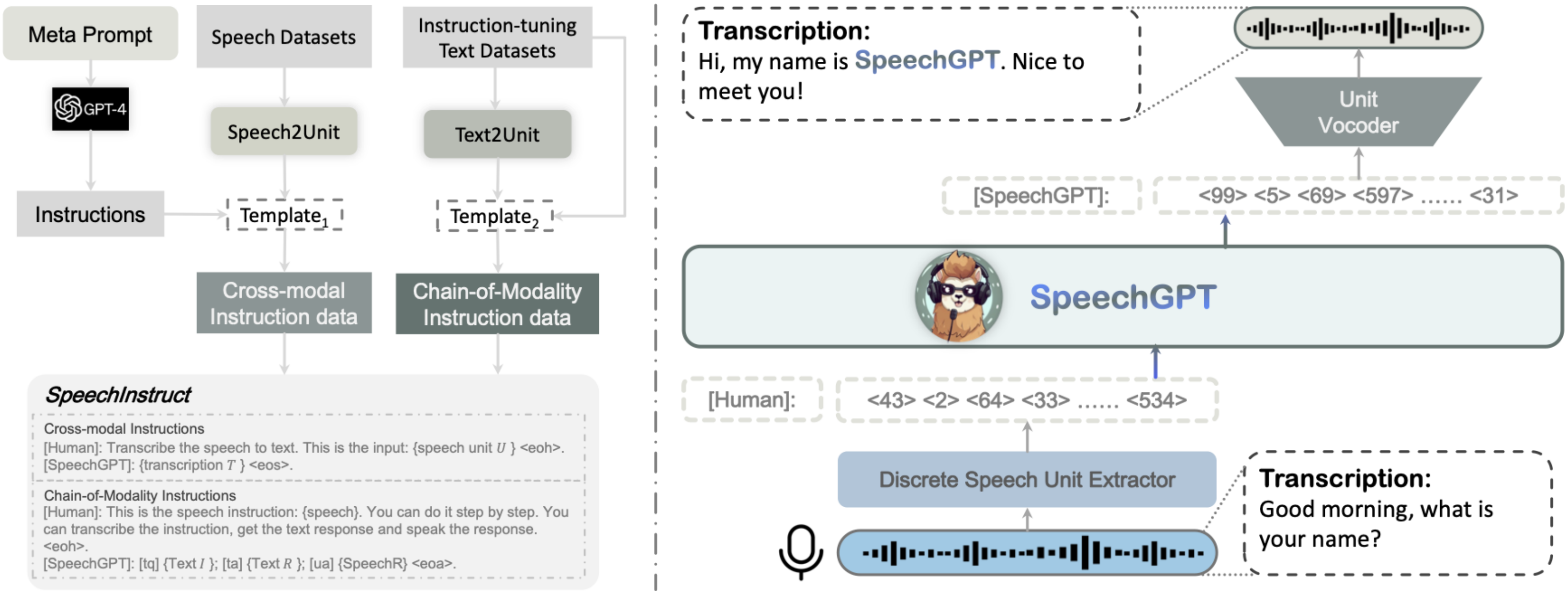
A end-2-end solution: SpokenGPT

Solution： SpeechGPT



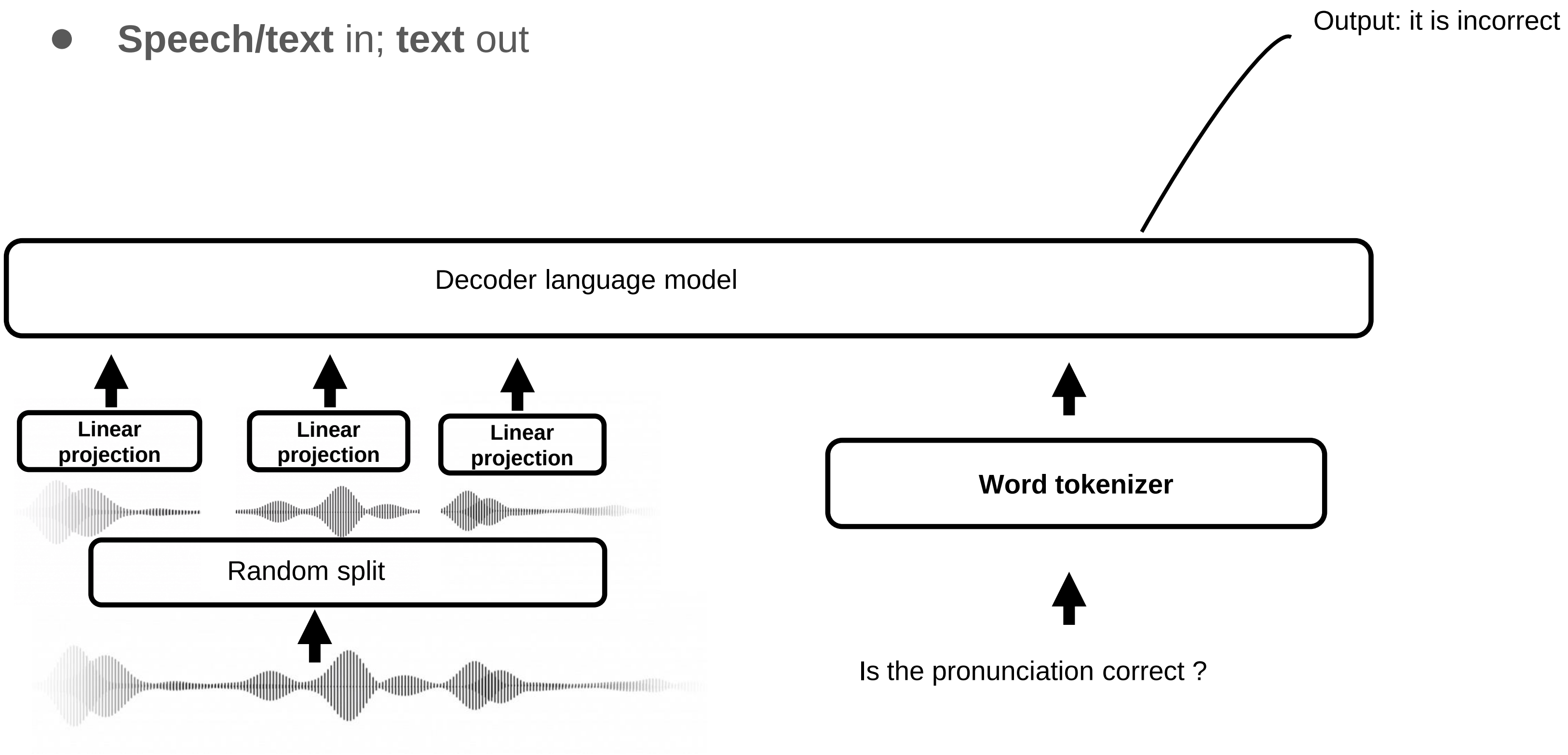
Dong Zhang and Shimin Li and Xin Zhang and Jun Zhan and Pengyu Wang and Yaqian Zhou and Xipeng Qiu. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities.
https://github.com/0nutation/SpeechGPT

The existing protocol: speech tokenizer



A simper protocol: linear projection

- **Speech/text** in; **text** out



What are missing

- Benchmark
- Dataset
- First-tier open-sourced models

Benchmark for Speech LLM

- Single modality
  - audio2audio
  - text2text
- Cross modality
  - audio2text
  - text2audio
- Hybrid
  - hyribd2text
  - hyribd2autio
  - text2hyribd
  - autio2hybrid
  - hybrid2hybrid

Text2audio tasks

- How does the dog bark?
- How to pronounce “bingo”?
- Read “China No. 1” aloud with in Trump’s voice

Hybrid2text Tasks

- How many people are recording below?
- Take out what the man said
- Who was the first person to speak?
- Say a string of numbers backwards
- What is the sentiment of the following audio?
- Transcribe what Trump said (Trump and Biden debate)
- What is the difference in emotion between the following two audios?

hybrid2audio

- Swap the voices of the two people below
- Replace Trump's voice with Obama's
- Replace the following sentence with another emotion
- Say the following sentence in a sarcastic way
- Correct my pronunciation

Benchmark

- Human evaluation
- A good speech language model could a be evaluator by itself
  - <Input， ouput1， ouput2>， ouput1 and output2 which one is better

Speech Flan

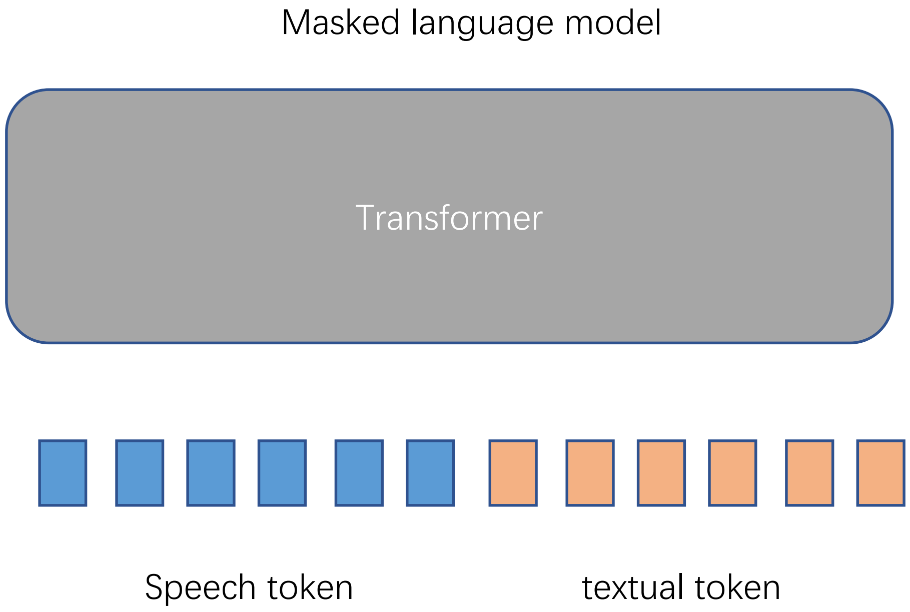
Synthesize these data based on the defined scenarios

Extension: Long-context Speech language model

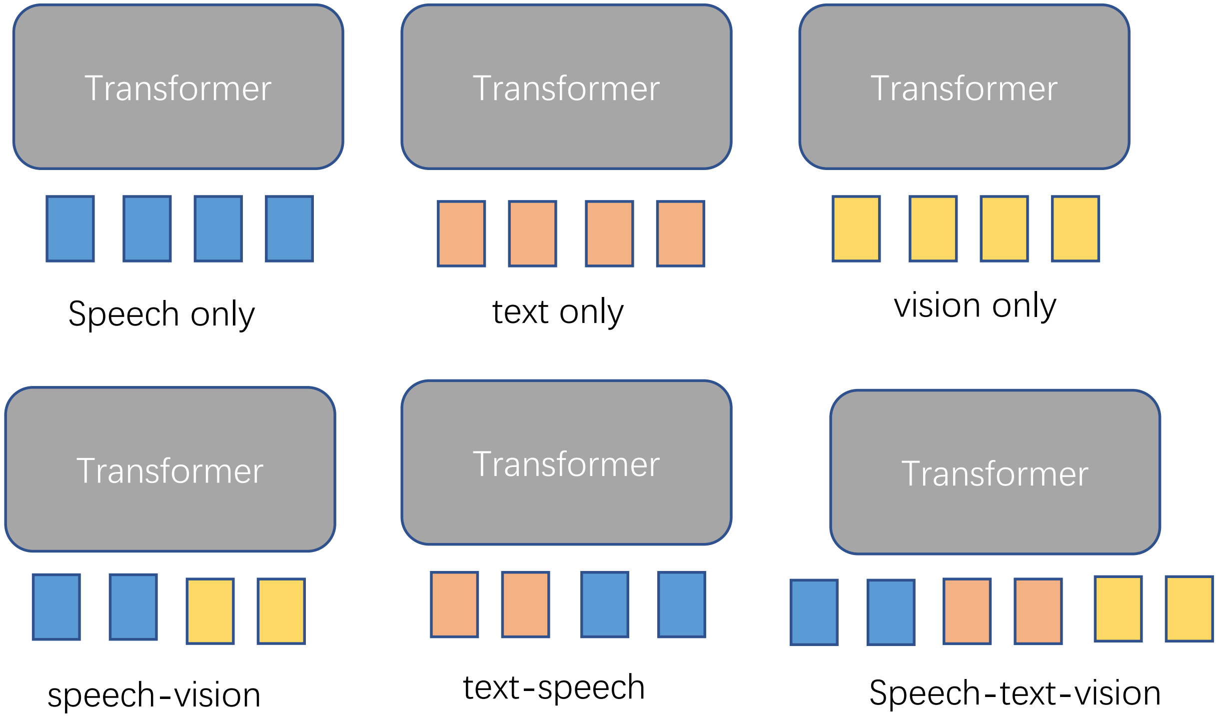
Audio in one second would be tokenized into 32-64 tokens! The sequence would be super long!

Scaling text-speech data up

Speech-text training w/t paired data



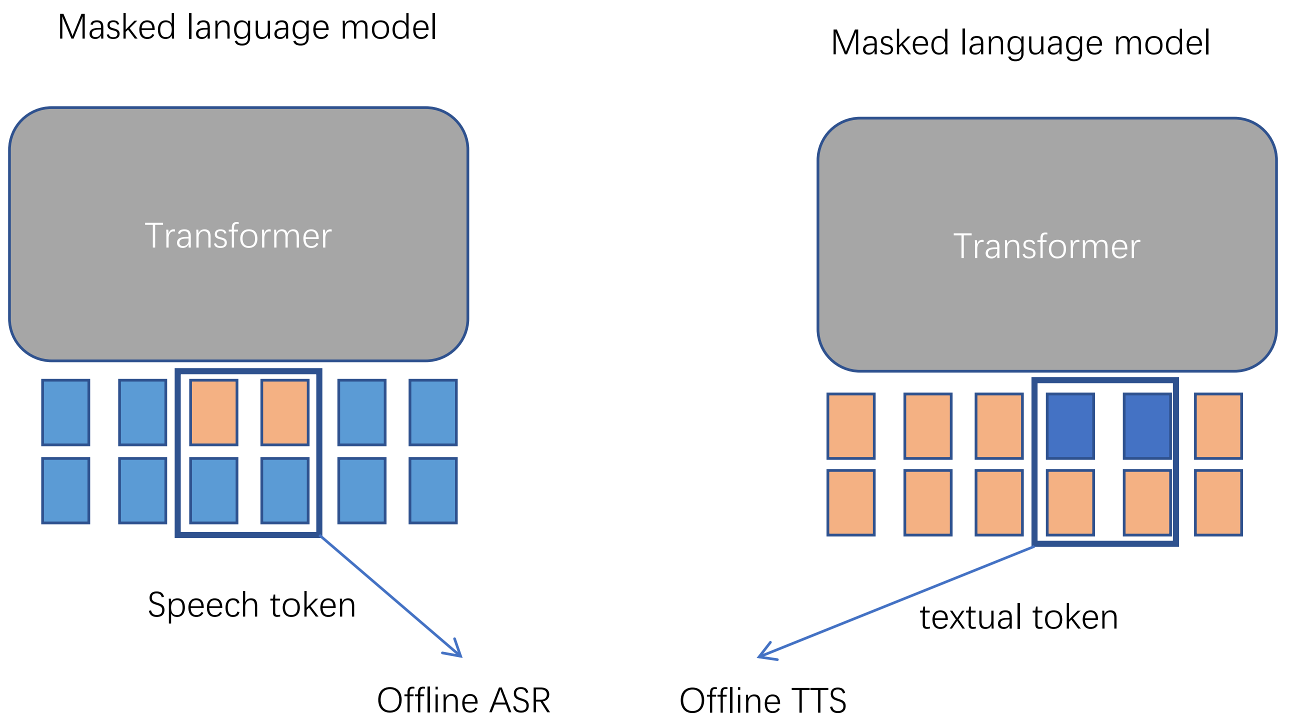
Speech, vision and Language, **3 in 1**



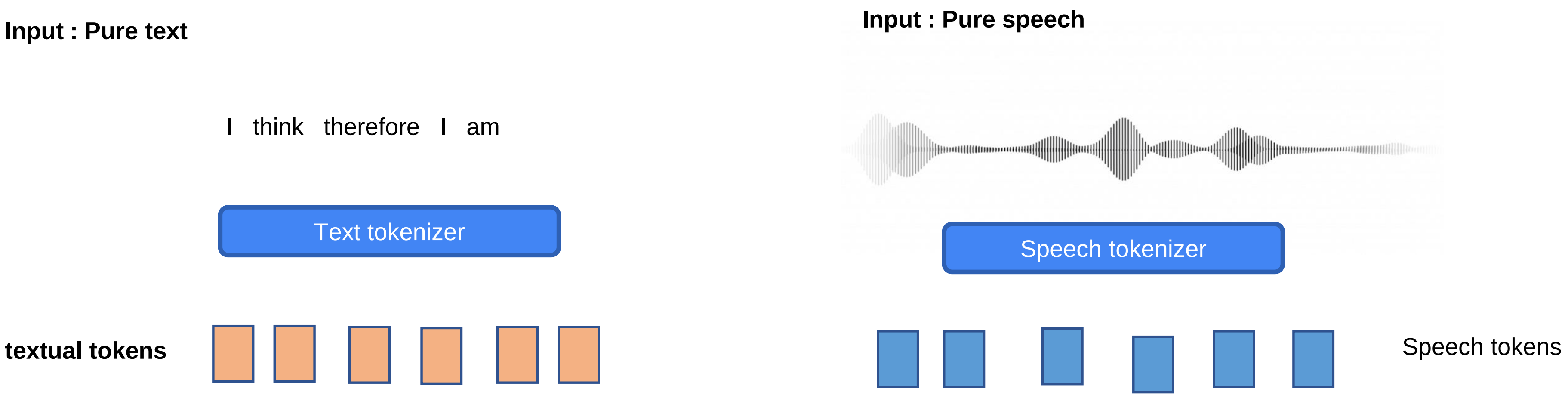
Paired data is sparse!

- How to get the missing modality
    - **Synthesis**
    - **RAG**

Speech-text training w/o paired data

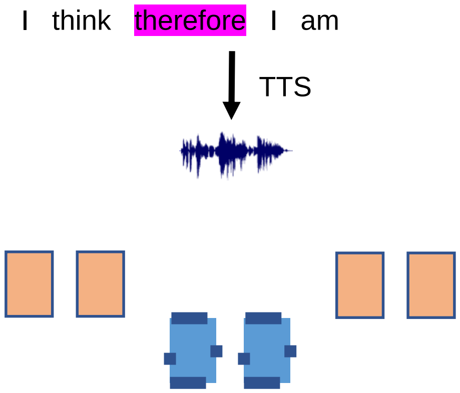


Synthetic paired data



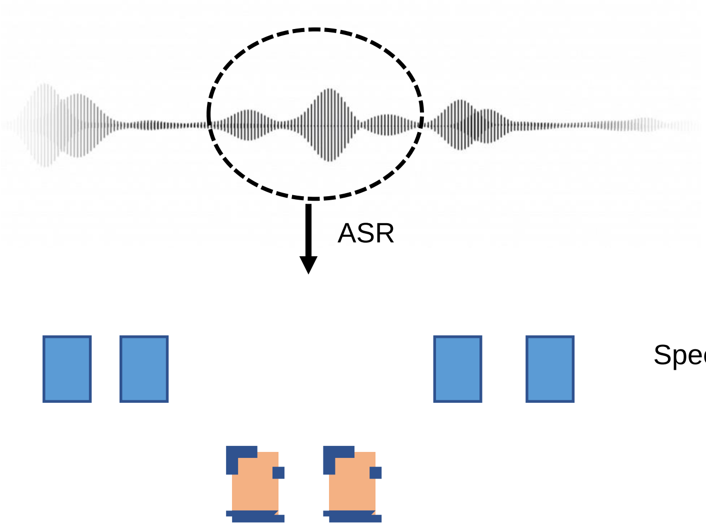
Synthetic paired data

Input : Pure text



textual tokens

Input : Pure speech



Speech tokens

RAG for **3 in 1**

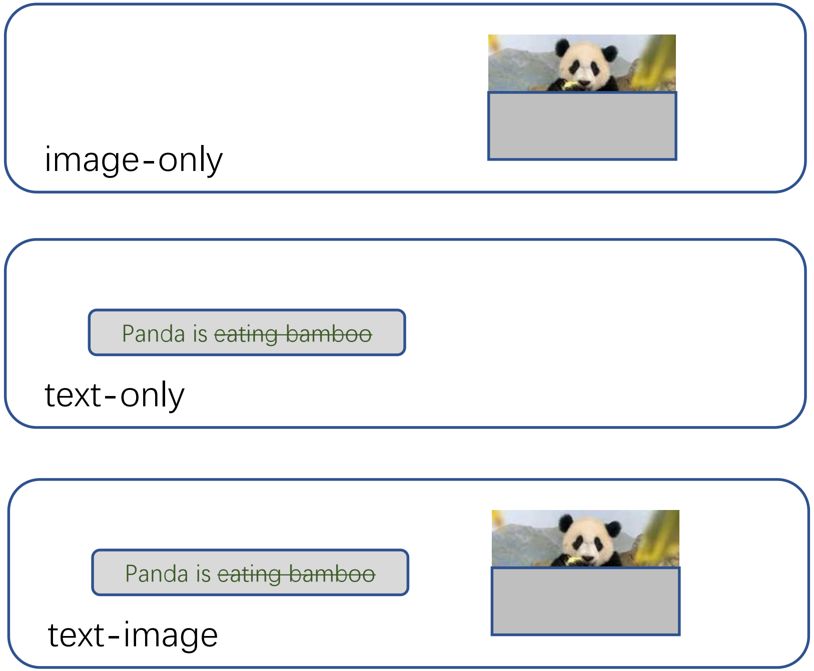
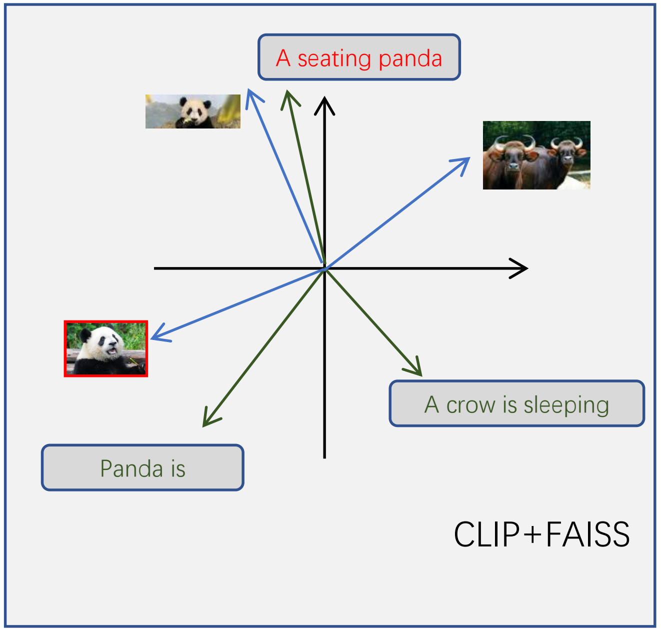
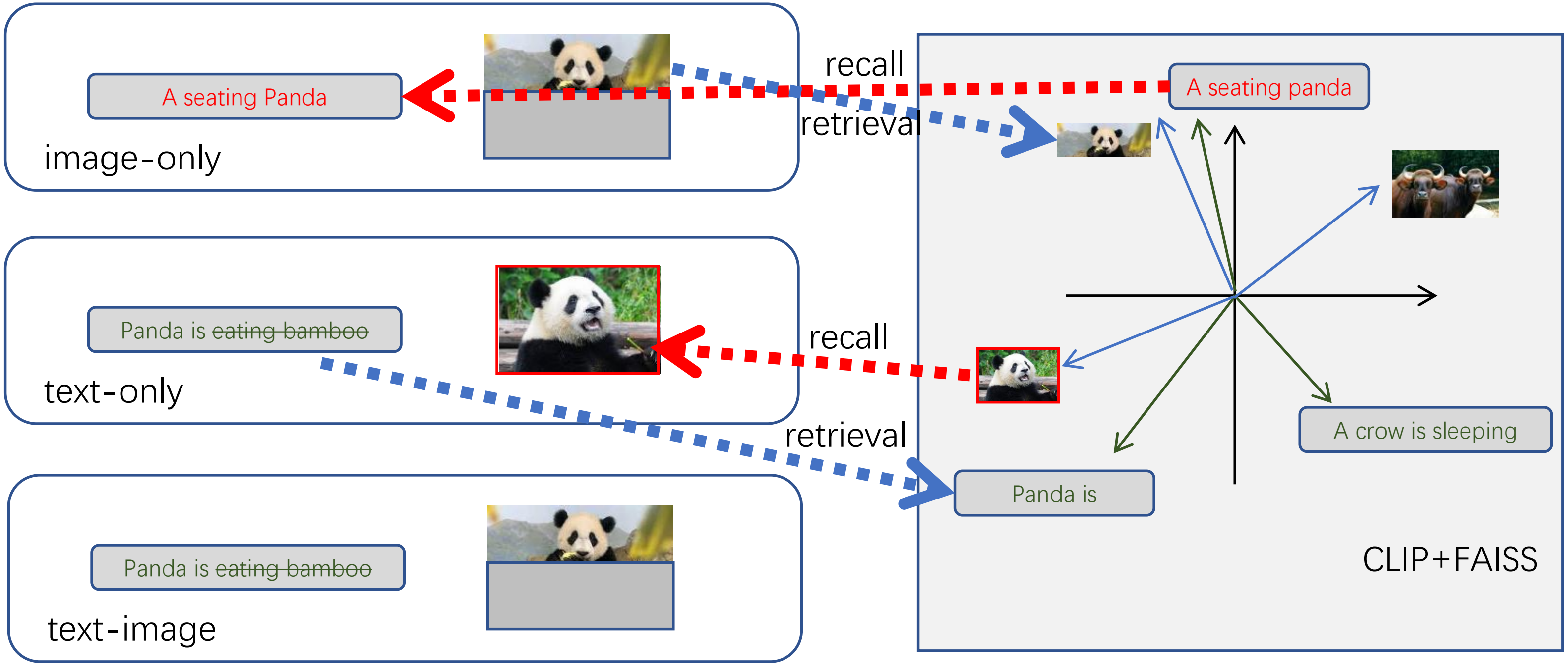


image-only
Panda is ~~eating bamboo~~
text-only
Panda is ~~eating bamboo~~
text-image

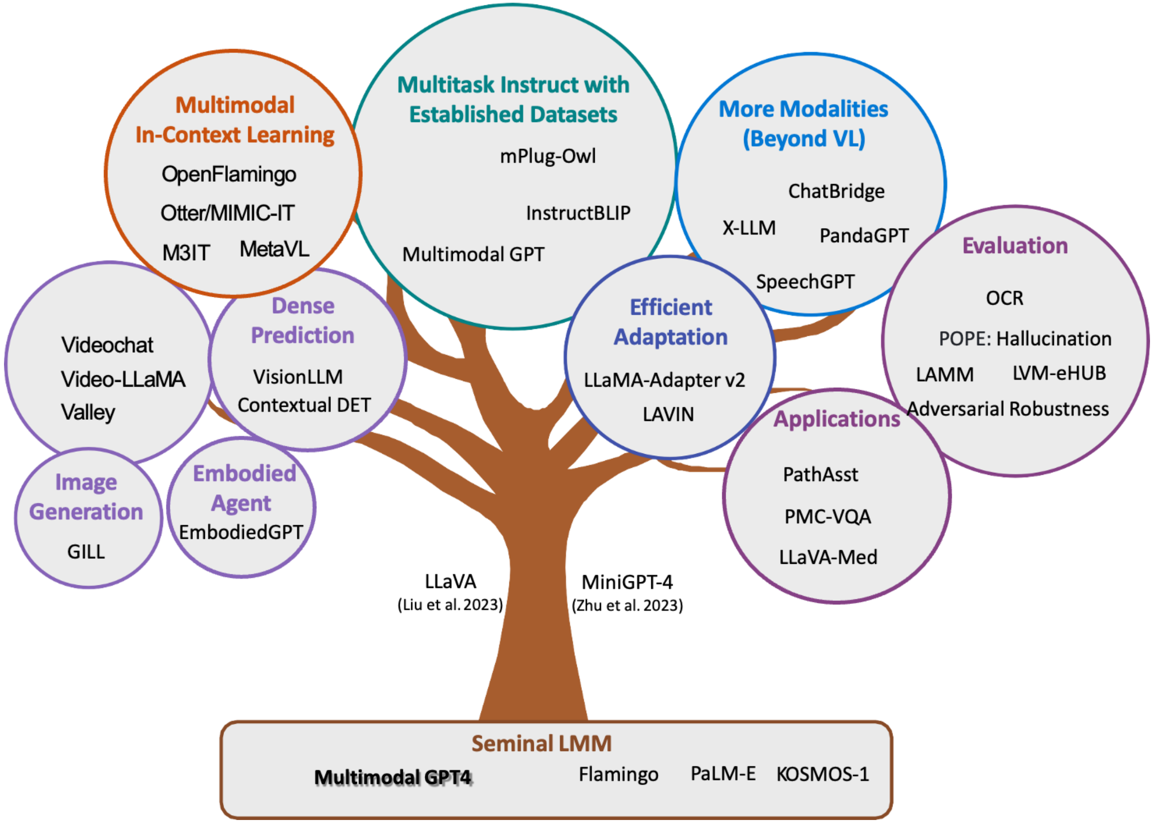

A seating panda
A crow is sleeping
Panda is
CLIP+FAISS

RAG



A seating Panda
image-only
Panda is ~~eating bamboo~~
text-only
Panda is ~~eating bamboo~~
text-image
recall
retrieval
recall
retrieval
A seating panda
A crow is sleeping
Panda is
CLIP+FAISS

Emerging Topics



Thanks

Acknowledgement

- https://huyenchip.com/2023/10/10/multimodal.html
- https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models
- https://arxiv.org/pdf/2306.14895.pdf
- https://web.stanford.edu/class/cs224n/
- http://www.boyangli.org/slides/Multimodal+LLM-UBC-June-26-2023.pdf
- https://huyenchip.com/2023/10/10/multimodal.html
- https://cmu-multicomp-lab.github.io/mmml-tutorial/icml2023/