

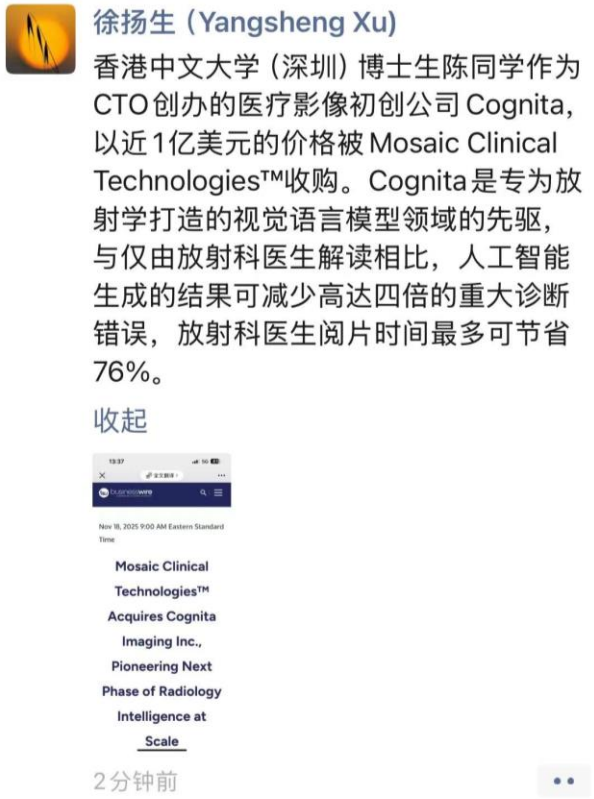
香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

CSC5051/MDS5110/CSC6052 : Natural Language Processing

Lecture 10: Multimodal LLMs

Fall 2026
Benyou Wang
School of Data Science

Before the lecture





Join the journey on “(multi-modal) LLMs for Society”

Contents

Foundations

why and how vision meets language

1 Why Multimodality?
Definition · modalities · why now · milestones

2 Connecting Vision and Language: CLIP and Flamingo
Contrastive pre-training · web data · Perceiver · gated cross-attention

Building multimodal LLMs

architecture and training

3 Multimodal LLMs: Architecture
Language as pivot · controller vs. joint model · encoders

4 Training Multimodal LLMs
Visual instruction tuning · caption data · multi-stage training

Beyond images

hearing, generating, acting, 3D

5 Speech Language Models
SSL units · neural codecs · interleaving · full duplex · omni

6 Generation and World Simulators
Unified generation · next-frame prediction · world simulators

7 Embodied Intelligence: VLA and World Models
VLA · System 1/2 · world models · WAMs · GPT-6 Astra · spatial

8 3D Understanding and Generation
Reconstruction · 3D generation · language-driven 3D · future

Outlook

what comes next

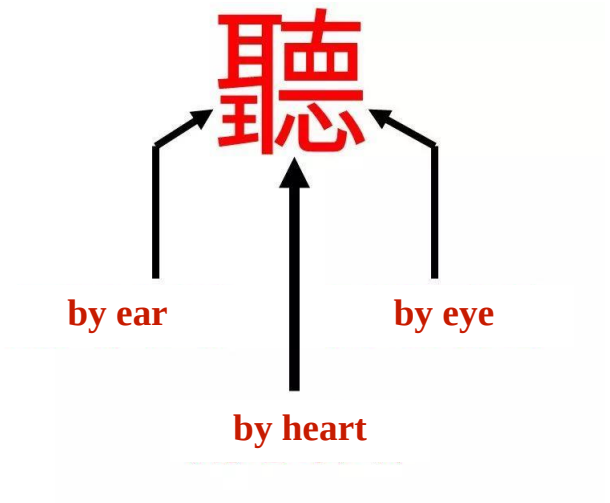
9 Research Directions and Outlook
Efficiency · evaluation · edge devices · 2026 trends

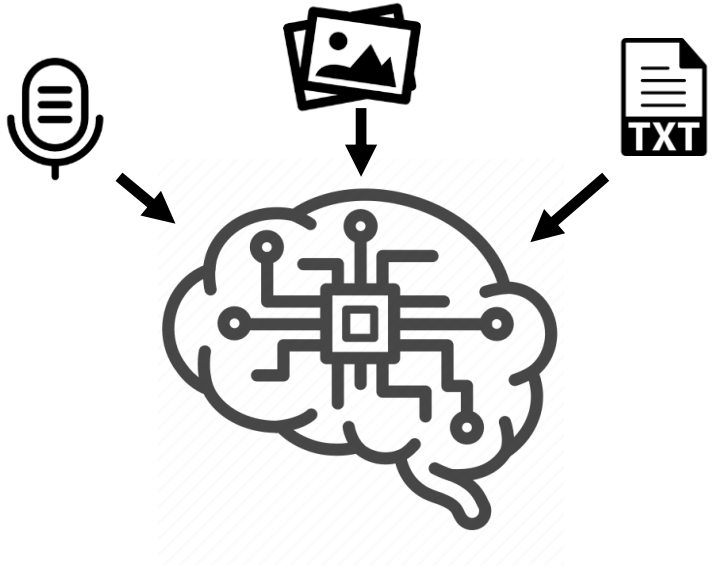


1 Why Multimodality?

Humans perceive the world through many senses at once. Text alone is a narrow window onto a world of images, sound and motion.

Even Listening Is Multimodal





The character 聽 (“to listen”) combines ear, eyes and heart: humans process multimodal information simultaneously

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

What is Multimodality?

Multimodal Behaviors and Signals

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

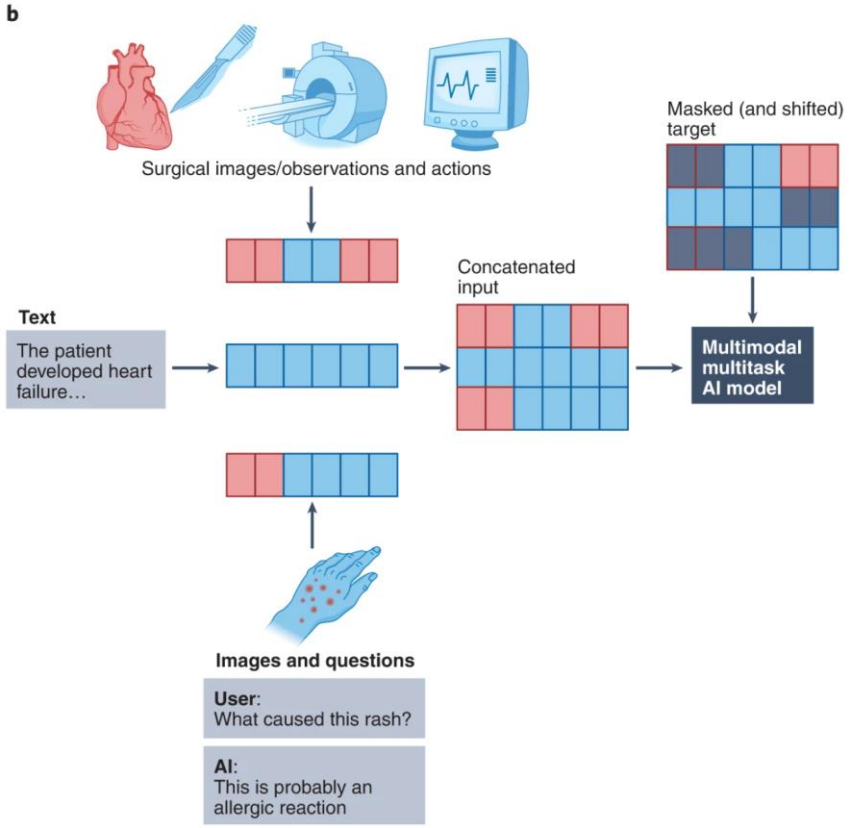
Acoustic
- **Prosody**
  - Intonation
  - Voice quality
- **Vocal expressions**
  - Laughter, moans

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**

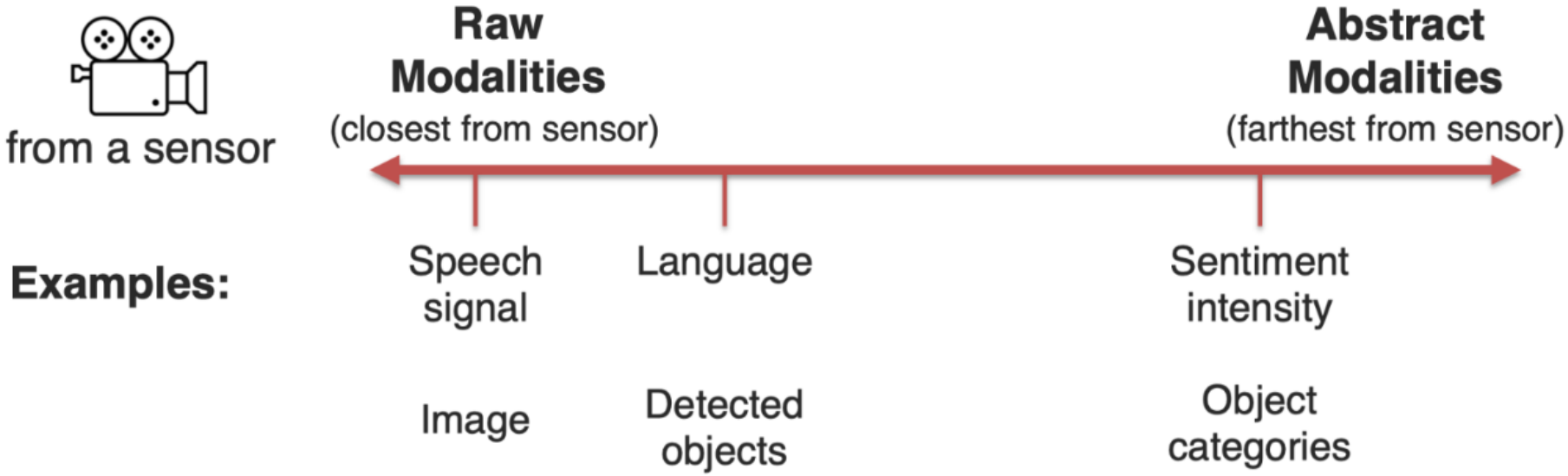


An example of how multimodality can be used in healthcare. Image from Multimodal biomedical AI

What is Multimodality?

Definition

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



What is Multimodality?

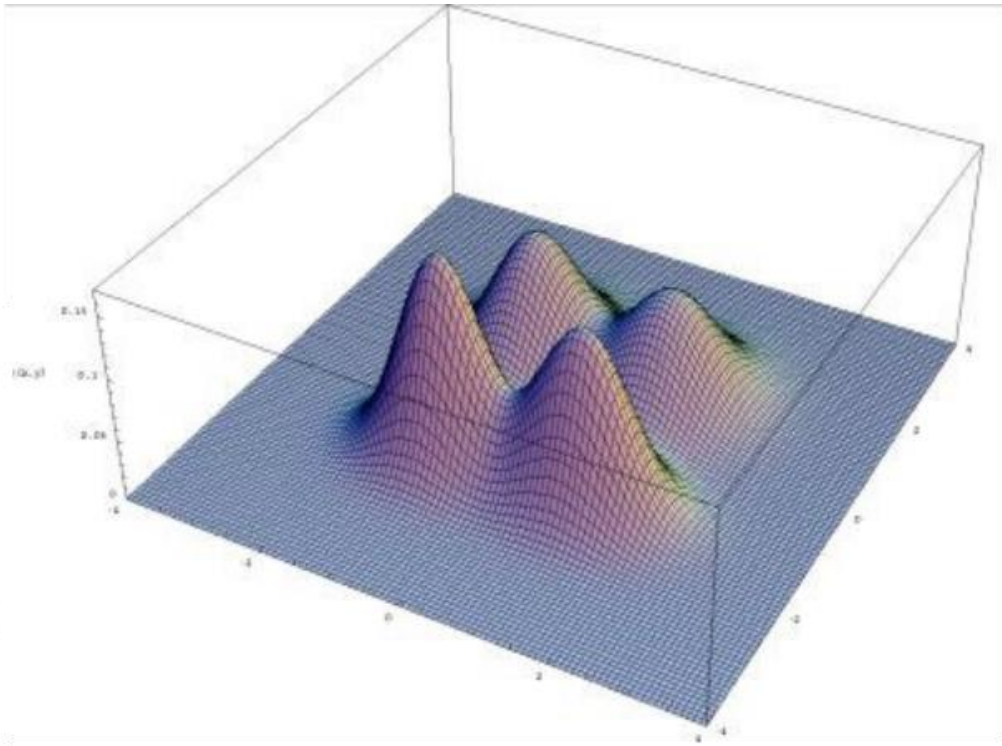
multimodal adjective

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.

Why 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.

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 first focus on two early 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. However, Flamingo's strong performance prompted some to consider it the GPT-3 moment in the multimodal domain.
- **Multimodal LLMs**

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

2 Connecting Vision and Language: CLIP and Flamingo

Two early ideas shaped every multimodal LLM: align images with text (CLIP), then let a language model talk about images (Flamingo).

In this part

2.1 Early Work I: CLIP

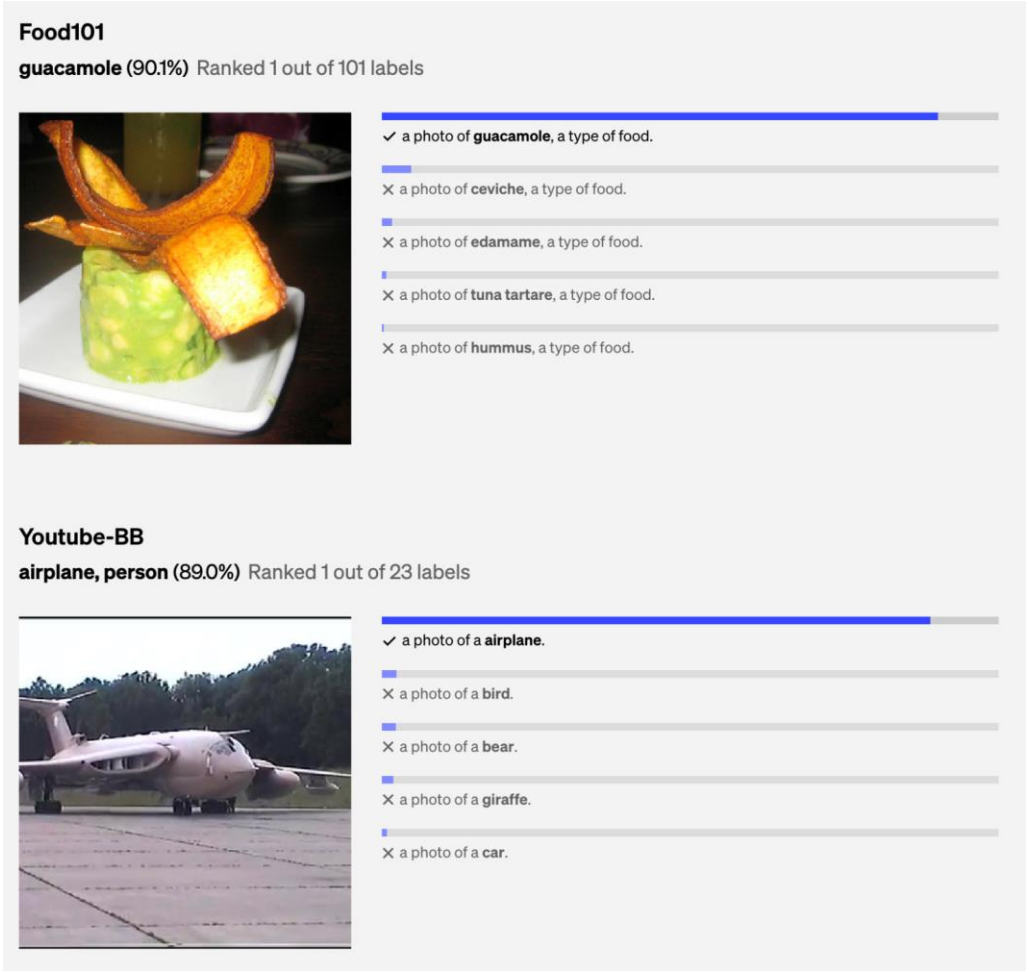
2.2 Early Work II: Flamingo

2.1 Early Work I: CLIP

A good practice for alignment: map images into the text space with a contrastive objective.

CLIP: Contrastive Language-Image Pre-training

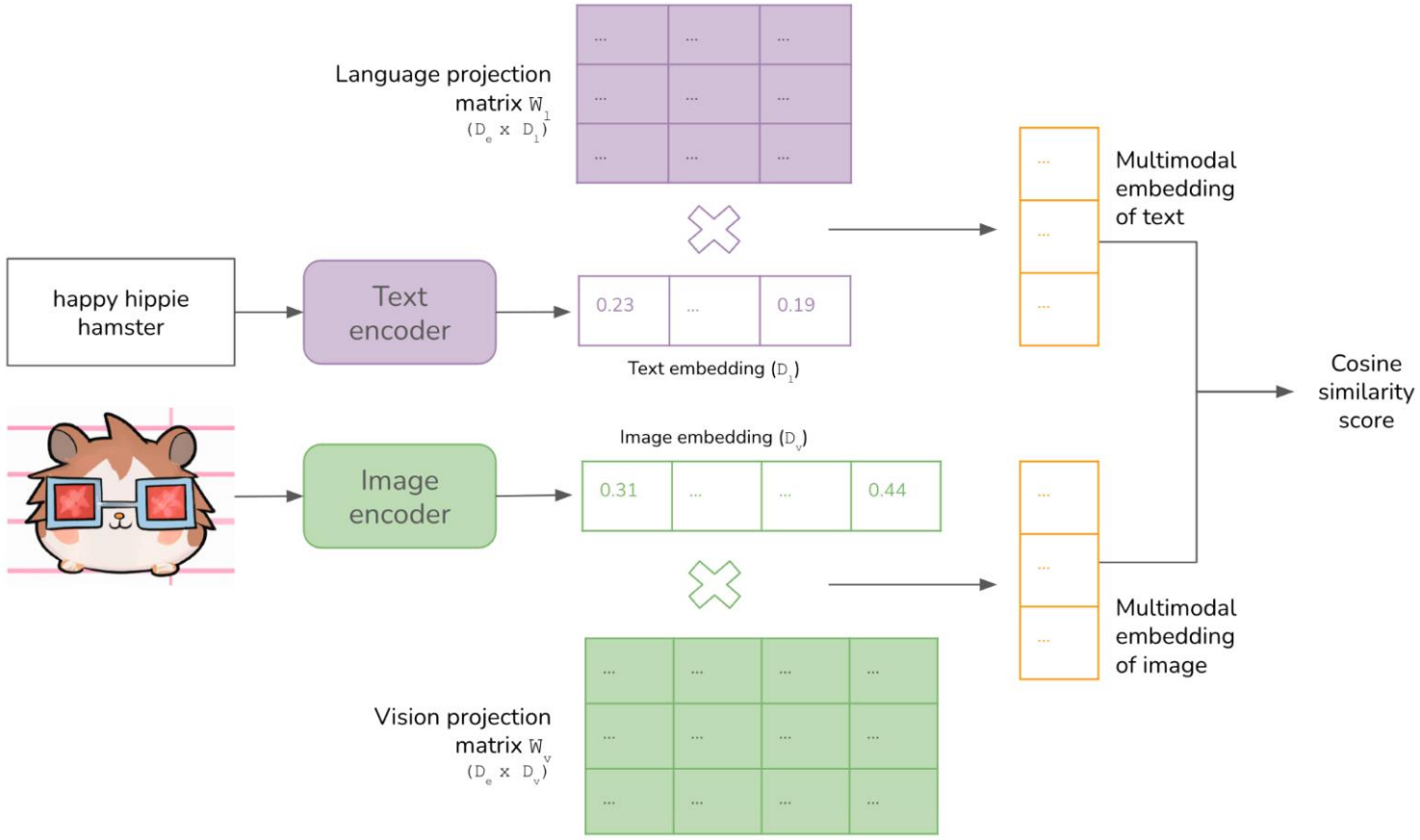
➔ CLIP leveraged **natural language supervision** and **contrastive learning**, which allowed CLIP to both scale up their data and make training more efficient. We'll go over why/how these two techniques work.



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.

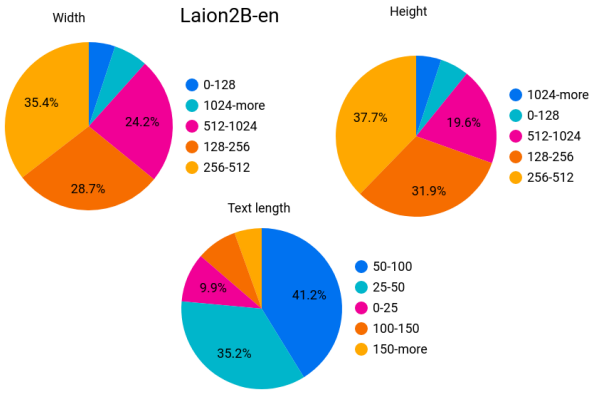
- [CLIP] Learning Transferable Visual Models From Natural Language Supervision (OpenAI, 2021)

Data Example: Wikipedia-based Image Text Dataset

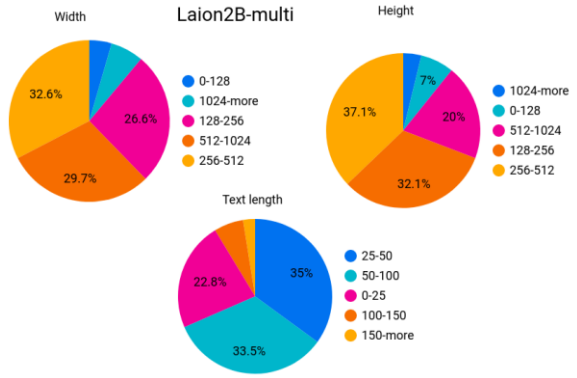


An Even Larger One: LAION-5B

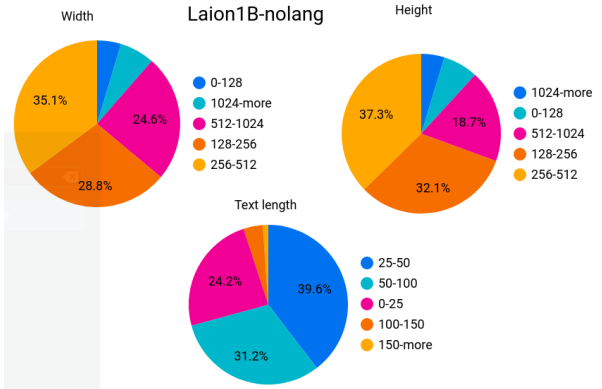
- **Dataset columns**
  - **URL**: the image url, millions of domains are covered
  - **TEXT**: captions, in english for en, other languages for multi and nolang
  - **WIDTH**: picture width
  - **HEIGHT**: picture height
  - **LANGUAGE**: the language of the sample, only for laion2B-multi, computed using cld3
  - **similarity**: cosine between text and image ViT-B/32 embeddings, clip for en, mclip for multi and nolang
  - **pwatermark**: probability of being a watermarked image, computed using our watermark detector
  - **punsafe**: probability of being an unsafe image, computed using our clip based detector



Average text length: 67



Average text length: 52

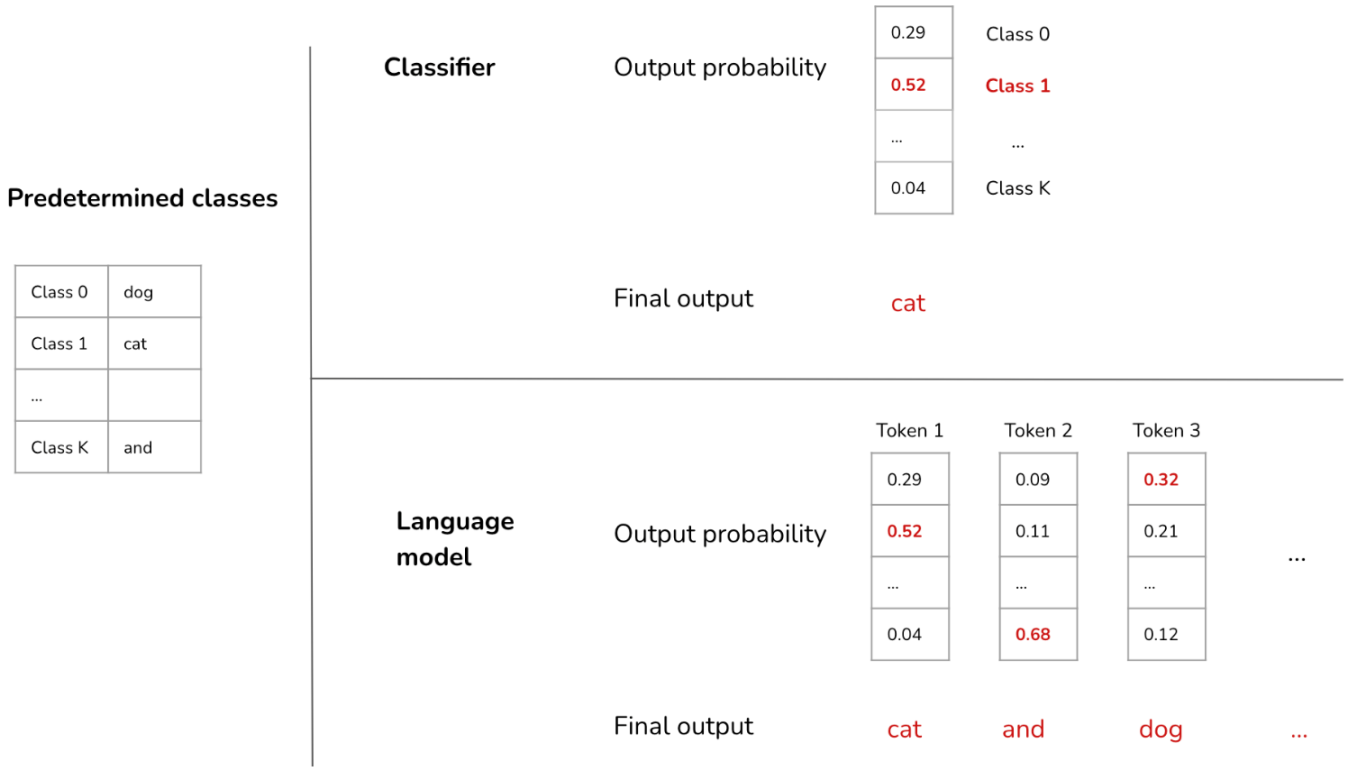


Average text length: 46

Existing Data

数据类型	数据种类	数据数量	压缩后 数据大小	压缩格式	采集时间	数据说明
开源图像数据集	LAION-5B	50亿图文对	240TB	tar	2023年6月	/
	Laion400	4亿图文对	12TB	tar	2023年6月	采集地址：https://laion.ai/laion-400-open-dataset/ 图片大小256*256
	OBELICS	3亿图片文对	26TB	tar	2024年2月	采集地址：https://huggingface.co/datasets/HuggingFaceM4/OBELICS
	MMC4	3.2亿张图文对	26TB	tar	2024年1月	fewer faces v1.1版本 https://github.com/allenai/mmc4/blob/main/README.md
	Coyo-700m	4.1亿图文对	50TB	tar	2024年10月	分辨率两边不小于300像素
	MINT-1T	26亿图文对	272TB	tar	2024年8月	采集地址： https://huggingface.co/datasets/mlfoundations/MINT-1T-HTML
	BLIP3-KALE	1.83亿图文对	12.5TB	tar	2024年12月	使用网络采集文本来增强合成的图像captions，以生成基于事实的图像captions，可用于训练更强大、知识更丰富的多模态模型。数据集全量 2.18亿个图文对，因下载时部分链接失效，实际落盘1.83亿图文对。
	CapsFusion-120M	1.2亿图文对	4TB	tar	2024年8月	采集地址：https://huggingface.co/datasets/BAAI/CapsFusion-120M
	Pixelprose	1500万图文对	3TB	tar	2024年8月	采集地址：https://huggingface.co/datasets/tomg-group-umd/pixelprose
商用图像数据集	谷歌艺术	400万图片	13.9TB	图片格式	2025年6月	最短边长大于1024P，以高清晰度图像为主
	LFI-Gallery	100万图片	5TB	图片格式	2025年6月	Leica Fotografie International的gallery（画廊）模块的所有图片。图片分辨率大于1080P比率较高
	马蜂窝（游记图片）	1.5亿张游记图片	110TB	图片格式	2024年7月	游记图片，以及图片和原始html游记 对应关系的meta信息。 另有游记文本数据集（详见社交互动sheet页）：HTML格式，保留所有原始信息，纯文本，含图片URL
	下厨房（封面图+步骤图）	348万张封面图片 1881万张步骤图片	7.5TB	图片格式	2025年6月	菜谱图片：为解析处理后的数据，提取了菜谱标题、详细烹饪步骤等信息。包括封面图和做菜步骤图，且图片已下载。 另有菜谱文本数据集（详见文本垂类sheet页）：353个csv文件保存了，菜谱的原始HTML信息，HTML包含 用料、详细烹饪步骤及配图URL等等信息
	Sogou图库	280万张图文对	269GB	gz	/	包含原图，缩略图，链接，文字说明
	生物百科	2200万张图片	12TB	zip	2025年6月	采集地址：https://www.inaturalist.org/ 含50万余个物种，每个物种图片含meta信息，并配有中英文维基百科、抖音百科、especies百科，以及各个百科页面的图片
	植物百科	1015万张图片	3.5TB	zip	2025年9月	植物图像数据库，知识图谱训练语料

Discriminative vs. generative for image classifications



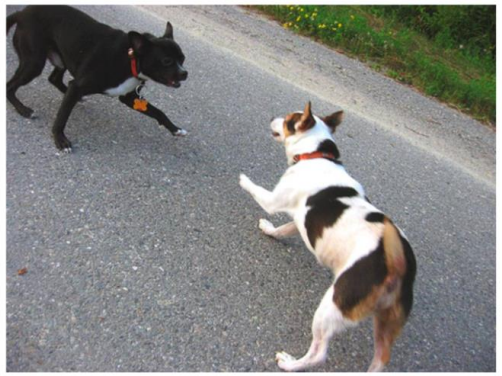
If a classifier outputs only one class for each input, a language model outputs a sequence of classes. Each generated class is called a token. Each token is from a predetermined list, the vocabulary, of the language model.

- [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

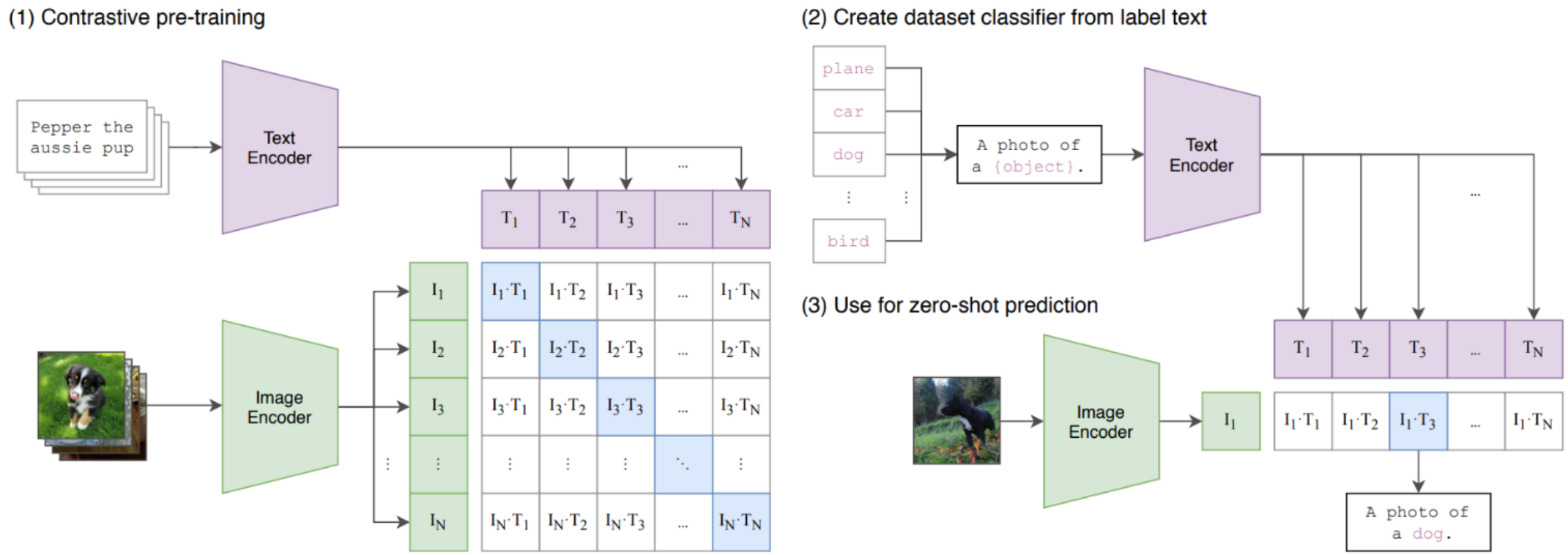
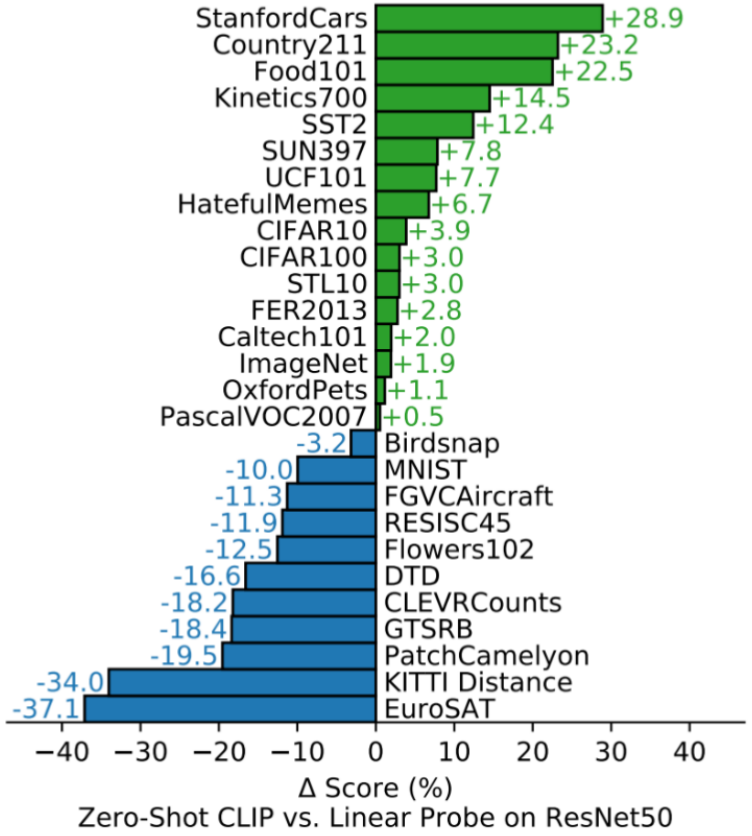


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the 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

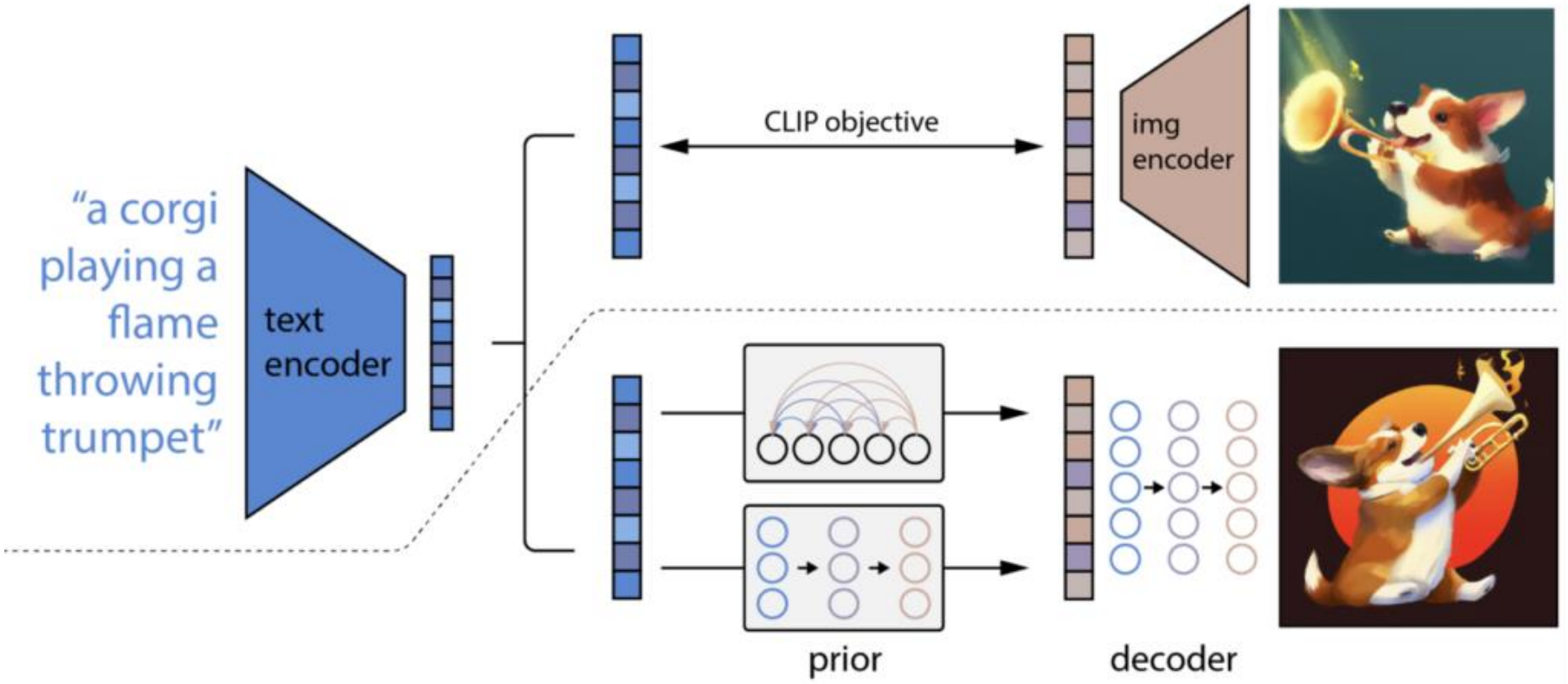
Since CLIP's training process was conceptually similar to image-to-text retrieval and text-to-image retrieval, CLIP "*displays significant promise for widely-applicable tasks like image retrieval or search*." However, "*on image retrieval, CLIP's performance relative to the overall state of the art is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

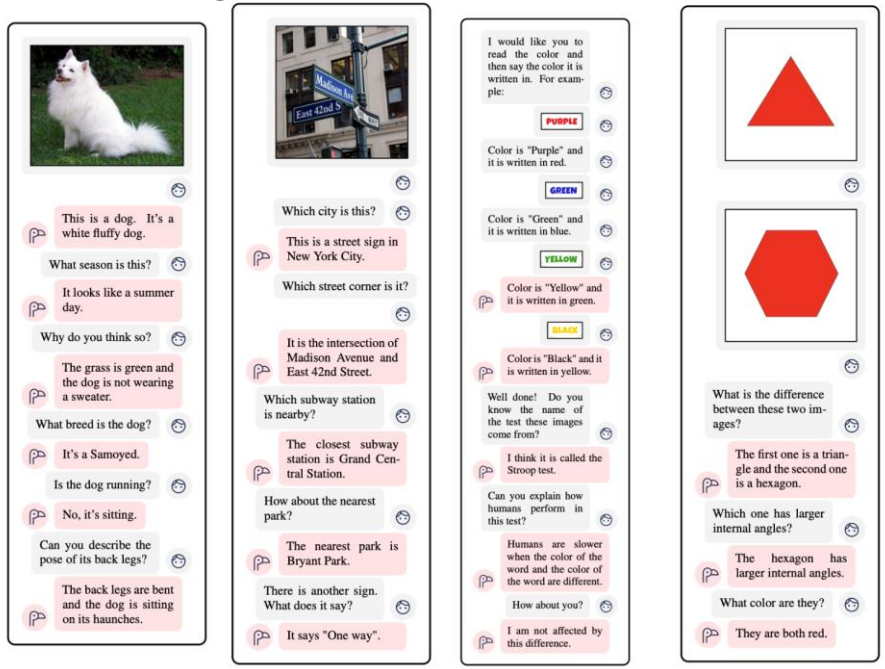
2.2 Early Work II: Flamingo

The dawn of large multimodal models: let a generative language model “talk about” images.

Flamingo: the dawns of LMMs

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

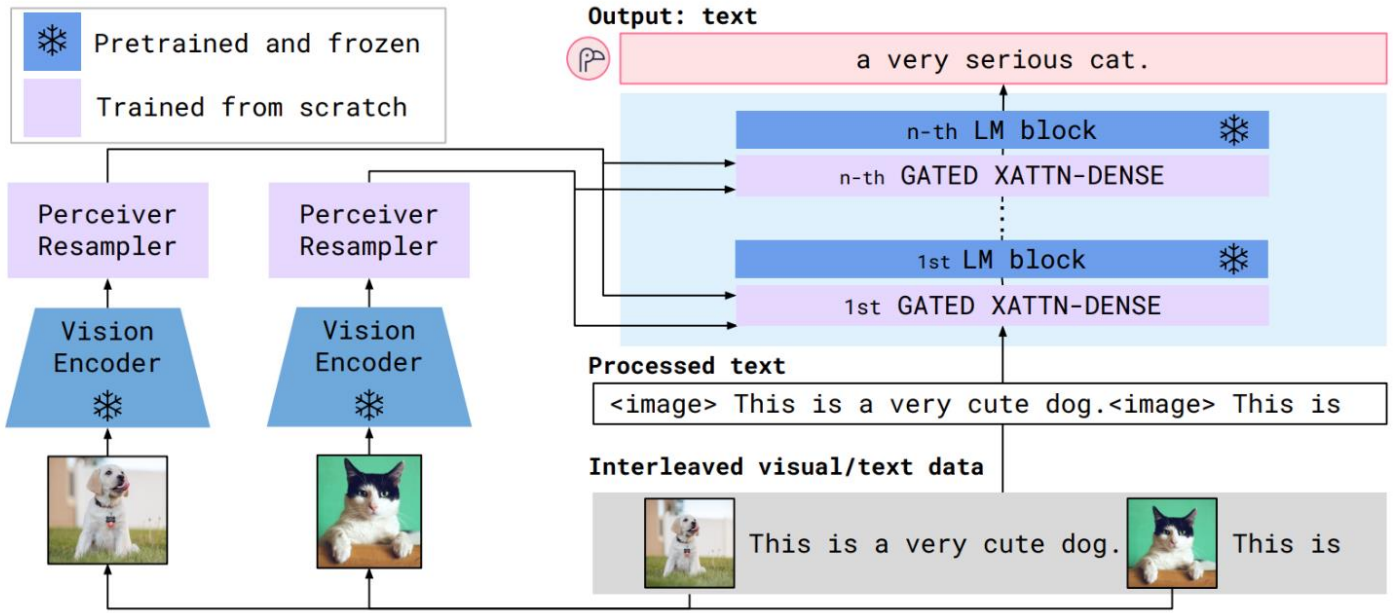


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

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.



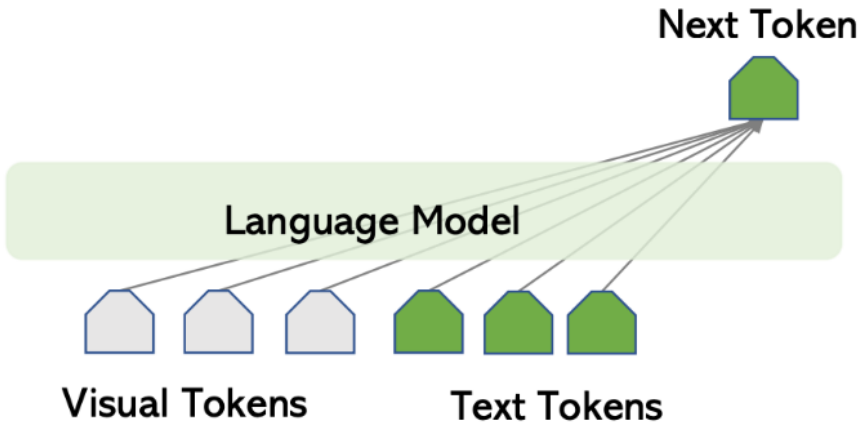
Image-Text Pairs dataset | Video-Text Pairs dataset | Multi-Modal Massive Web (M3W) dataset

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 language model

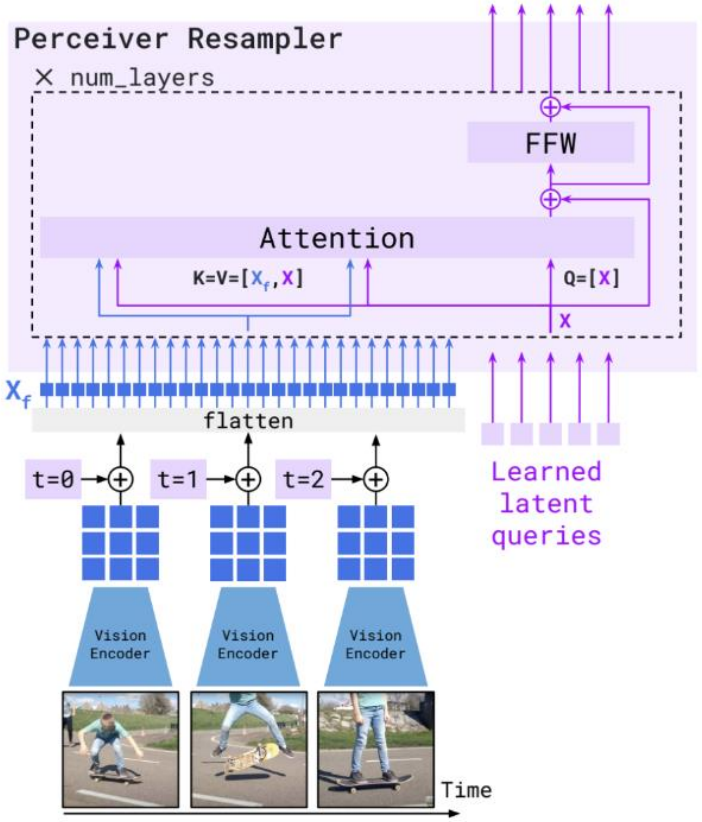
Flamingo uses Chinchilla as their language model. More specifically, they freeze the 9 pretrained Chinchilla LM layers. A traditional language model predicts the next text token based on the preceding text tokens. Flamingo predicts the next text token based on both the preceding text and visual tokens.



- 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.




```
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.

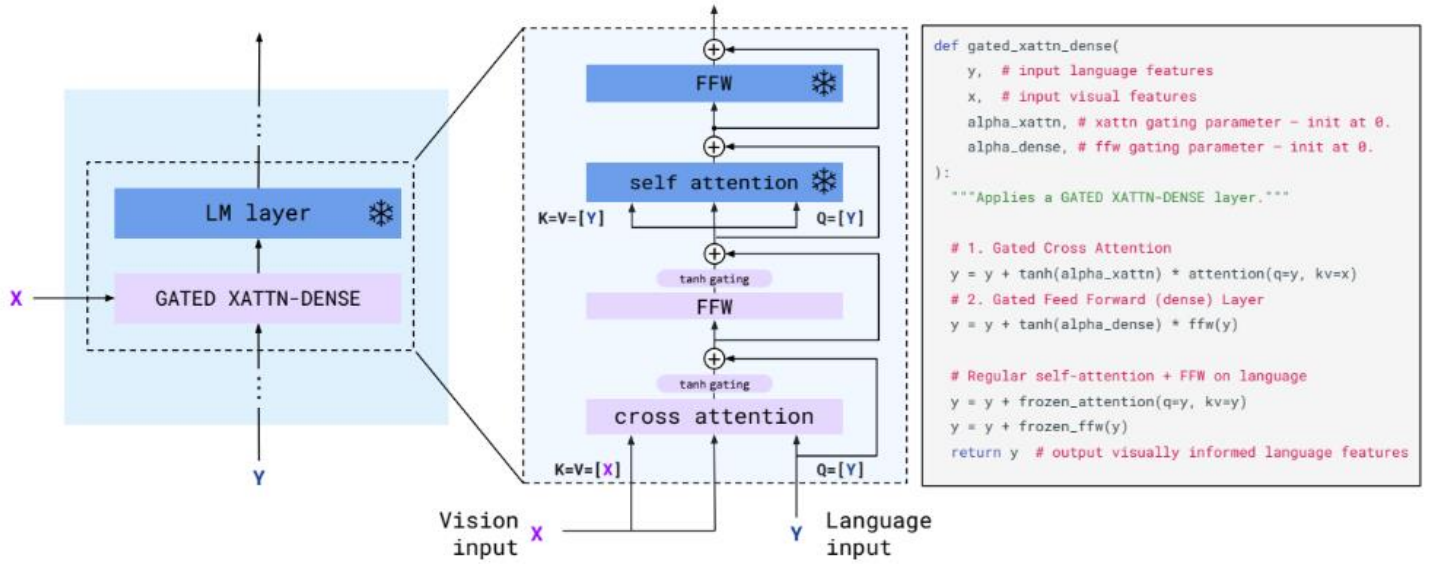


Figure 4: **GATED XATTN-DENSE layers.** To condition the LM on visual inputs, we insert new cross-attention layers between existing pretrained and frozen LM layers. The keys and values in these layers are obtained from the vision features while the queries are derived from the language inputs. They are 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: autoregressive language models

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

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

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)

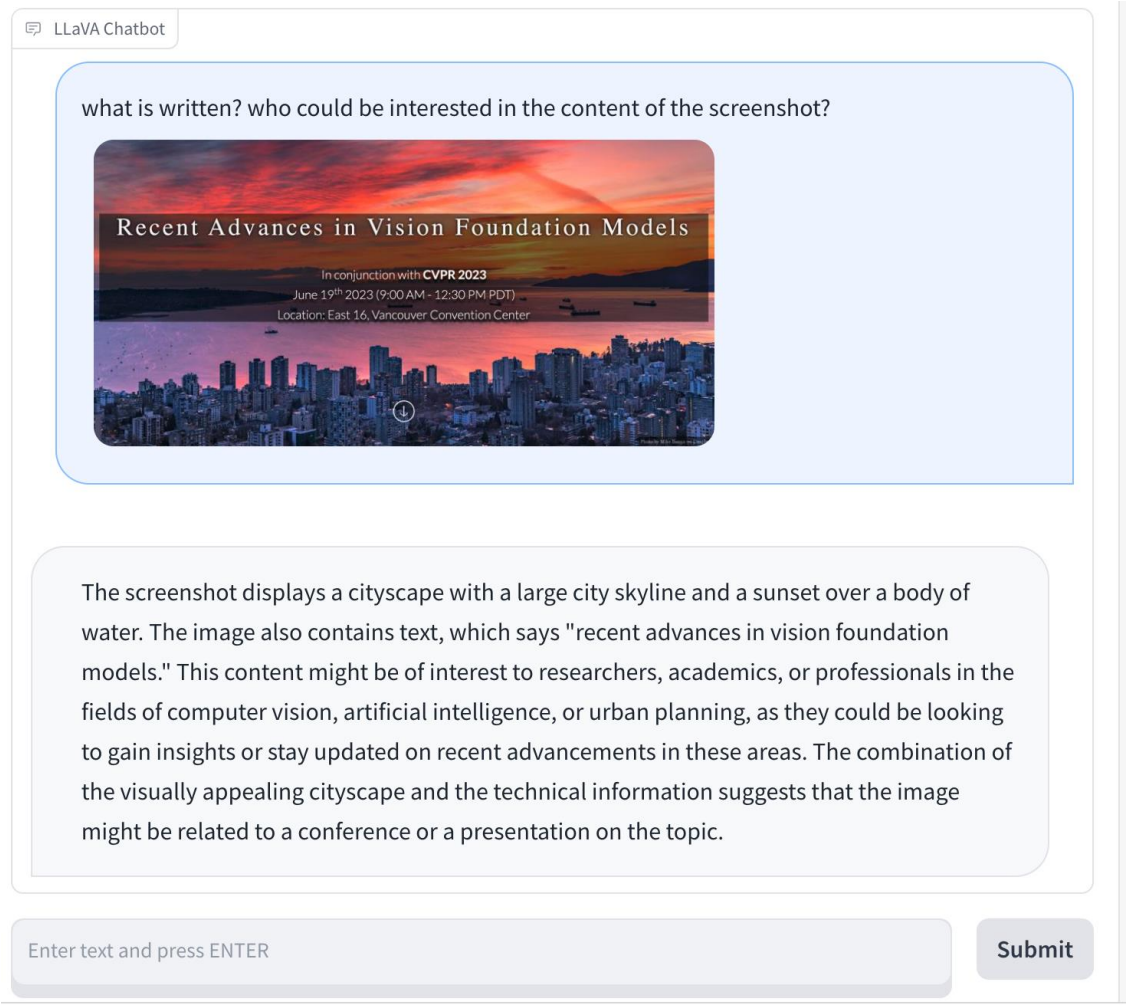
3 Multimodal LLMs: Architecture

Most multimodal LLMs keep language as the pivot: encode other modalities, project them into the LLM's token space, and let the LLM reason.

Multimodal LLMs

— Welcome to the Era of Large Language Models. (Large) language models really improve the cognition-level abilities (e.g., reasoning and planning), but it also

Multimodal LLMs



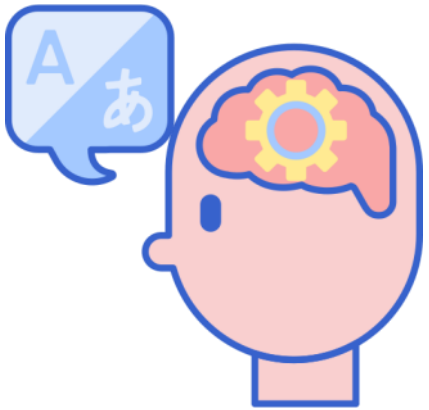
Visual chat with LMM. Generated by LLaVA: https://llava-vl.github.io/

Preliminary Idea: Intelligence over Language

Emergent phenomena have extensively already occurred in language-based LLMs.



These LLMs now generally possess very powerful **semantic understanding capabilities**. This also implies that **language is a crucial modality for carrying intelligence**.



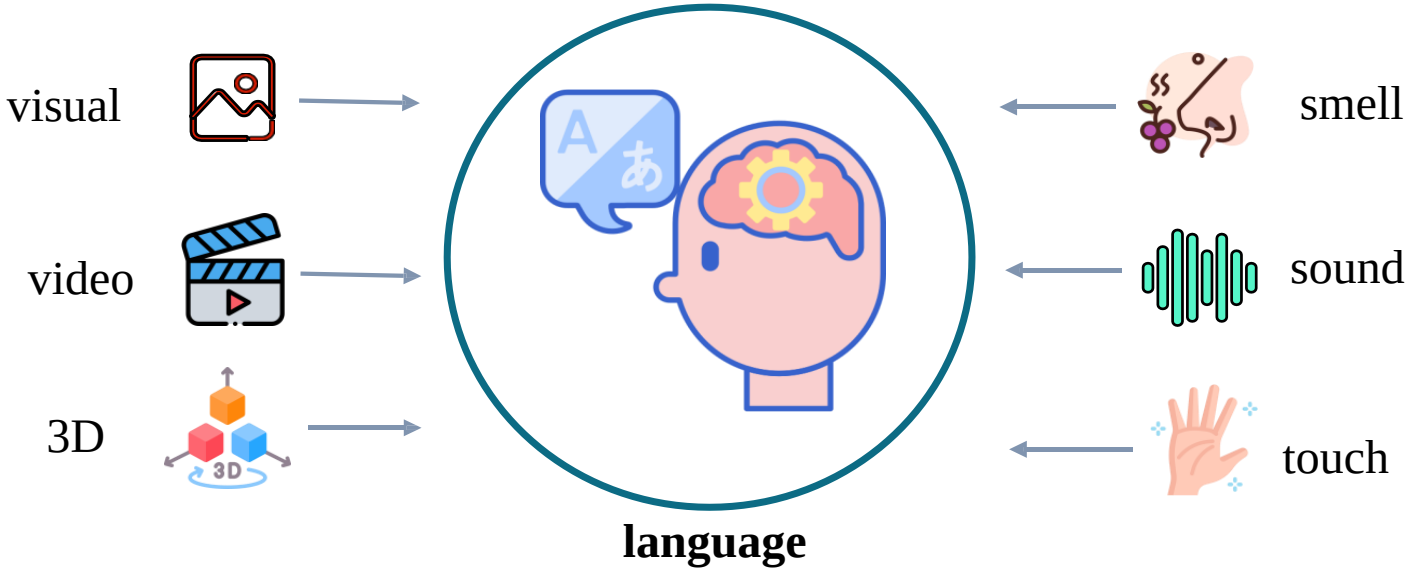
language

Preliminary Idea: Language Intelligence as Pivot

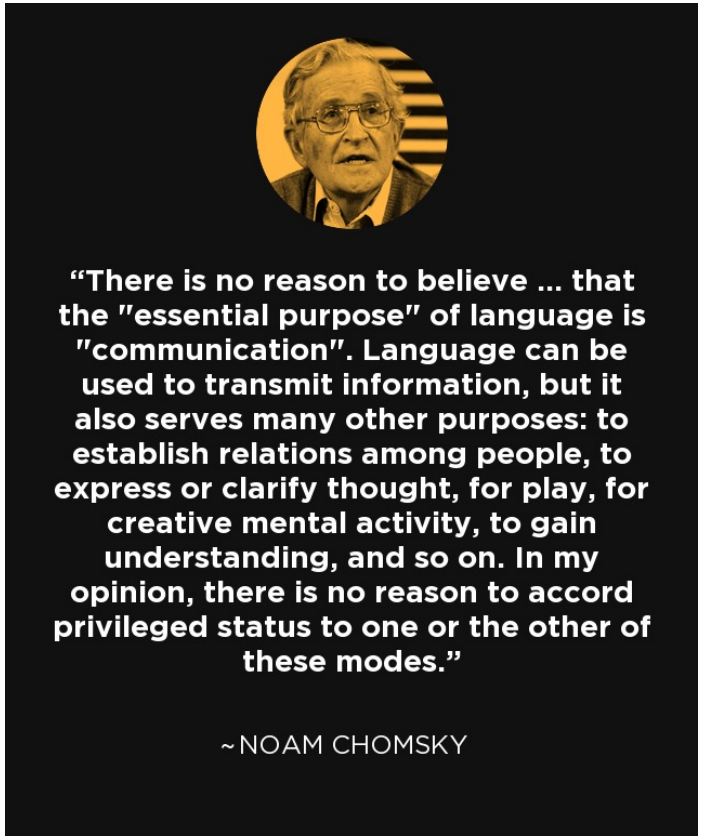
Given this premise, **nearly all CURRENT MLLMs are built based on language-based LLMs** as the core decision-making module (i.e., the brain or central processor).

By adding additional external non-textual modality modules, LLMs are enabled with multimodal abilities.

- Extend the capability boundary, next milestone towards more advanced intelligence
- More applications



Language for communication and also thought



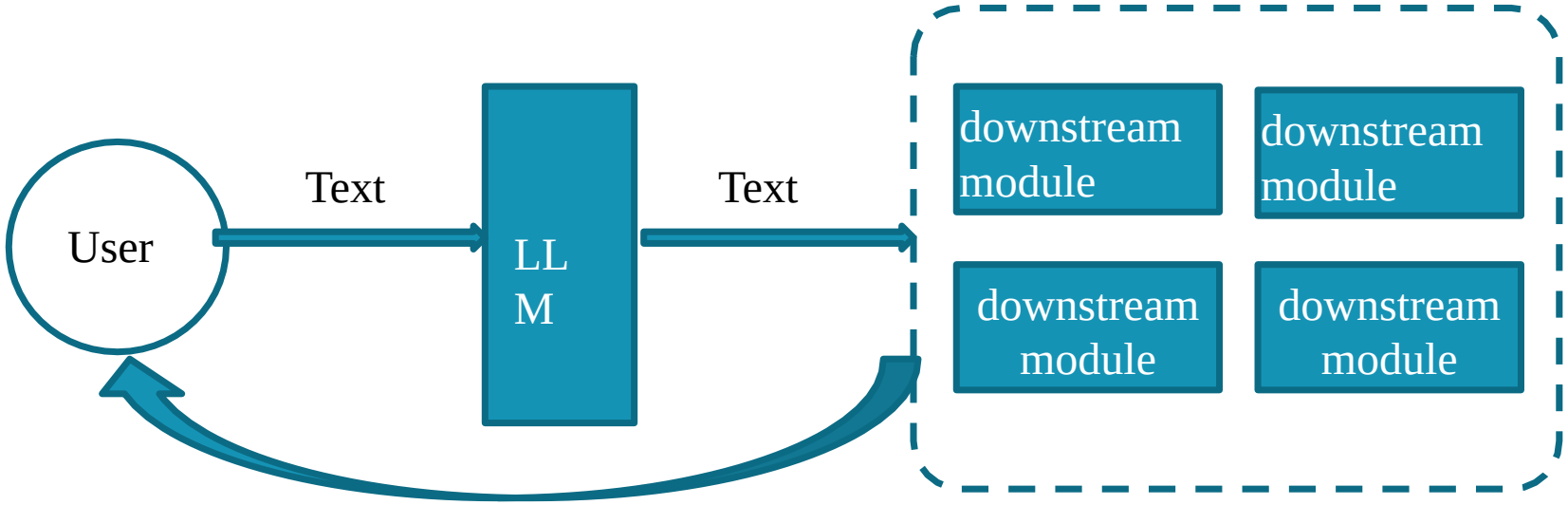
A recent perspective paper from Nature: Language is primarily a tool for communication rather than thought.
https://www.nature.com/articles/s41586-024-07522-w

Type-I: LLM as Discrete Scheduler/Controller

The role of the LLM is to **receive textual signals** and **instruct textual commands to call downstream modules**.

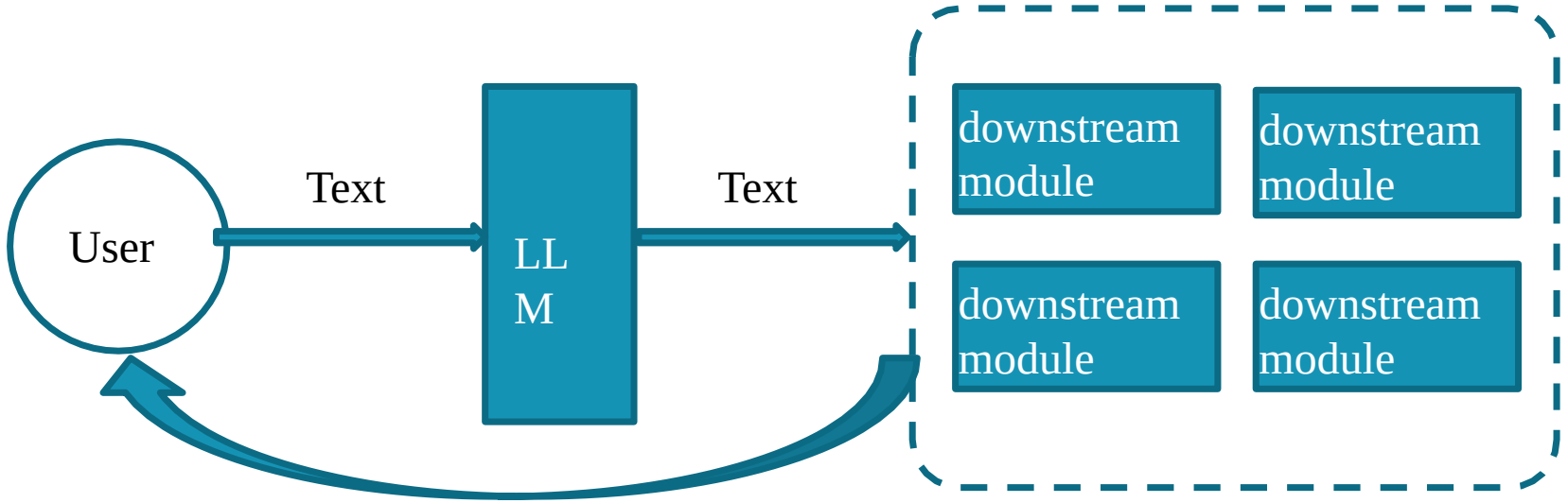
- Key feature:

  All message passing within the system, such as “multimodal encoder to the LLM” or “LLM to downstream modules”, is facilitated through **pure textual** commands as the medium.

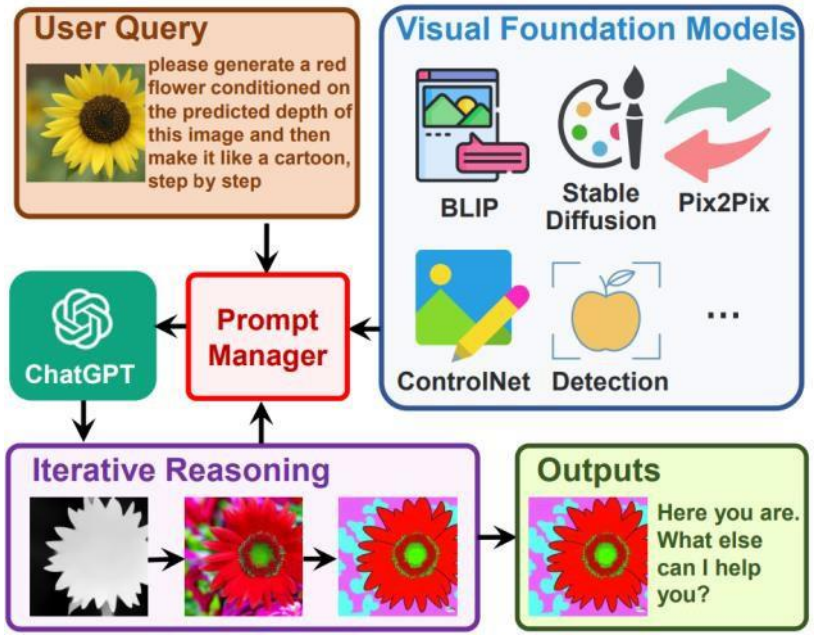


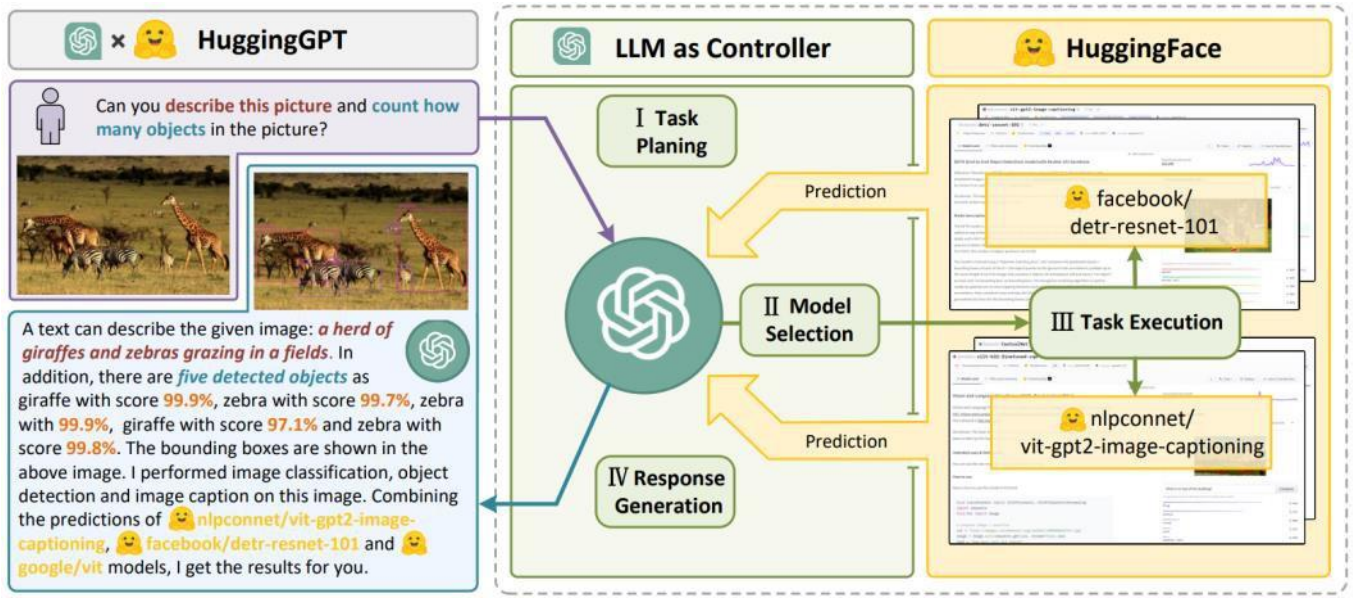
Type-I: LLM as Discrete Scheduler/Controller

- Representative MLLMs:
  - Visual-ChatGPT
  - HuggingGPT
  - MM-REACT
  - ViperGPT
  - AudioGPT
  - LLaVA-Plus
  - …



Type-I: LLM as Discrete Scheduler/Controller

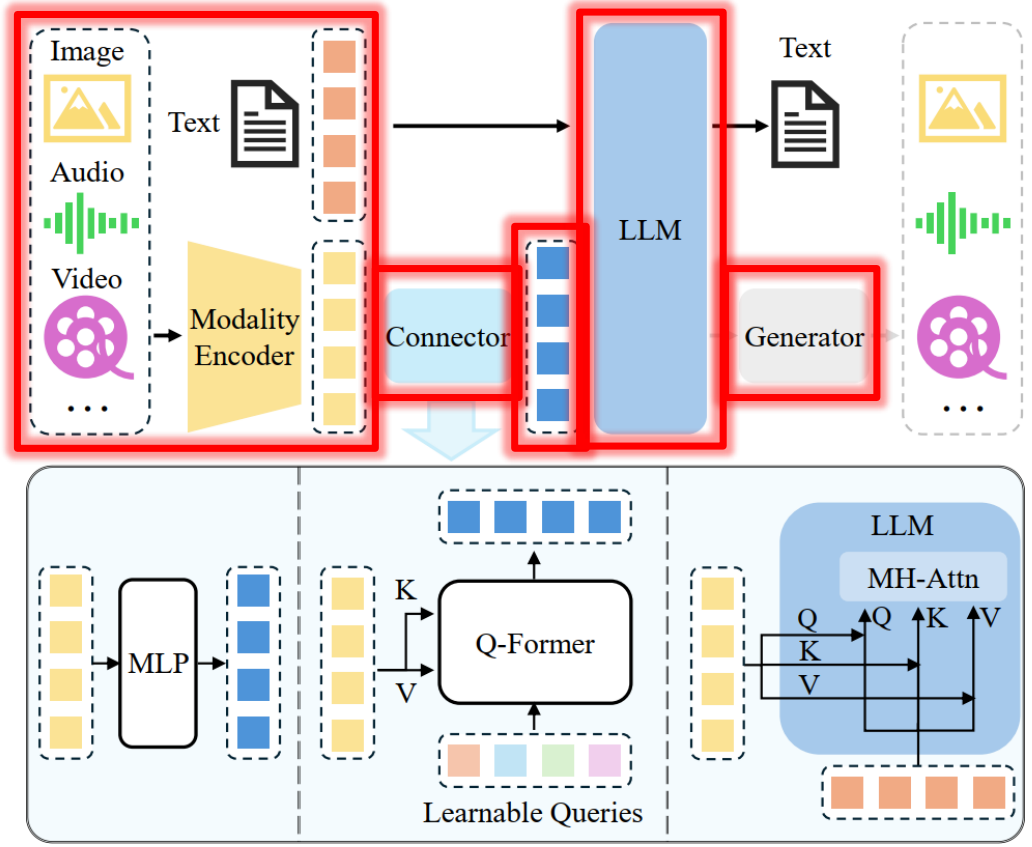




- Quick to build (without training), flexible extension to many tool features
- Information loss in text medium, the bottle-neck

Type-II: LLM as Joint Part of System

- > 90% MLLMs belong to this category.
- Higher upper-bound, better integrated into a unified model

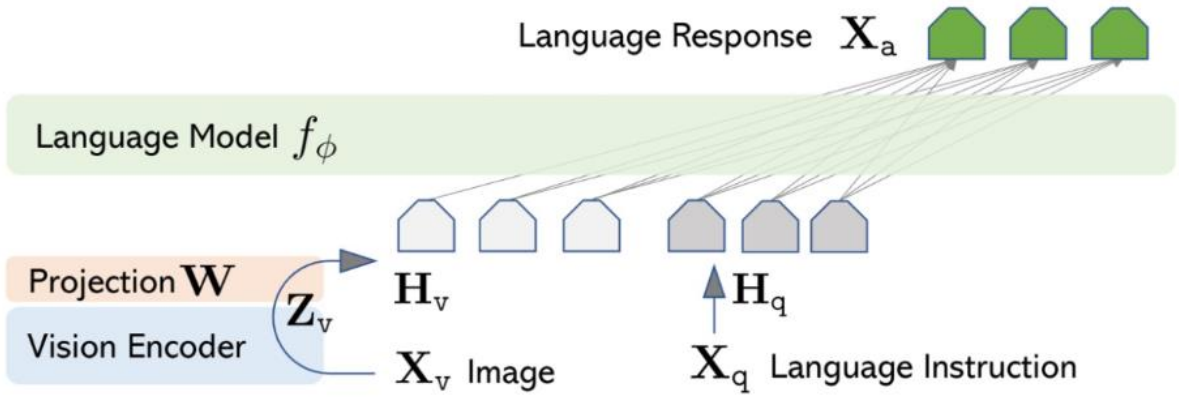


[1] A Survey on Multimodal Large Language Models. [https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models, 2023.](https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models)

Architecture of LLMs

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.

Other Multimodal LLMs: Qwen-VL

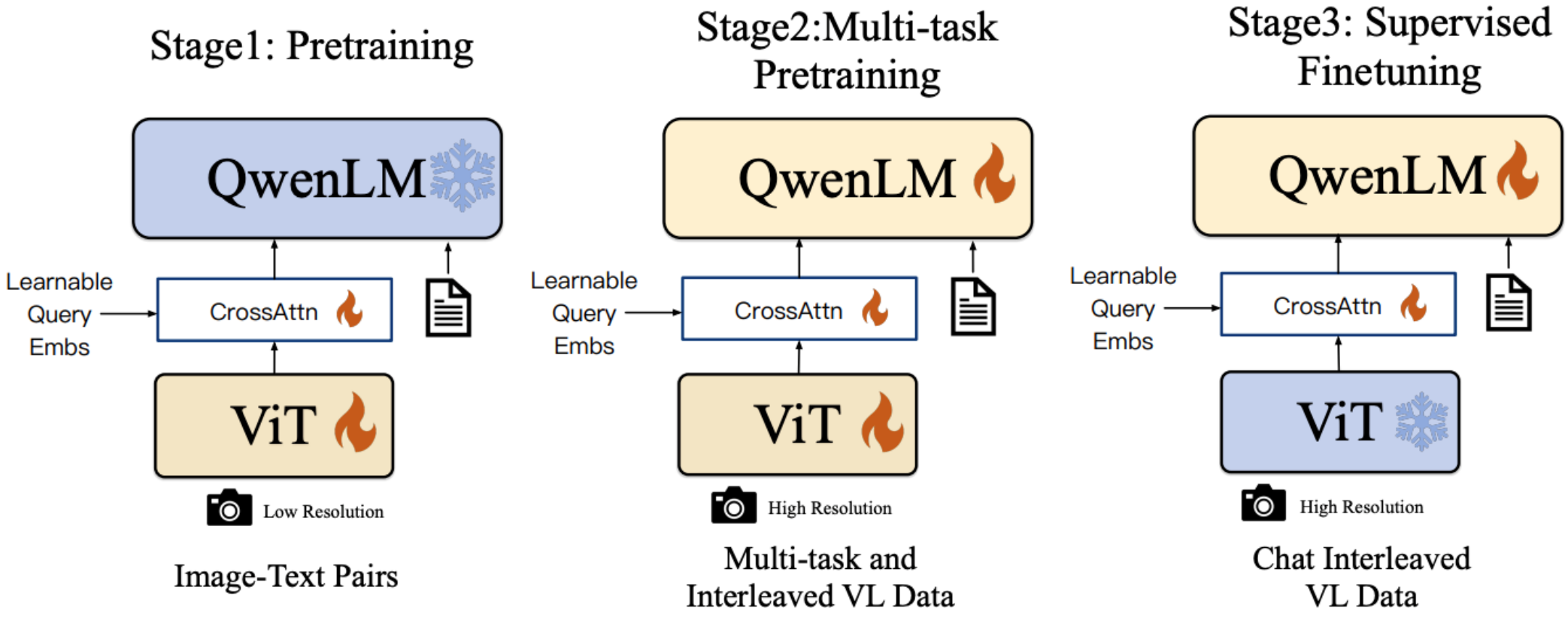
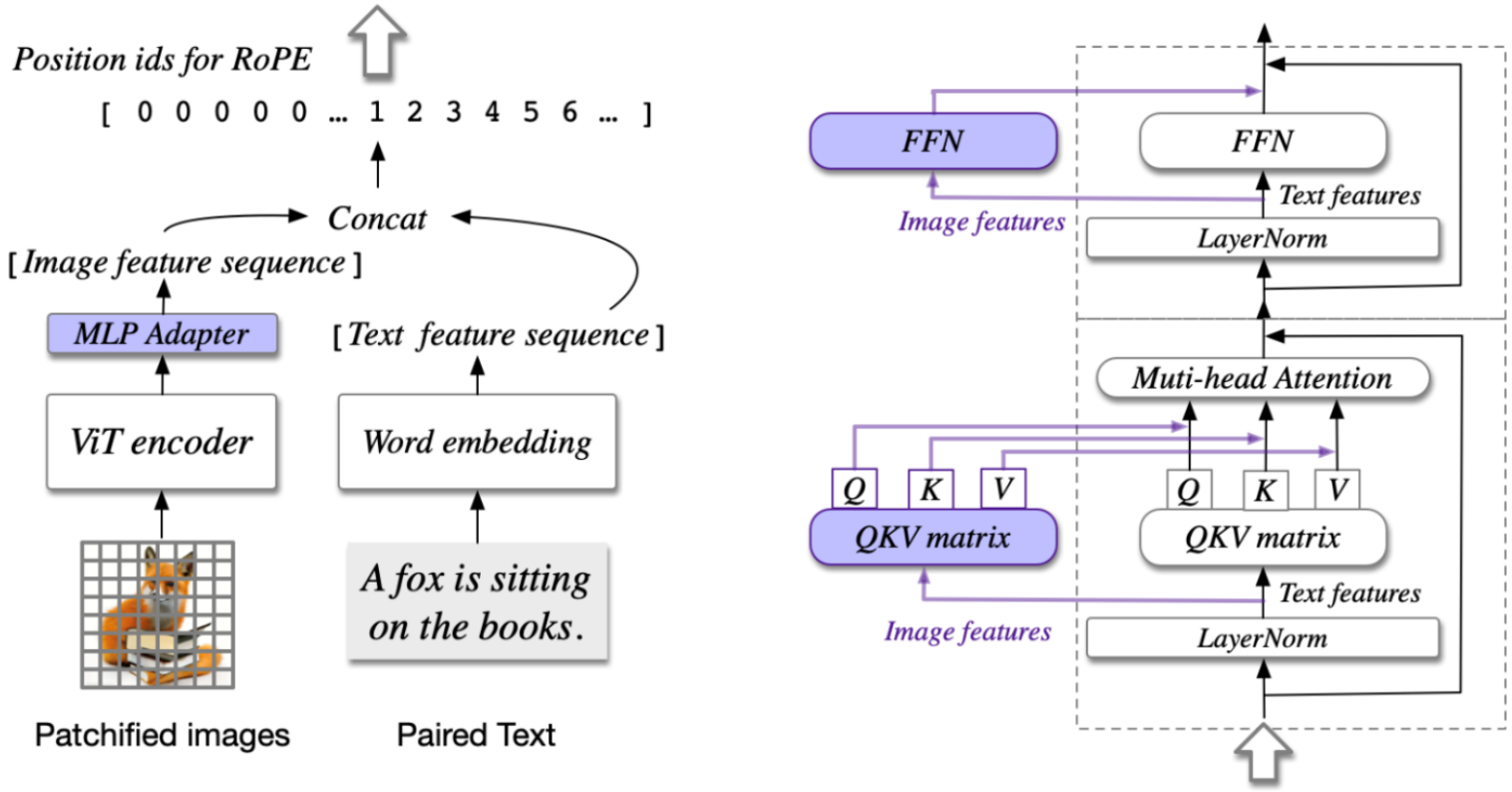


Figure 3: The training pipeline of the Qwen-VL series.

Other Multimodal LLMs: CogVLM-Chat

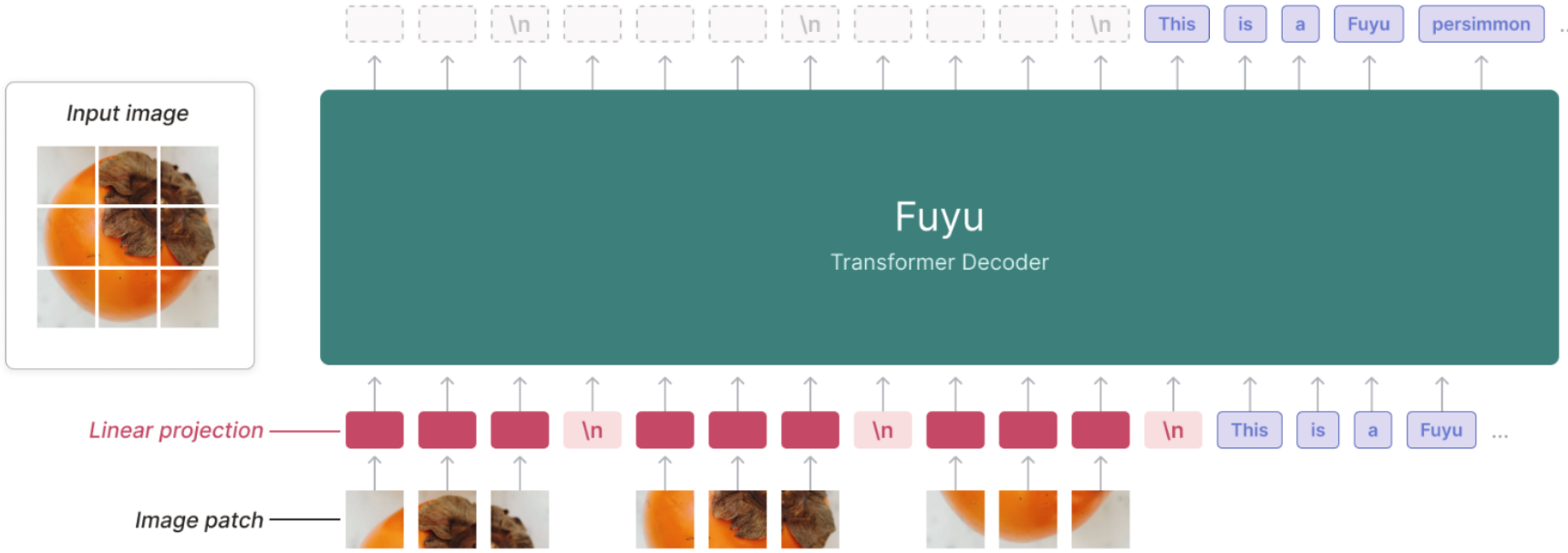


(a) The input of visual language model

(b) The visual expert built on the language model

Figure 3: The architecture of CogVLM. (a) The illustration about the input, where an image is processed by a 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



Native Multimodal Models: GPT-4o & Gemini

- Paradigm Shift: From Modular to Native Multimodal
  - GPT-4o (OpenAI, 2024): Omni-model processing text, image, audio, video natively in a single end-to-end architecture
  - Gemini 2.0 (Google, 2024): Native multimodal from the ground up, not CLIP+LLM stitching
  - Key advantage: Eliminates information bottleneck between vision encoder and LLM
  - Enables real-time multimodal interaction (voice + vision + text simultaneously)
  - Meta Muse Spark （ https://about.fb.com/news/2026/04/introducing-muse-spark-meta-superintelligence-labs ）
  - Qwen3.5-Plus / Qwen3.5， Native Multimodal Agents

Open-Source SOTA Multimodal LLMs

- Qwen2.5-VL (Alibaba): Dynamic resolution, video understanding, agent capabilities
- InternVL2.5 (Shanghai AI Lab): Scaling to 78B, strong OCR and document understanding
- DeepSeek-VL2 (DeepSeek): MoE architecture for efficient multimodal reasoning
- LLaVA-OneVision: Unified single-model for image, video, multi-image tasks
- MiniCPM-V 2.6: Edge-efficient MLLM rivaling GPT-4V on mobile devices

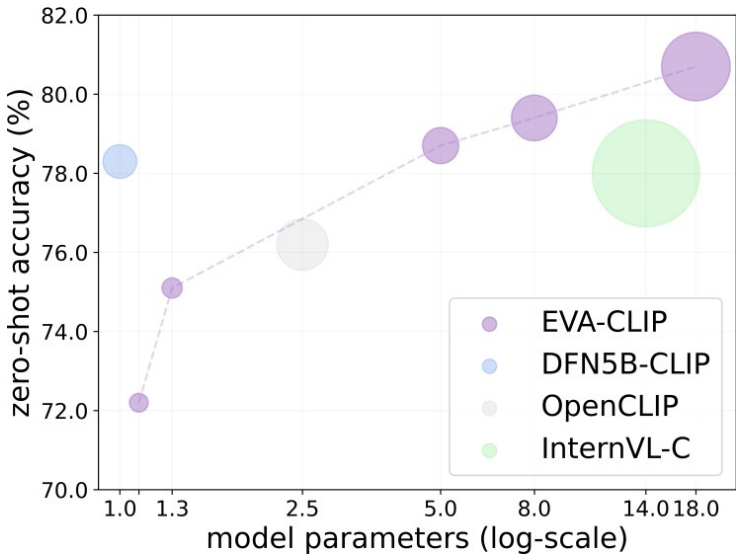
Unified Understanding and Generation

- Beyond separate understanding and generation pipelines
- Janus-Pro (DeepSeek): Decouples visual encoding for understanding vs. generation, unified transformer backbone
- Emu3 (BAAI): Next-token prediction for all modalities (text, image, video) using discrete tokens （Published in Nature）
- Chameleon (Meta): Early-fusion token-based multimodal model
  - Key insight: Separate encoders for understanding (semantic) vs. generation (low-level detail)
  - Our work: Janus, ShareGPT-4o-Image (see Slide 90)

Visual encoders or tokenizer

Visual Encoder

- **CLIP-ViT** is the most popular choice for vision-language models.
  - Providing image representations well aligned with text space.
  - Scale well with respect to parameters and data.
- **SigLIP** is gaining increasing popularity (smaller and stronger)



Visual Encoder

- Limitations of existing pretrained ViTs:
  - Fixed low-resolution (224x224 or 336x336) in square shape
- High-resolution perception is essential, especially for OCR capability!



Low resolution encoding misses fine-grained visual details!

Visual Encoder

- High-resolution Multimodal LLMs
  - Image slice-based: Split high-resolution images into slices
  - Representatives:
    - GPT-4V, LLaVA-NeXT, MiniCPM-V 2.0/2.5, LLaVA-UHD, mPLUG- DocOwl 1.5, SPHINX, InternLM-XComposer2-4KHD, Monkey

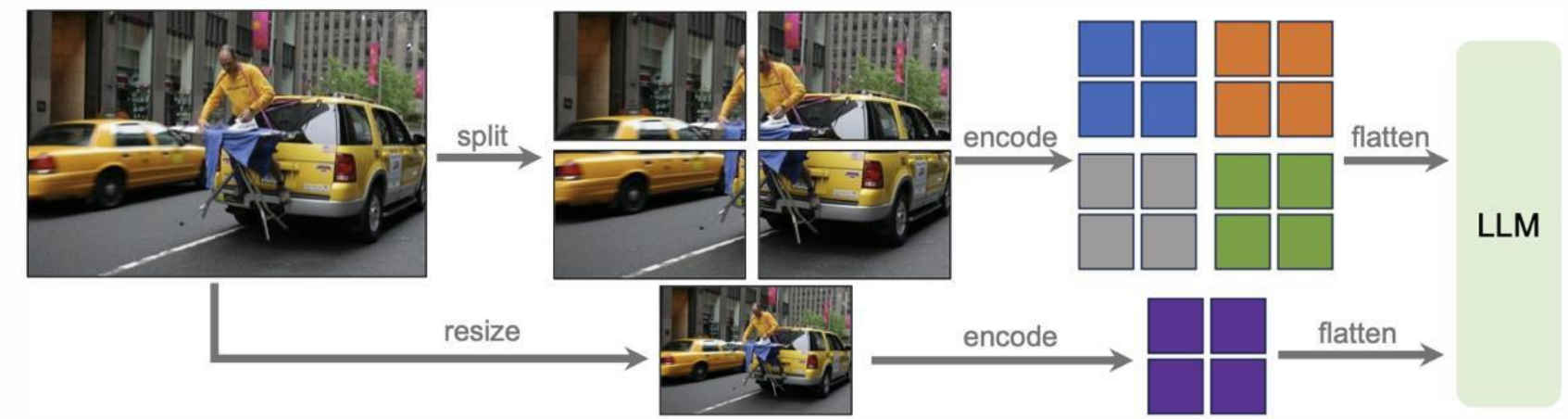
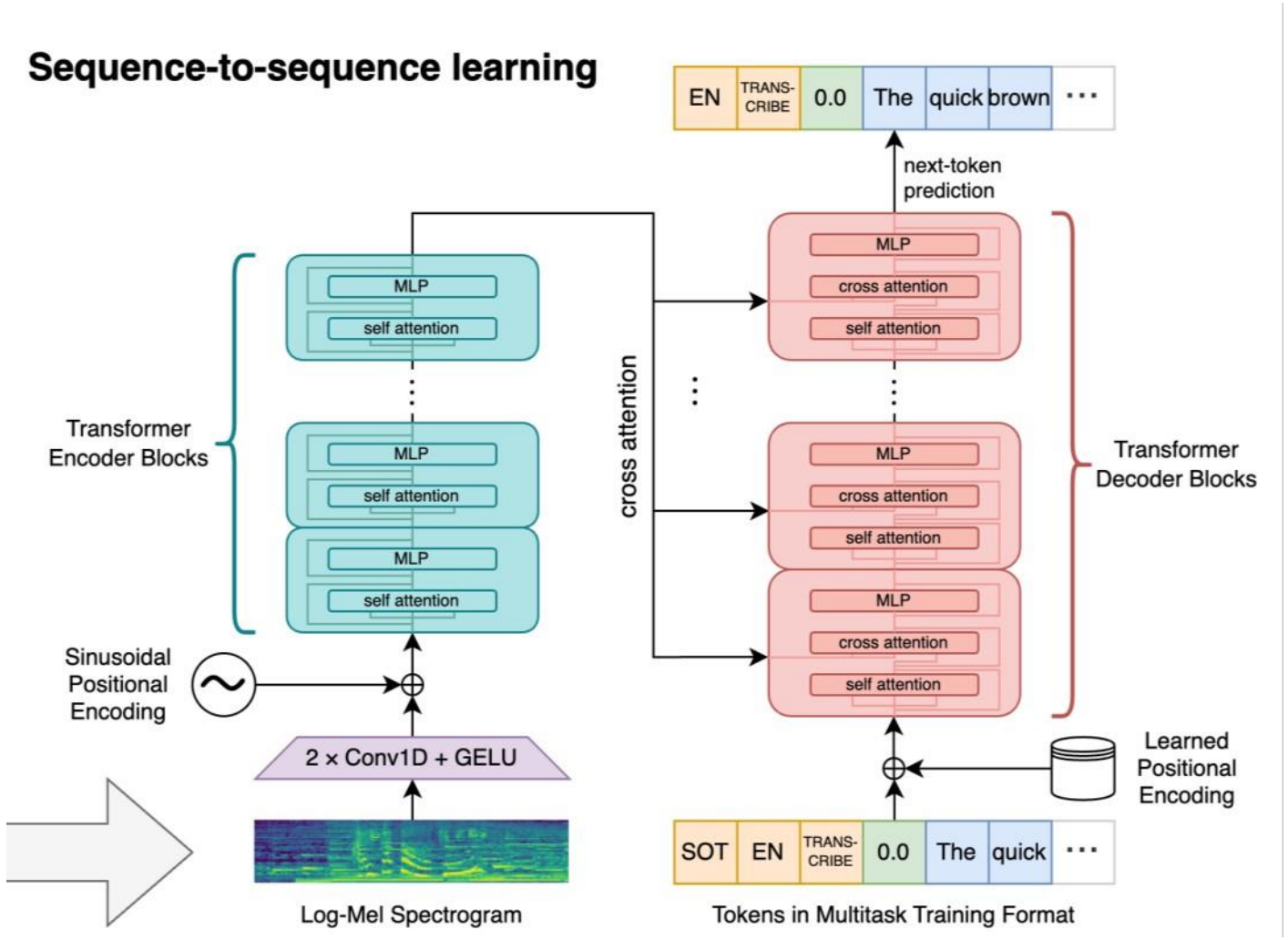


Illustration of dynamic high resolution scheme: a grid configuration of 2 × 2

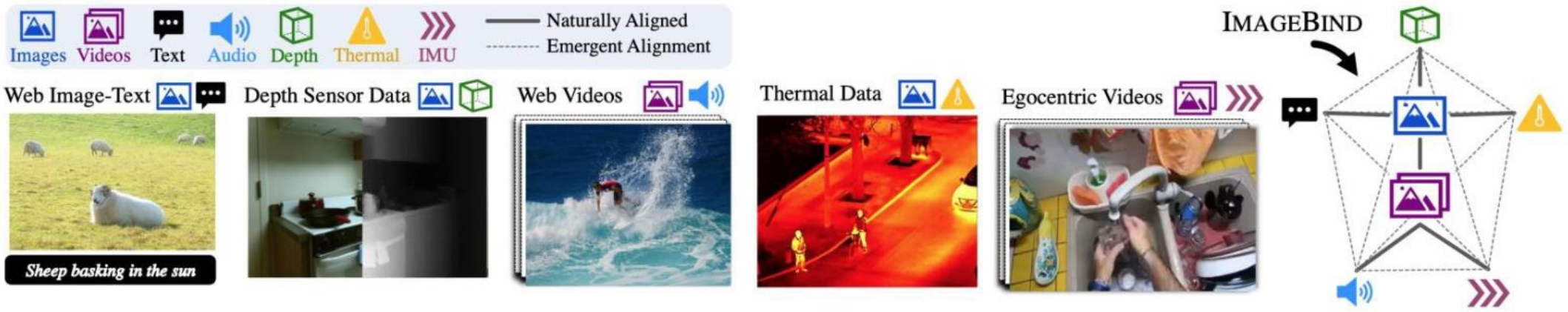
Non-Visual Encoder

- **Audio:**
  - Whisper
  - AudioCLIP
  - HuBERT
  - BEATs
- **3D Point:**
  - Point-BERT



Unified Multimodal Encoder

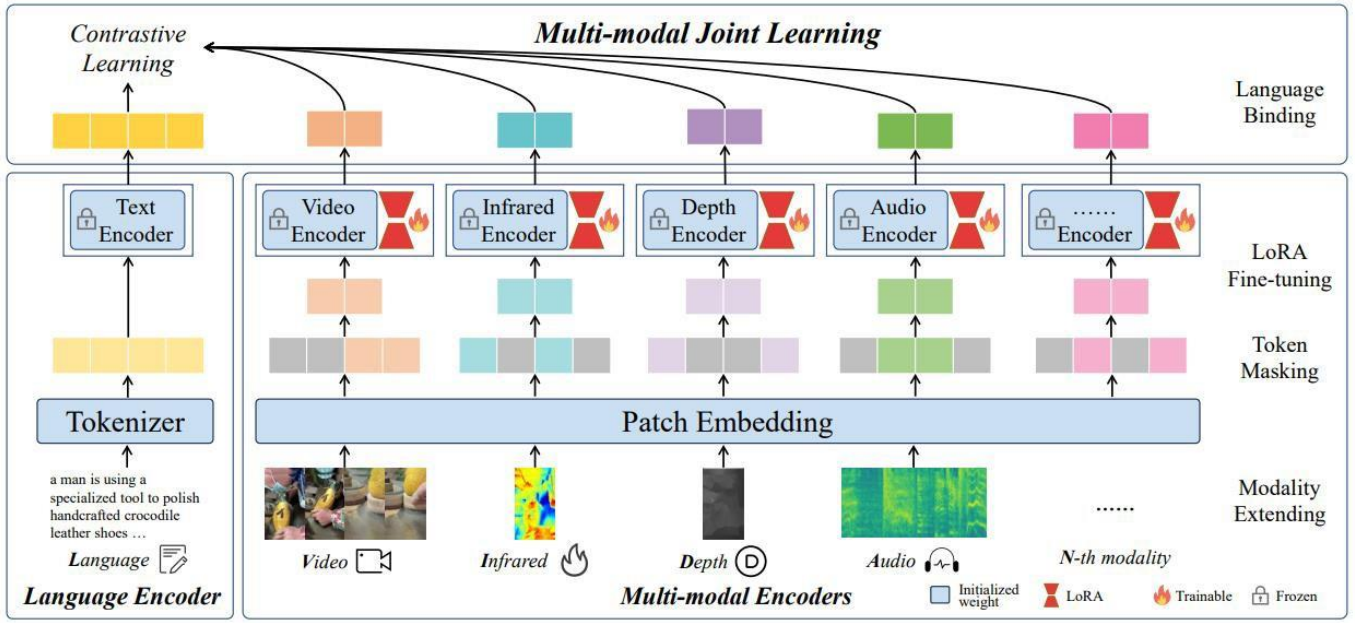
- ImageBind:
  - Embedding all modalities into a joint representation space of **Image**.
  - Well aligned modality representations can benefit LLM understanding



[1] ImageBind: One Embedding Space To Bind Them All. 2023

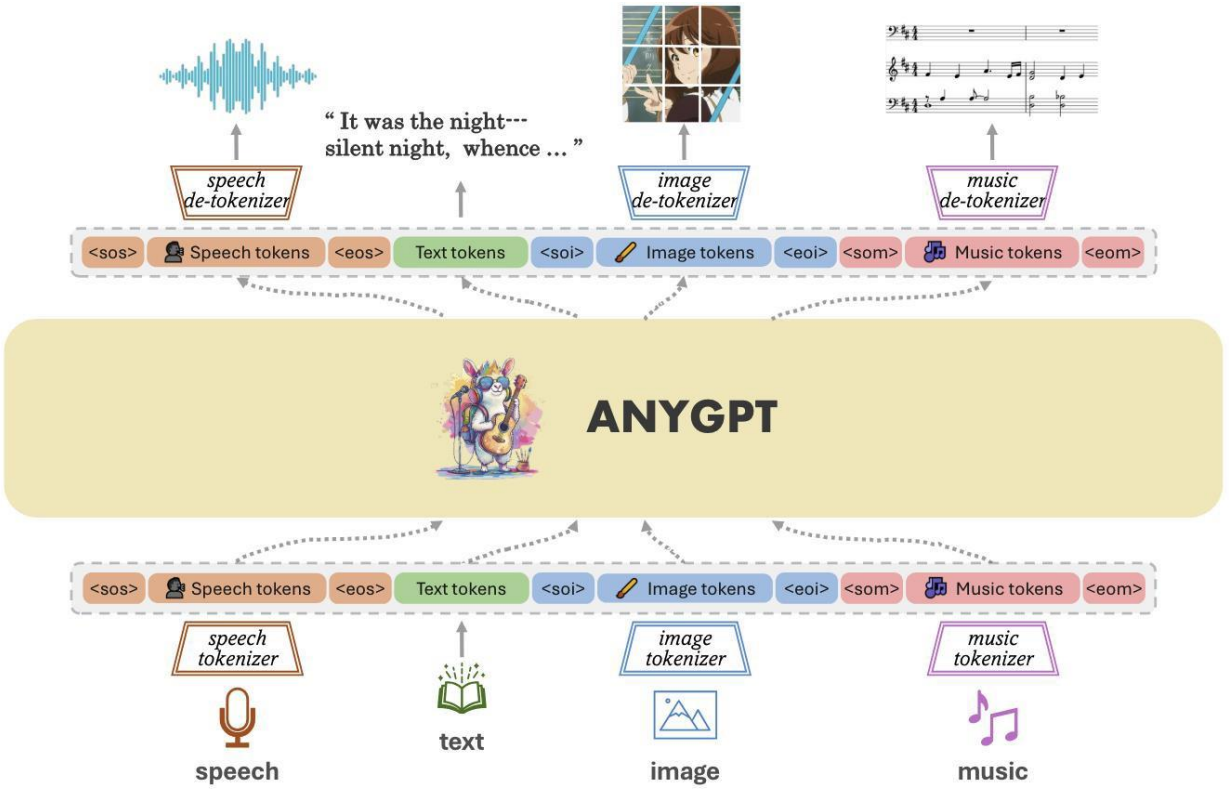
Unified Multimodal Encoder

- LanguageBind:
  - Embedding all modalities into a joint representation space of **Language**.
  - Well aligned modality representations can benefit LLM understanding



[1] LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. 2023

Tokenization

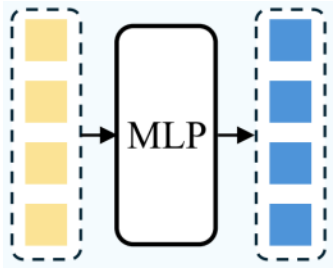


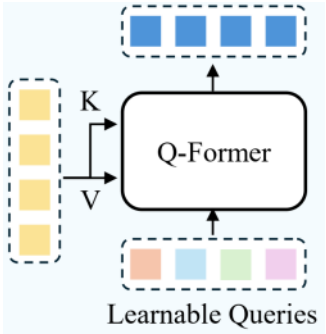
- Represent multimodal signals as discrete tokens in a codebook
  - **Advantages**: support unified multimodal signal understanding and generation in an auto-regressive next-token prediction framework

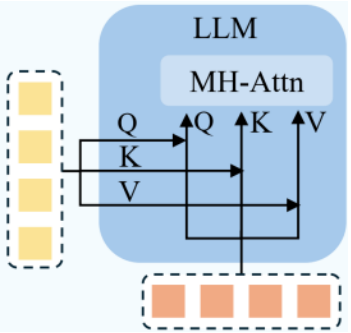
[1] AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. 2023

Tokenization

- Methods to Connect Multimodal Representation with LLM
  - Projecting multimodal (e.g., image) representations into LLM semantic space
    - Q-Former: BLIP-2, InstructBLIP, VisCPM, VisualGLM
    - Linear projection: LLaVA, MiniGPT-4, NExT-GPT
    - Two-layer MLP: LLaVA-1.5/NeXT, CogVLM, DeepSeek-VL, Yi-VL
    - Perceiver Resampler: Flamingo, Qwen-VL, MiniCPM-V, LLaVA-UHD
    - C-Abstractor: HoneyBee, MM1



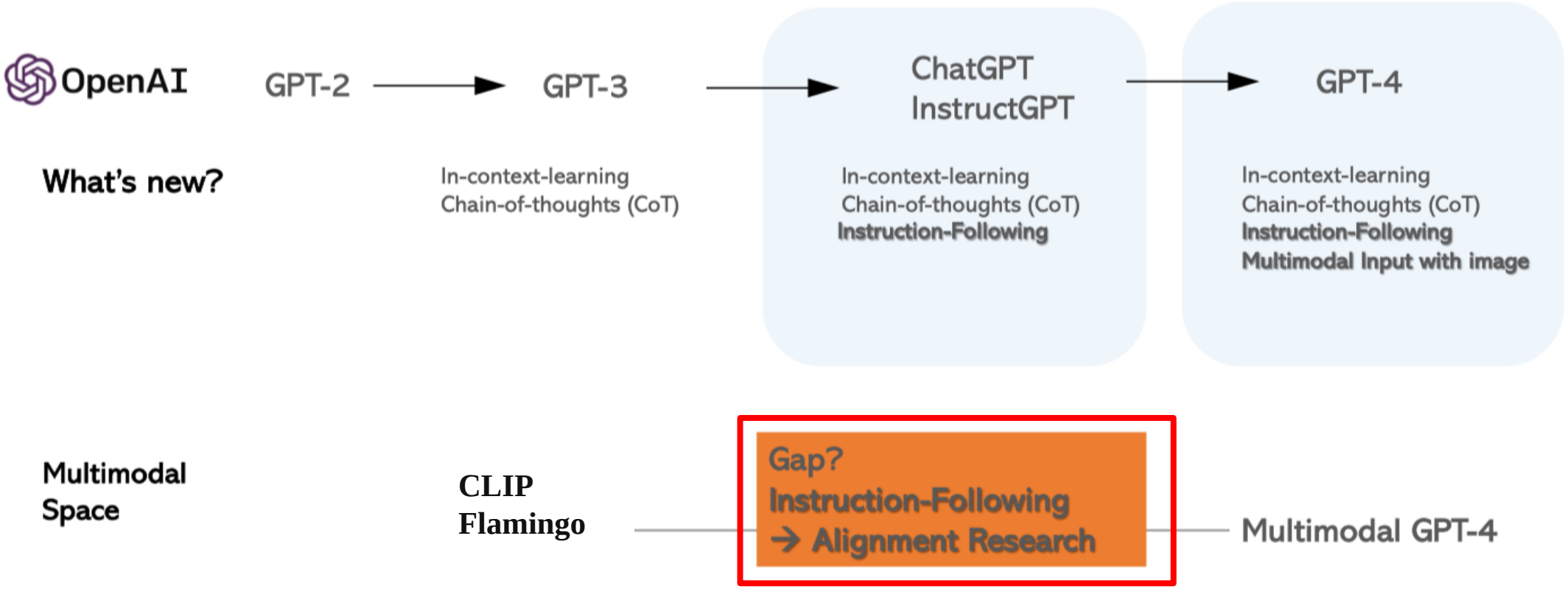




4 Training Multimodal LLMs

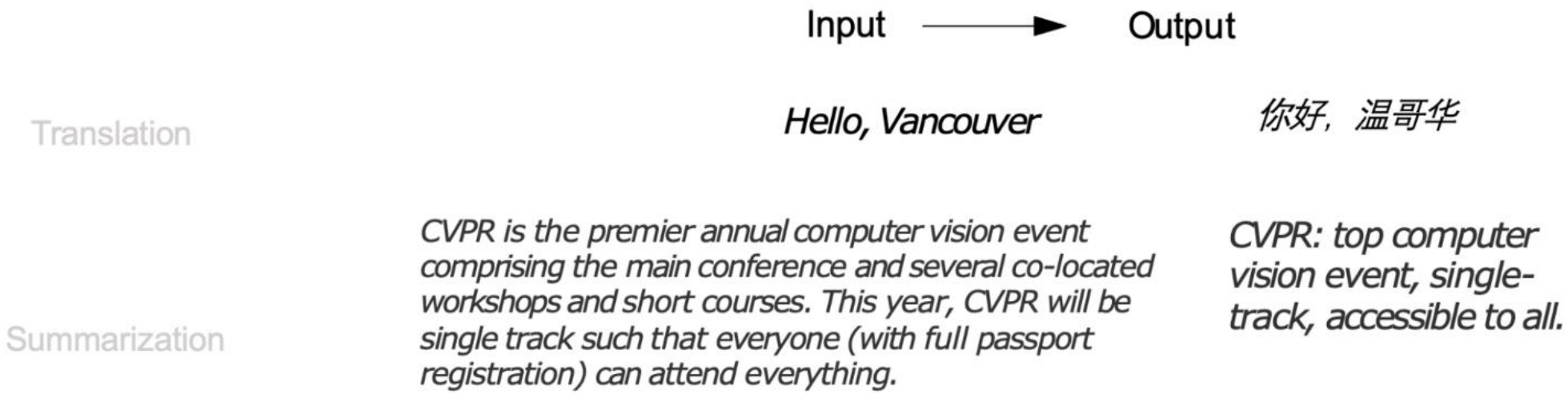
Alignment first, instruction following second — and data quality matters more than architecture.

Recap on Language Modeling



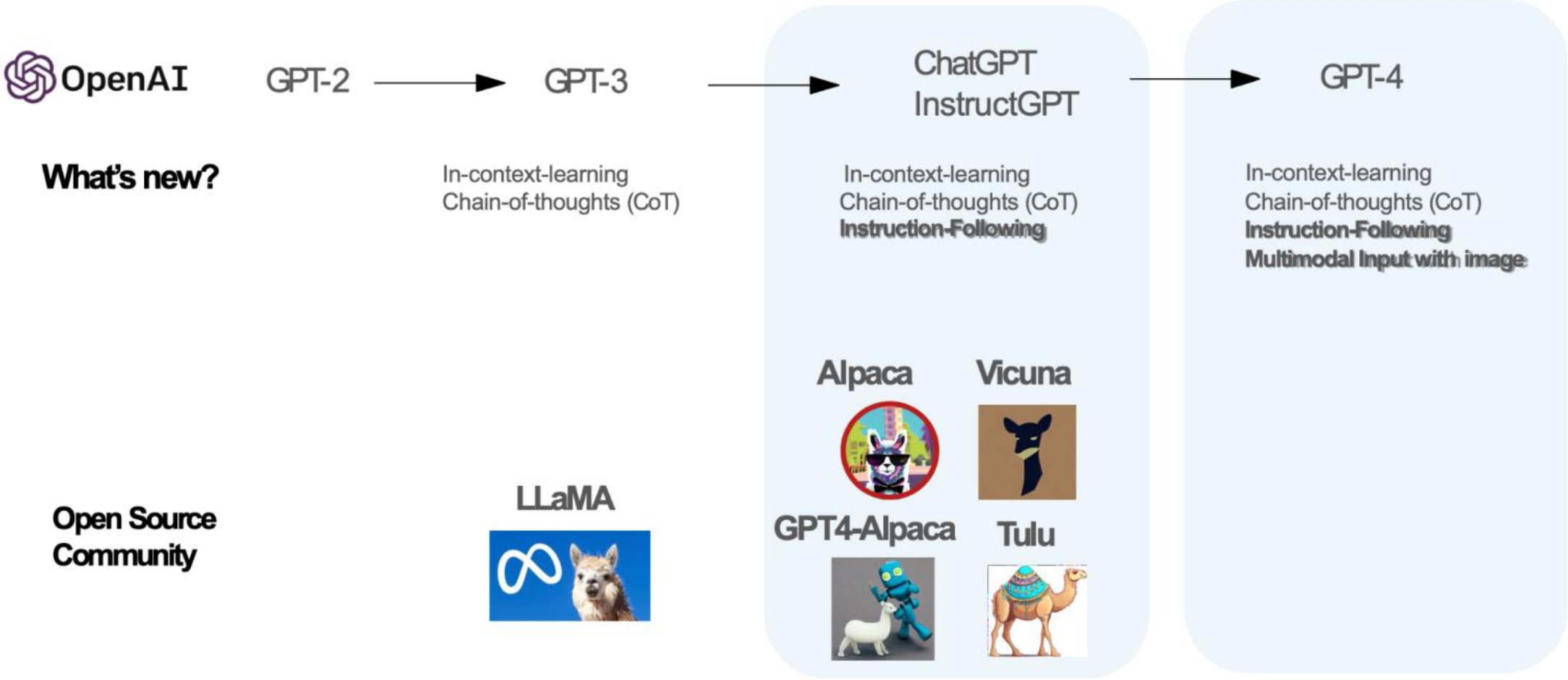
See lecture 3 and lecture 5 for more details

Instruction Tuning in LLM



- 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
Teacher	GPT-4 (text-only)	GPT-4 (text-only)

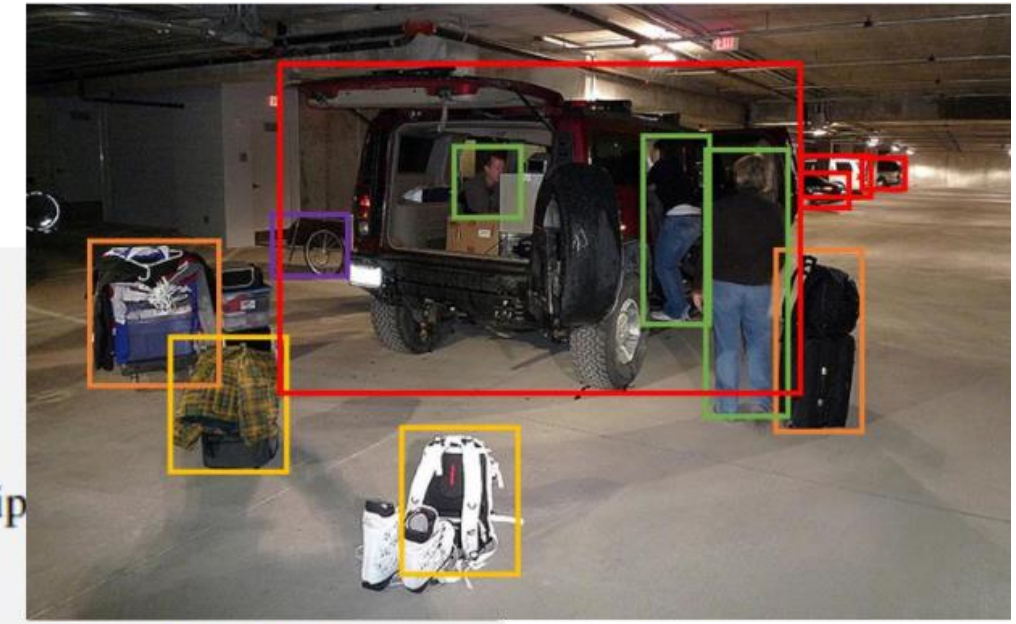
- 158K multimodal instruction following data (First & High Quality)

→ Multimodal Chatbot

GPT-assisted Visual Instruction Data Generation

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

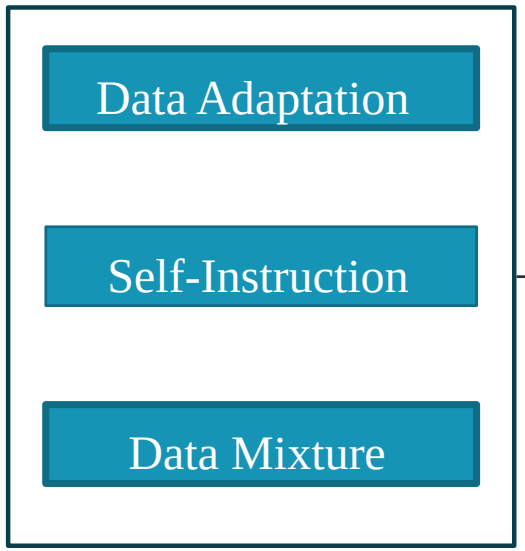
→ Text-only GPT-4



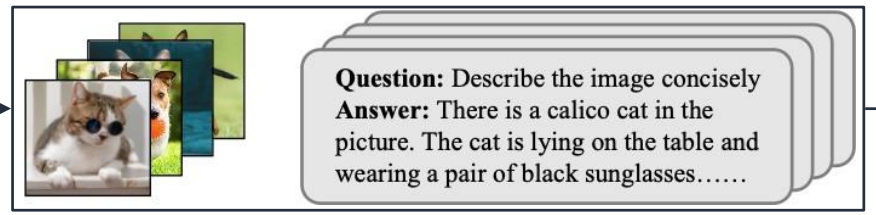
Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people 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]

MLLM Instruction Tuning Framework

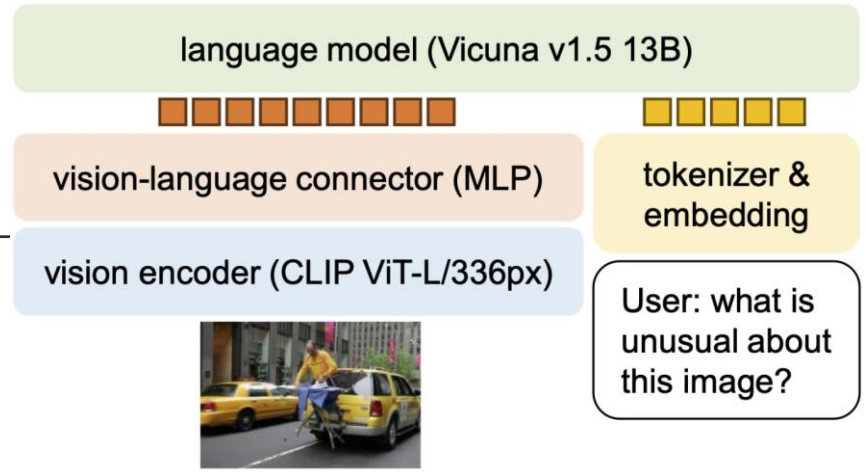
Data Construction



Visual Instruction-following data



Visual Instruction Tuning Framework
Example: LLaVA-1.5



Multi-stage training in Multimodal LLMs

Training Data

- **Image-text paired data (i.e. captions)**
- **Visual instruction tuning data (Visual QA)**

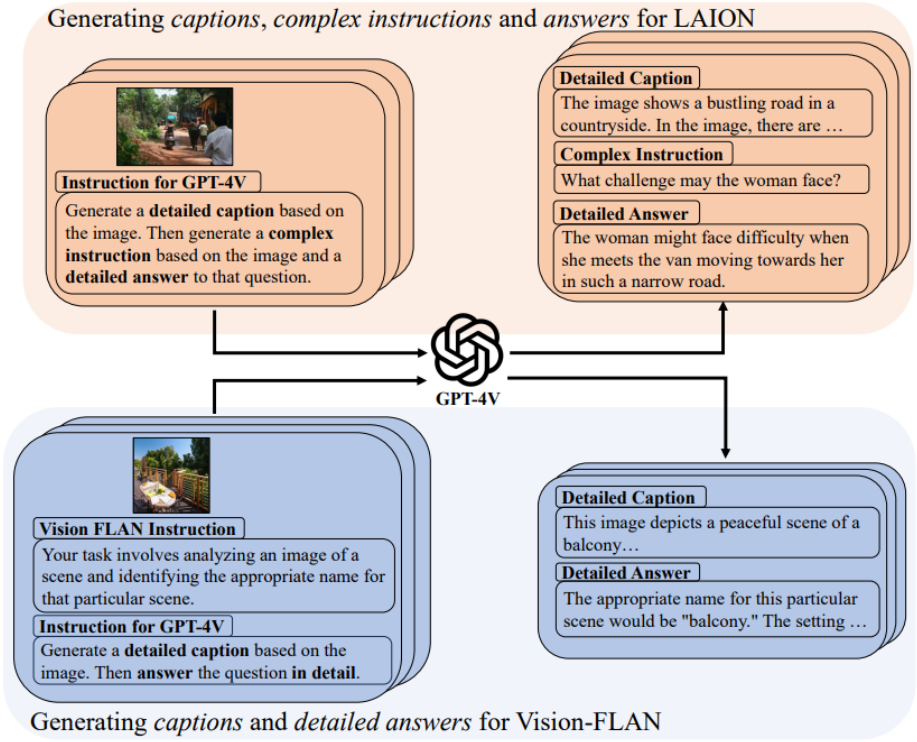


Figure 1: Pipeline for scaling up high-quality data. Prompts in the figure are shown for demonstration purpose. See the detailed prompt in Appendix A.2 and A.3.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, **Benyou Wang**. ALLaVA: Harnessing GPT4V-synthesized Data for A Lite Vision-Language Model. https://arxiv.org/abs/2402.11684

Pretraining Alignment Data (i.e. image caption)



COCO: Young children standing on a platform waiting for a train to arrive.
Adults and children watching a train slowly leave.
A family near a railroad track watching the train pass.
People waiting on a platform as a train pulls up.
A train station with a green chain on the tracks and children waiting for it to go by.

Coarse-grained Image-text

LLaVA: At a train station, a group of people, including both young children and adults, are standing on a platform waiting for a train to arrive. The train is already present on the tracks, partially visible on the right side of the image. Some of the people watch the train closely, while others seem to be patiently anticipating its departure.

There is a total of eight individuals waiting for the train, with one child in the middle of the platform and the others scattered around. A backpack can be found on the far left side of the platform, suggesting that someone may have set it down while waiting.

Coarse-grained Image-text

ShareGPT4V: The image captures a moment at a train station. A green train is on the tracks, moving away from the platform labeled "Platform 2". The train's motion is observed by people standing on the platform, their attention drawn to the departing vehicle.

A red signal stands on the tracks, its vibrant color contrasting with the green of the train. Signs are prominently displayed around the platform. One warns "Beware of Trains", a cautionary reminder of the powerful machines that frequent these tracks. Another sign advises "Please Use The Footbridge To Cross The Line. If You Require Assistance Please Ask The Station Staff", guiding passengers to safely navigate the station.

The sky above is a clear blue, providing a serene backdrop to the bustling activity of the station. In the distance, trees can be seen, adding a touch of nature to this man-made setting. The image is a snapshot of everyday life at a train station, capturing both its routine operations and its inherent charm.

Fine-grained Image-text

[1] ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. 2023.

Pretraining Alignment Data (i.e. image caption)

- **Coarse-gained Image-text**
  Data volume is large, the captions are shorts and noisy.

- **Fine-gained Image-Text**

  High quality, longer and more accurate descriptions, fine-gained alignment between different modalities.

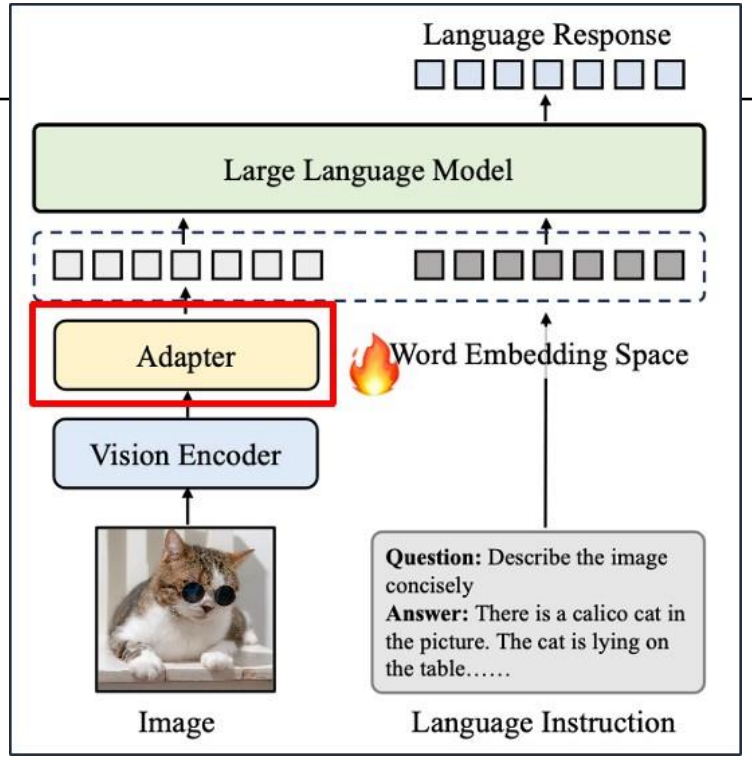
Dataset	Samples	Taxonomies
CC-3M	3.3M	Coarse-grained Image-Text
CC-12M	12.4M	Coarse-grained Image-Text
LAION-5B	5.9B	Coarse-grained Image-Text
SBU-Captions	1M	Coarse-grained Image-Text
ShareGPT4V-PT	1.2M	Fine-grained Image-Text
LVIS-Instruct4V	111k	Fine-grained Image-Text
ALLaVA	709k	Fine-grained Image-Text
MSR-VTT	200k	Video-Text
WavCaps	24k	Audio-Text
MMC-Instruction	600k	Chart-Text

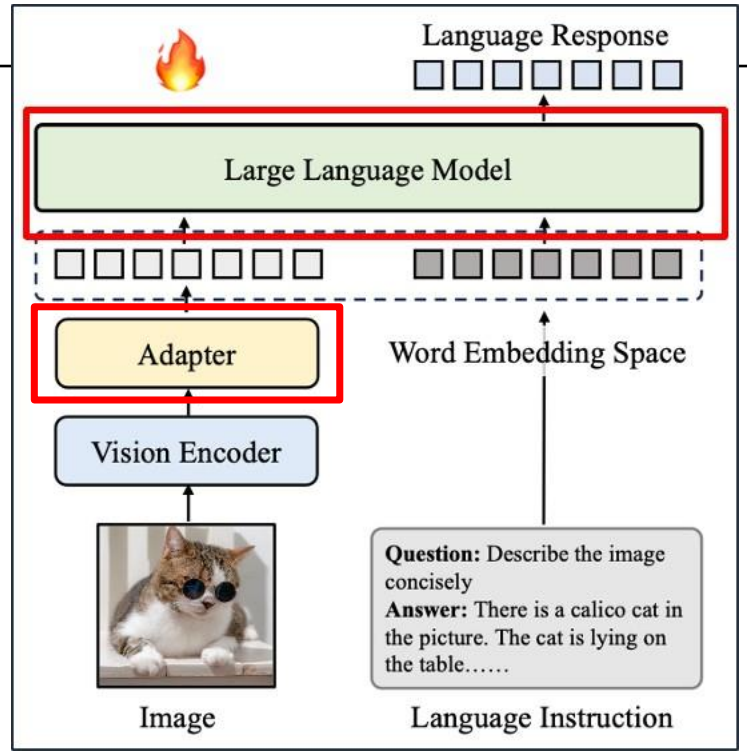
Instruction Tuning Data (visual QA)

Dataset	Size	Modalities	Constructions
LLaVA-Instruct-158k	158k	Image, Text	ChatGPT-generated
LRV-Instruction	400k	Image, Text	GPT4-generated
MMC-Instruction	600k	Chart, Text	GPT4-generated/adapted
Clotho-Detail	3.9k	Text, Audio	GPT4-generated
MACAW-LLM	119k	Image, Video, Text	GPT-3.5-turbo-generated
MIMIC-IT	2.8M	Image, Video, Text	ChatGPT-generated
StableLLaVA	126k	Image, Text	StableDiffusion & ChatGPT-generated
LAMM	196k	Image, PointCloud, Text	GPT4-generated
VIGC-LLaVA	1.8M	Image, Text	Model-generated
ALLaVA-4V	1.3M	Image, Text	GPT-4V

The Typical Training Paradigm

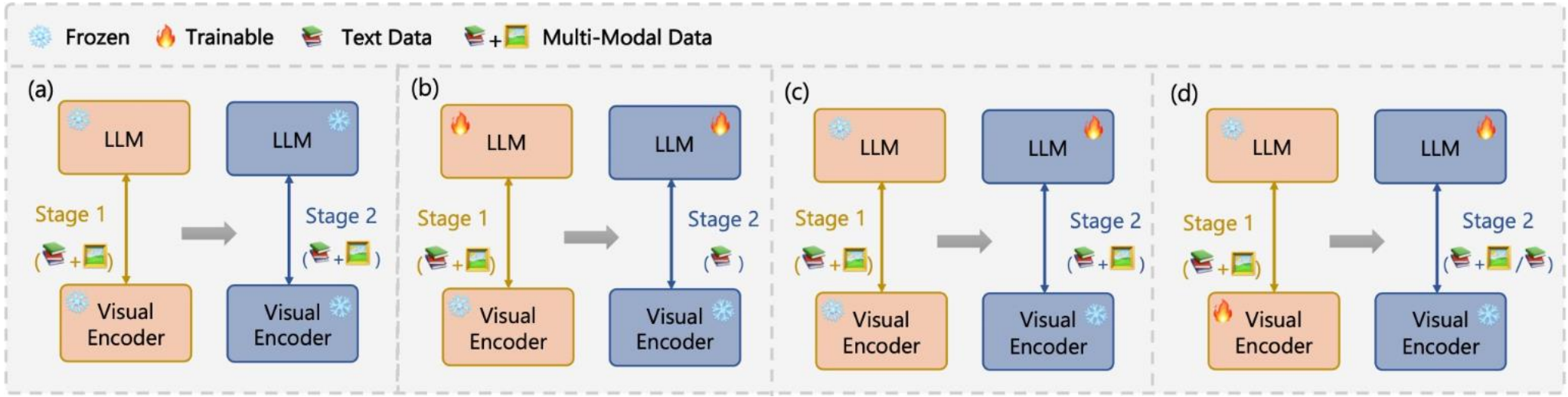
Stage1: Pretraining Stage		Stage2: Instruction Tuning Stage
Align different modalities, provide world knowledge		Teach models to better understand the instructions from users and fulfill the demanded tasks.





Training Paradigms

Training paradigms of popular multimodal large language models.

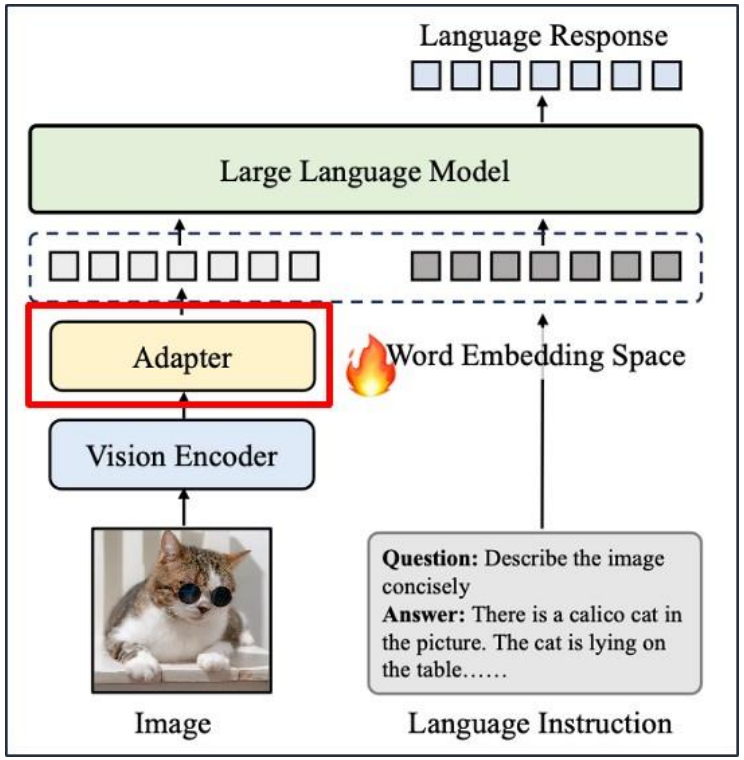


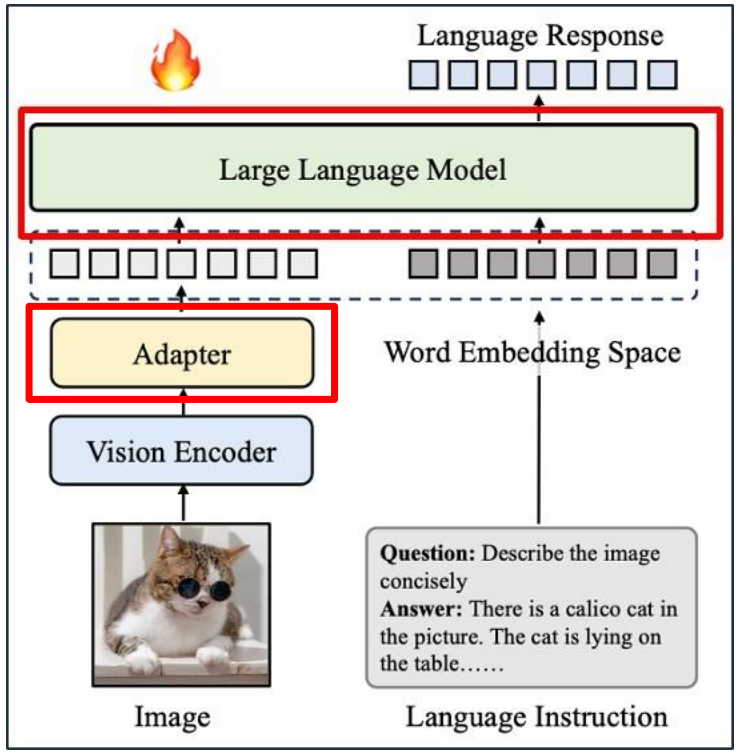
MiniGPT4 Kosmos-1 LLaVA mPLUG-Owl

1 *mPLUG-Owl: Language Models with Multimodality. 2023.*
2 *Visual Instruction Tuning. NeurIPS 2023.*
3 *MINIGPT-4: ENHANCING VISION-LANGUAGE UNDERSTANDING WITH ADVANCED LARGE LANGUAGE MODELS. 2023.*
4 *Language Is Not All You Need: Aligning Perception with Language Models. 2023.*

Pretraining Paradigms

Stage1: Pretraining Stage		Stage2: Instruction Tuning Stage
Align different modalities, provide world knowledge		Teach models to better understand the instructions from users and fulfill the demanded tasks.





5 Speech Language Models

Speech is the most natural interface, but it is continuous, fast and full of non-textual information.
How do we turn sound into tokens an LLM can model?

Why Speech Is Hard for Language Models

Continuous

A waveform has 16,000–24,000 samples per second. There are no natural “words” to tokenize.

Long

One second of speech ≈ 2–3 words of text, but 50–75 frames of a typical encoder.

Rich

Beyond words: speaker identity, emotion, prosody, laughter, background sound.

Real-time

Conversation needs replies within ~200–300 ms, overlaps and interruptions.

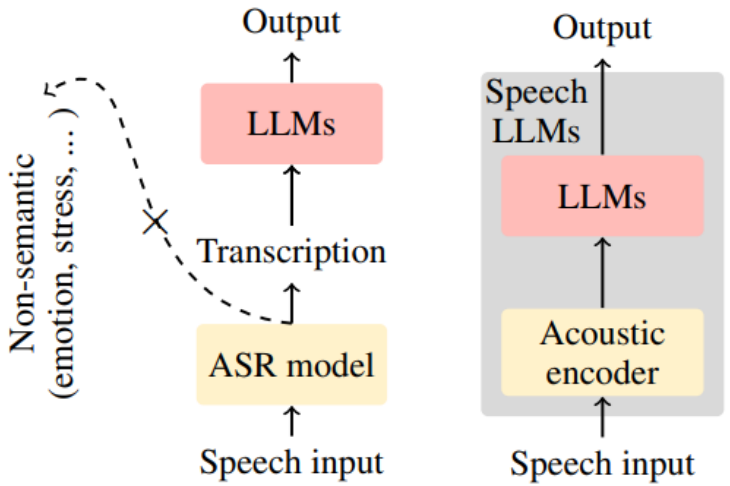
The central design question: what is a “speech token”? It decides what the LM can hear, say, and how fast.

Text ≈ 3–4 tokens per second of speech · HuBERT units: 50 per second · EnCodec: 75 frames × up to 32 codebooks per second

Hsu et al. HuBERT (2021) · Défossez et al. EnCodec (2022) · Défossez et al. Moshi (2024).

An Interesting Question: Speech Without Words?

- Can speech-to-speech models work in general without involving language (words)?



Some people argues that speech-to-speech is cool; but without language, it probably cannot think (reason and plan)!

An Interesting Question: A Speech-to-Speech Stress Test

- Can an end-to-end speech-to-speech model answer this question well?

Count from one to one hundred, pausing whenever you encounter a prime number.

Output 1, 2 [pause], 3 [pause], 4, 5 [pause], 6, 7 [pause], 8, 9, 10, 11 [pause], 12, 13 [pause], 14, 15, 16, 17 [pause], 18, 19 [pause], 20, 21, 22, 23 [pause], 24, 25, 26, 27, 28, 29 [pause], 30, 31 [pause], 32, 33, 34, 35, 36, 37 [pause], 38, 39, 40, 41 [pause], 42, 43 [pause], 44, 45, 46, 47 [pause], 48, 49, 50, 51, 52, 53 [pause], 54, 55, 56, 57, 58, 59 [pause], 60, 61 [pause], 62, 63, 64, 65, 66, 67 [pause], 68, 69, 70, 71 [pause], 72, 73 [pause], 74, 75, 76, 77, 78, 79 [pause], 80, 81, 82, 83 [pause], 84, 85, 86, 87, 88, 89 [pause], 90, 91, 92, 93, 94, 95, 96, 97 [pause], 98, 99, 100

Without textual planning, speech2speech might not solve this problem.

Cascade vs. End-to-End Spoken Dialogue

Cascade (2023): ASR → LLM → TTS



✘ loses emotion and speaker cues at the text bottleneck ✘ latency adds up (1–3 s) ✘ strict turn-taking

End-to-end (2024–): one model hears and speaks tokens



✔ keeps paralinguistics ✔ streaming, ~200 ms ✔ can listen while speaking (full duplex)

Everything now hinges on the tokenizer in the middle — the next four slides.

Radford et al. Whisper (2022) · OpenAI GPT-4o (2024) · Défossez et al. Moshi (2024) · Zeng et al. GLM-4-Voice (2024).

Semantic Tokens: Self-Supervised Units (HuBERT, wav2vec 2.0)



Discrete units



How HuBERT learns

- Mask spans of audio; predict the cluster ID of the masked frames (like BERT's masked LM).
- Targets come from k-means on MFCCs, then on its own features — iterate.

What units keep — and lose

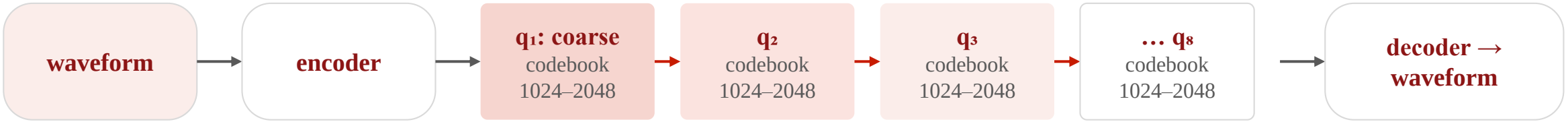
✔ phonetic content; LM-friendly; “textless NLP” (GSLM, 2021)

✘ speaker, emotion and fine prosody are mostly gone

→ need a vocoder (e.g. unit-HiFi-GAN) to speak

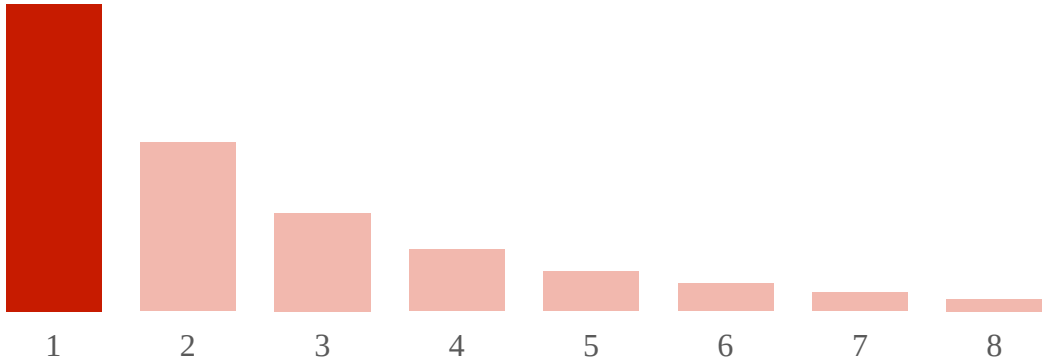
Baevski et al. wav2vec 2.0 (2020) · Hsu et al. HuBERT (2021) · Lakhotia et al. Generative Spoken Language Modeling (2021).

Acoustic Tokens: Neural Codecs and Residual Vector Quantization



each stage quantizes the residual left by the previous one

Reconstruction error after k codebooks



illustrative: each codebook removes part of the remaining error

Bitrate = frame rate × codebooks × bits

- SoundStream (2021): the first RVQ codec, 3–18 kbps
- EnCodec (2022): 75 Hz × 2–32 codebooks → 1.5–24 kbps
- Mimi (2024): 12.5 Hz × 8 × 11 bits = 1.1 kbps
- Drop late codebooks → lower bitrate, worse quality: a quality "dial"

Zeghidour et al. SoundStream (2021) · Défossez et al. EnCodec (2022) · Défossez et al. Moshi/Mimi (2024).

Semantic vs. Acoustic Tokens — and Hybrid Tokenizers

Semantic tokens

✔ content, easy for an LM to model long-range

✘ lose voice, emotion; need a vocoder

HuBERT / w2v-BERT units

Acoustic tokens

✔ reconstruct the exact voice and sound

✘ many tokens, hard long-range modeling

SoundStream, EnCodec

AudioLM (2022): generate in a hierarchy

semantic tokens (what is said)

coarse acoustic (who, how)

fine acoustic (details)

Hybrid: one codec, semantic first codebook

SpeechTokenizer (2023)

8-level RVQ at 50 Hz; level 1 distilled from HuBERT → content; levels 2–8 → timbre and details

Mimi (Kyutai, 2024)

12.5 Hz, streaming; 1 semantic VQ distilled from WavLM + 7 acoustic RVQ levels; 80 ms frames

Borsos et al. AudioLM (2022) · Zhang et al. SpeechTokenizer (2023) · Défossez et al. Moshi (2024).

Speech Tokenizers at a Glance

Tokenizer	Year	Frame rate	Codebooks	Bitrate	Type
HuBERT k-means units	2021	50 Hz	1 (K≈500)	≈0.45 kbps	semantic
SoundStream	2021	75 Hz	RVQ, ≤24	3–18 kbps	acoustic
EnCodec	2022	75 Hz	RVQ, 2–32	1.5–24 kbps	acoustic
SpeechTokenizer	2023	50 Hz	RVQ, 8	4 kbps	hybrid (level 1 ← HuBERT)
Mimi (Moshi)	2024	12.5 Hz	1 + 7	1.1 kbps	hybrid, streaming
GLM-4-Voice tokenizer	2024	12.5 Hz	1 (16K codes)	0.175 kbps	semantic (ASR-supervised)

Trend: fewer tokens per second (75 → 12.5 Hz), closer to text's 3–4 tokens/s — so the LM spends its capacity on meaning, not samples.

Numbers from the original papers; bitrate = frame rate × codebooks × $\log_2$(codebook size). HuBERT bitrate assumes K = 500.

Teaching an LLM to Speak: Interleaving Text and Speech

1 Extend the vocabulary	Add speech tokens to a text LLM; continue pre-training on speech *keeps text knowledge only if text data stays in the mix*
2 Interleave at word level	…the cat [speech: s_{12} s_{88} s_{5}] on the mat… *SpiritLM (2024): switches modality mid-sentence → transfer from text to speech*
3 Think in text, speak in speech	text tokens run slightly ahead of the speech tokens they produce *GLM-4-Voice alternates text and speech chunks; Moshi's "inner monologue"*

Text is the LLM's strongest modality — good speech LMs reuse it as a scaffold instead of replacing it.

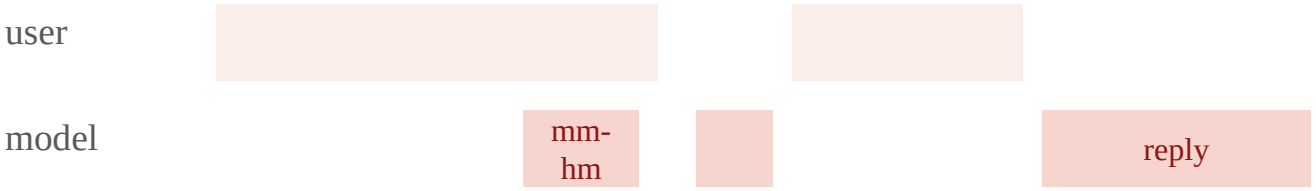
Nguyen et al. SpiritLM (2024) · Zeng et al. GLM-4-Voice (2024) · Défossez et al. Moshi (2024).

Full-Duplex Dialogue: Listening While Speaking (Moshi)

Half duplex: take turns



Full duplex: two streams modeled at once



Moshi's design

- Mimi tokens at 12.5 Hz for both speakers
- Inner monologue: time-aligned text tokens before its own speech tokens
- RQ-Transformer: a big temporal model + a small depth model over codebooks

Latency: 160 ms theoretical, ~200 ms in practice — within human turn-taking gaps (~200–300 ms).

Since then: Seeduplex, MiniCPM-o, DuplexOmni (2026) bring full duplex to omni models.

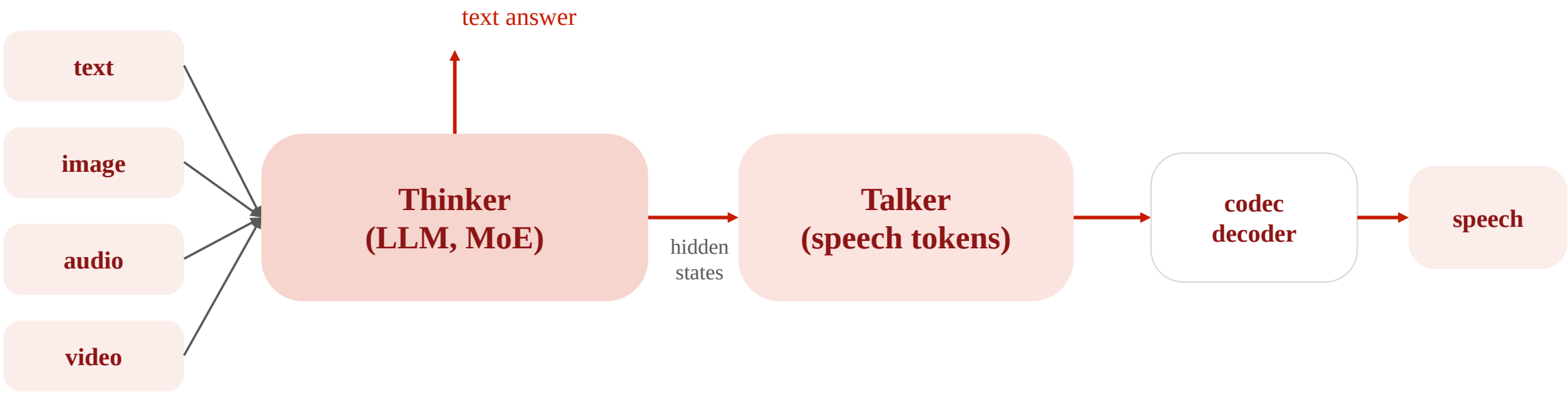
Défossez et al. Moshi: A Speech-Text Foundation Model for Real-Time Dialogue (Kyutai, 2024) · DuplexOmni (arXiv:2606.09186, 2026).

Seeduplex: Full-Duplex Voice Interaction (ByteDance Seed)



ByteDance Seed. Seeduplex: full-duplex voice interaction (official WeChat article, 2025).

Omni Models: The Thinker–Talker Architecture



- **GPT-4o (2024)**: one network for text, vision and audio; ~320 ms average voice latency.
- **Qwen2.5-Omni (2025)**: Thinker–Talker split; time-aligned positions (TMRoPE) for audio + video.
- **Qwen3-Omni (Sep 2025) → Qwen3.5-Omni (Apr 2026)**: MoE Thinker and Talker, ~234 ms latency, RVQ audio tokenizer, 256K context.

OpenAI GPT-4o (2024) · Qwen Team: Qwen2.5-Omni (2025), Qwen3-Omni (arXiv:2509.17765), Qwen3.5-Omni (arXiv:2604.15804).

End-to-End Speech LLMs: The 2024–2025 Landscape

- End-to-End Speech LLMs (expanding on the Speech2Speech discussion)
  - Mini-Omni (2024): Real-time end-to-end speech interaction with streaming output
  - GLM-4-Voice (Zhipu AI): Full-duplex voice conversation with emotional control
  - Echo-X: Speech-to-speech generation preserving speaker identity and emotion
  - Qwen2-Audio / Qwen2.5-Omni: Multi-turn audio understanding and generation

Our Take: What Omni Models Must Do Differently

1. Time is the native axis

Audio and video are continuous streams. Models must be always-on and streaming, not turn-based.

2. Don't squeeze everything through text

Emotion, speaker and environment sounds must survive tokenization — text is a scaffold, not a bottleneck.

3. Decouple thinking from talking

A slow reasoner and a fast speaker run asynchronously: think while listening, speak while thinking.

4. Evaluate interaction, not transcripts

Turn-taking, backchannels, interruptions and latency matter as much as WER or accuracy.

5. Data is the bottleneck

Real full-duplex conversations are scarce; synthetic TTS dialogue risks flat, uniform prosody.

6. An omni model is an agent in a stream

Hearing and seeing a live world is the first step toward acting in it (Part 7).

Lecture perspective, informed by Moshi (2024), Qwen-Omni (2025–2026), Seeduplex and DuplexOmni (2026).

Voice Will Be the Ultimate Interface

For every individual — in business (ToB) or consumer (ToC) products — the ultimate user interface is voice:

Voice-driven

speak instead of typing and tapping

Always available

start anytime, anywhere

Always listening

ambient, proactive help

Unlimited memory

remembers your context across years

Increasingly invisible

the device disappears; the assistant remains

Open questions: privacy of always-on listening, trust, and when an assistant should stay silent.

6 Generation and World Simulators

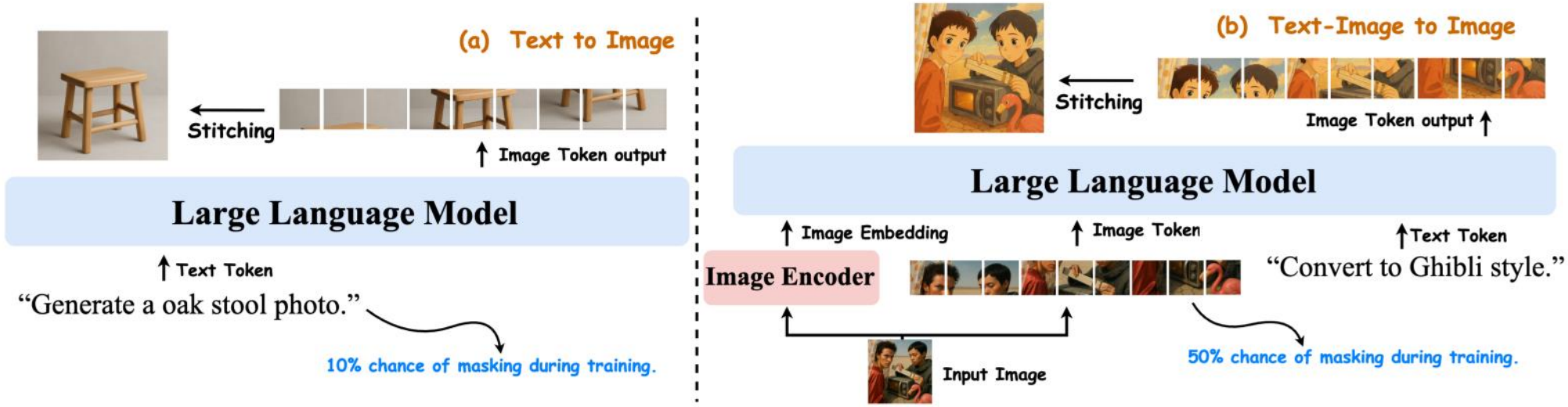
Understanding is half the story: the same models increasingly generate images, video — and whole simulated worlds.

Generated by Nano Banana



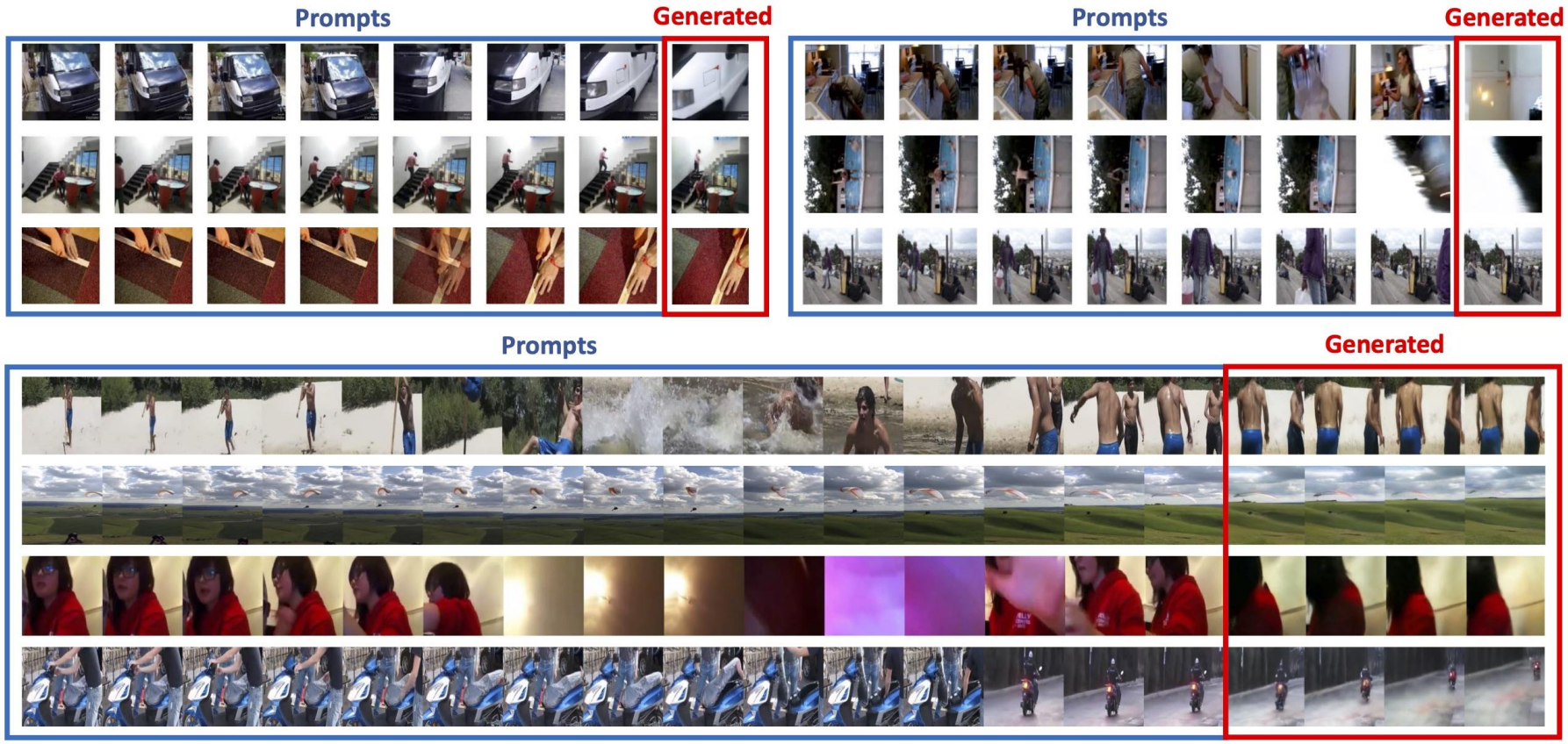
Prompt: “Draw a bar in a future world, like a spaceship: robots pour drinks and chat, outer space outside the window, and the screen shows doomsday scenes on Earth.”

Janus



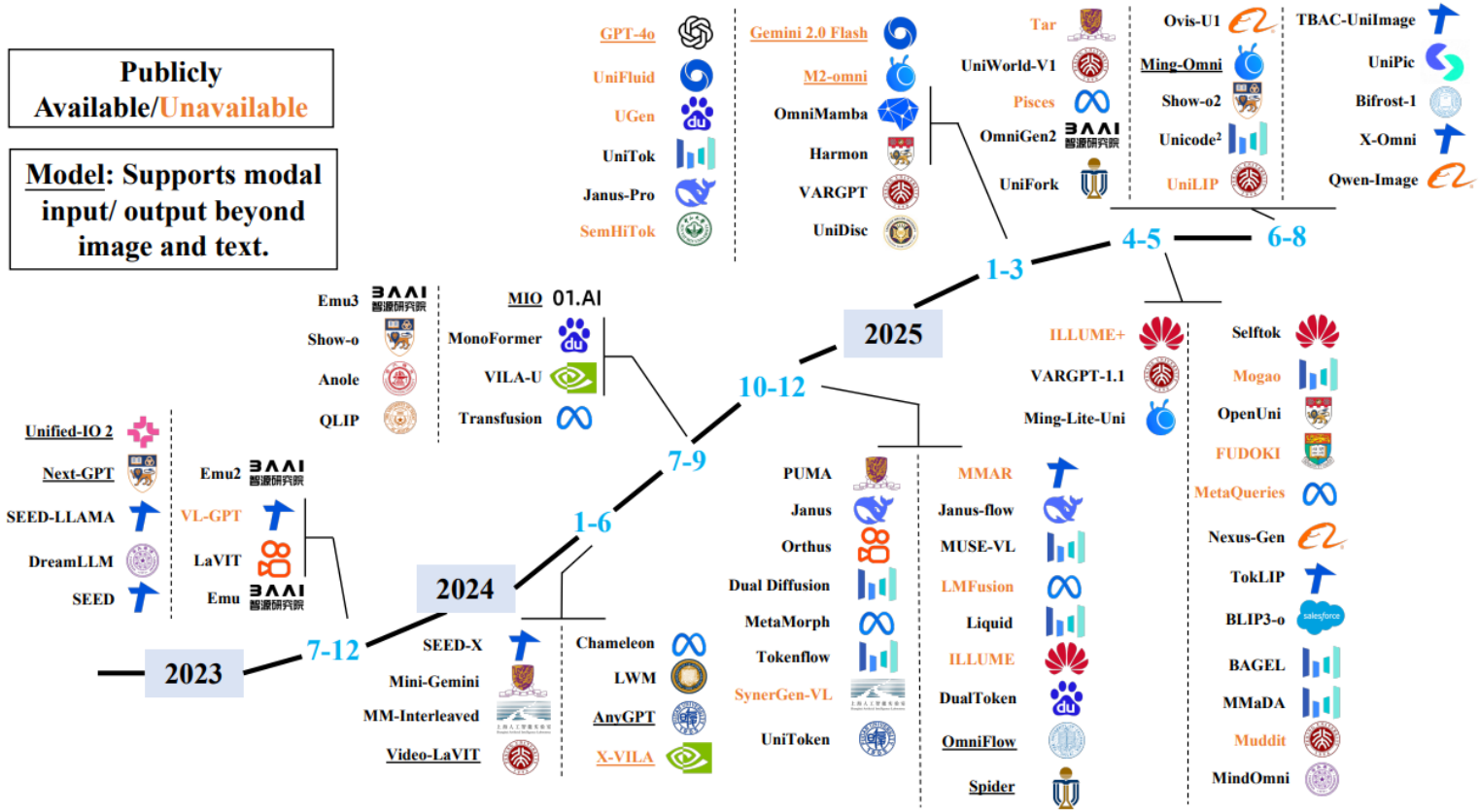
Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, Benyou Wang. ShareGPT-4o-Image: Aligning Multimodal Models with GPT-4o-Level Image Generation.

Large vision models using next-frame prediction



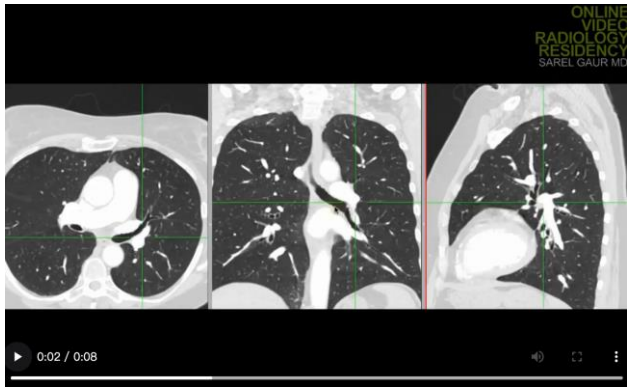
Frame predictions. LVM predicts the next frame (marked in red) given previous video frames as prompt. The results reveal the LVM can predict the video frames while considering dynamic objects and camera motion.

Unify multi-modal understanding and generation



Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Jiakui Hu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang. Unified Multimodal Understanding and Generation Models: Advances, Challenges, and Opportunities.

Medical Sora: Towards a Medical World Simulator



Medical image analysis



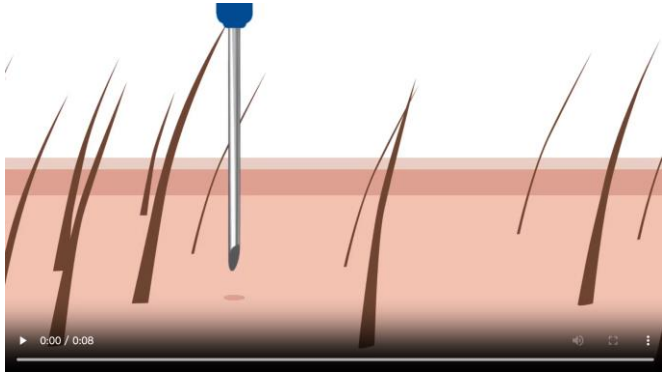
medical care



Instrument Usage



Operation



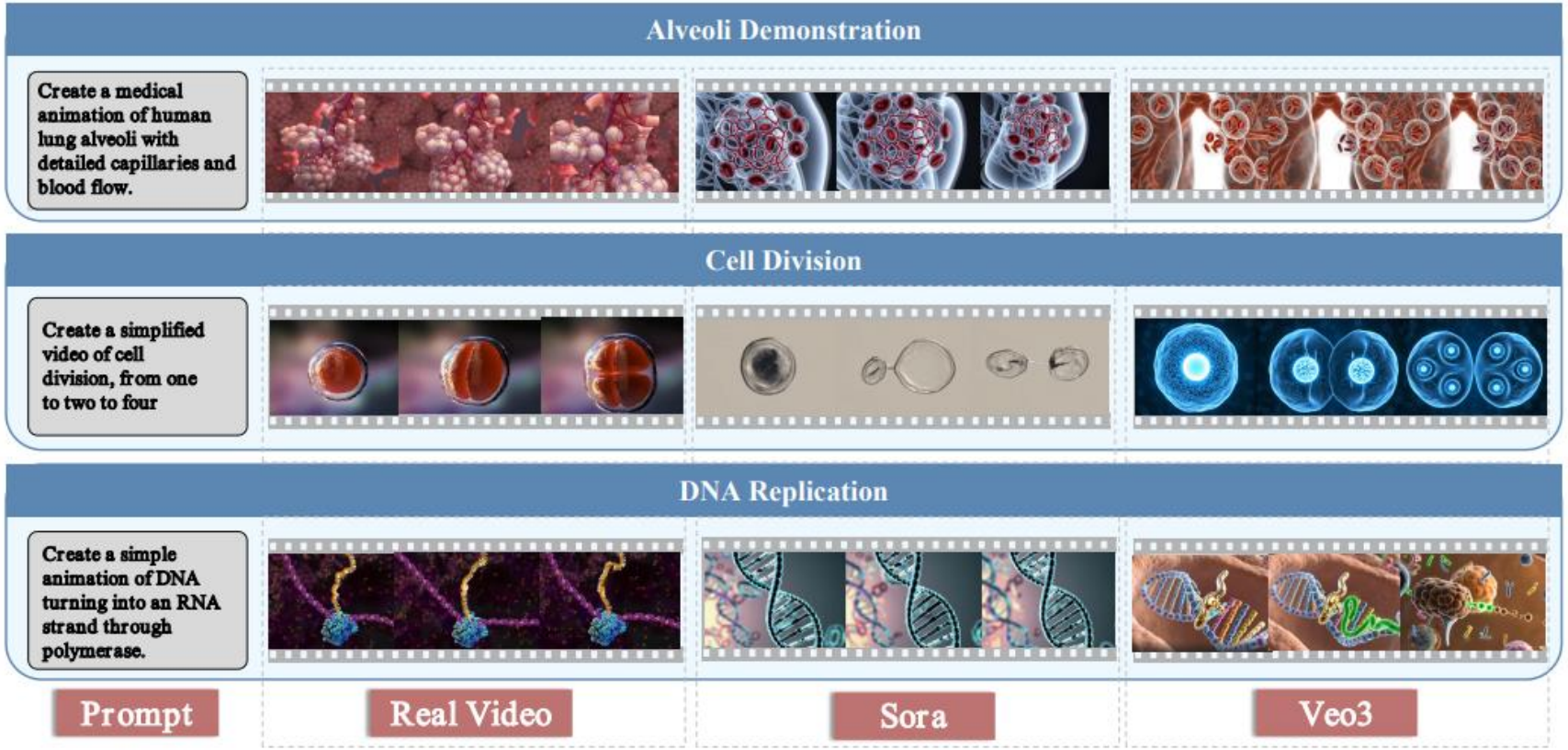
Medical Popularization



Medical Scenario

A medical world simulator offers immense application potential.

Micro world simulator



Rongsheng Wang, Minghao Wu, Hongru Zhou, Zhihan Yu, Zhenyang Cai, Junying Chen, Benyou Wang. MicroVerse: A Preliminary Exploration Toward a Micro-World Simulation.

7 Embodied Intelligence: VLA and World Models

To act in the physical world, a model must turn perception and language into motor commands — and predict what its actions will cause.

From Seeing to Acting: Why Embodiment Is Hard

Moravec's paradox

What is hard for humans (chess, calculus) is easy for AI; what is easy for a child (folding laundry, pouring water) is hard.

An LLM can explain how to make coffee — but cannot make it.

The data gap

- Text: ~10^{13} tokens on the web.
- Robots: Open X-Embodiment (2023) pooled ~1M trajectories from 22 robot types — tiny by comparison.

Three routes to physical intelligence

Borrow web knowledge

start from a VLM; add an action head → VLA

Learn physics from video

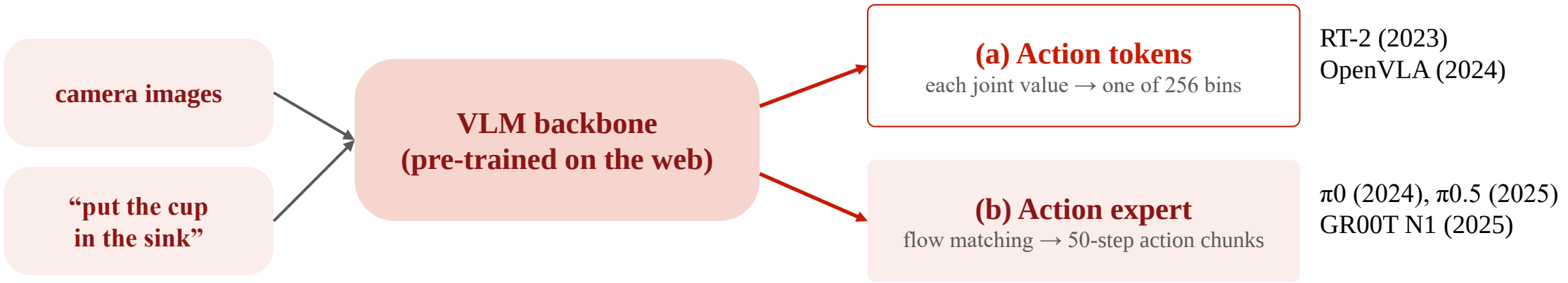
predict the future → world models, WAMs

Generate data

simulators and generated 3D worlds (Part 8)

Moravec, Mind Children (1988) · Open X-Embodiment Collaboration (2023).

Vision-Language-Action (VLA) Models



RT-2 (2023)
OpenVLA (2024)

π0 (2024), π0.5 (2025)
GR00T N1 (2025)

Why (b) won for dexterous control

- Continuous, smooth actions at high frequency (~50 Hz)
- Multi-modal action distributions (many valid grasps)
- Chunks: predict 0.5–1 s of motion at once → fewer slow VLM calls

Training recipe

- Co-train on web VQA + robot data so the model keeps its world knowledge
- Cross-embodiment pre-training, then task fine-tuning
- Increasingly: RL from real-world experience

Brohan et al. RT-2 (2023) · Kim et al. OpenVLA (2024) · Black et al. π0 (2024) · Physical Intelligence π0.5 (2025) · NVIDIA GR00T N1 (2025).

VLA Models: A Fast-Moving Timeline

Jul 2023
RT-2
actions as text tokens; web + robot co-training

Jun 2024
OpenVLA
7B open model, ~970K OXE episodes

Oct 2024
π0
flow-matching action expert

Mar 2025
Gemini Robotics · GR00T N1
frontier VLM → robot; dual-system humanoid model

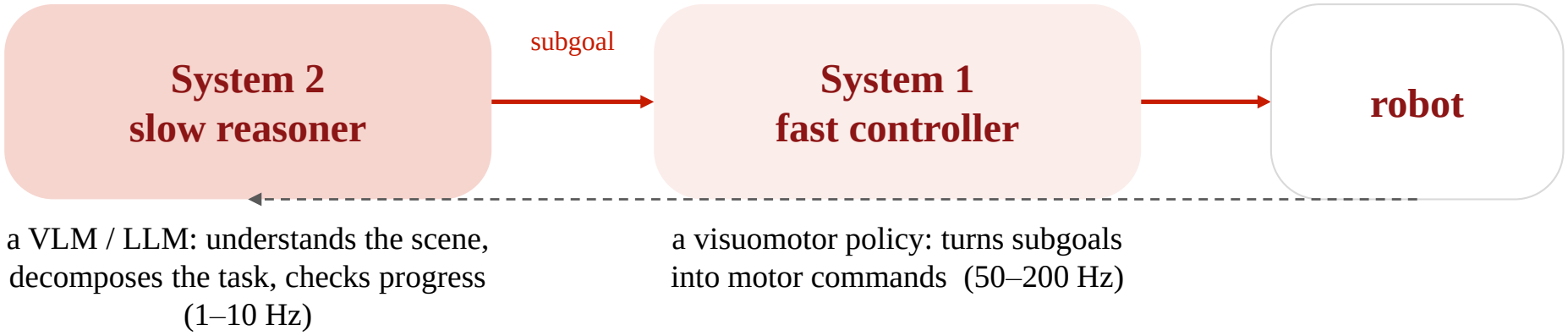
Apr 2025
π0.5
open-world generalization in new homes

2026
GR00T N1.7 · N2 · Gemini Robotics 2
humanoid whole-body control; N2 predicts the next observation too

The trend: from discrete tokens to continuous action experts — and now toward models that also predict the future.

Google DeepMind (2023, 2025, Jul 2026) · Kim et al. (2024) · Physical Intelligence (2024–2025) · NVIDIA GR00T (2025–2026; N2 figures as announced).

System 2 Plans, System 1 Acts



Examples

- Figure Helix and NVIDIA GR00T N1: a 7B-scale VLM + a small fast policy
- Gemini Robotics-ER (reasoning) + Gemini Robotics (action); version 2 in July 2026
- Generic frontier models as System 2: e.g. GPT-6 Astra planning for a π0.5 controller (next slide)

Kahneman, Thinking, Fast and Slow (2011) · Figure AI Helix (2025) · NVIDIA GR00T N1 (2025) · Google DeepMind Gemini Robotics (2025–2026).

A General Model as a Robot Brain? The Case of GPT-6 Astra

What it is

OpenAI frontier model released Sep 3, 2026; launched for computer use, coding and science — not as a robotics model.

External researchers then tested it on robots.

What evaluations report

- 12 embodied benchmarks: tabletop, humanoid, mobile, dexterous hands
- Best as System 2 in a closed loop with a fast System 1 (e.g. π0.5)
- *"Crossed the threshold of embodied competence, not yet practical embodied intelligence"*

What this means

Reasoning transfers

general models absorb spatial and task reasoning from scale

The bottleneck moves down

low-level control, latency and contact-rich skills remain hard

Cost matters

demonstration cost and execution time limit practical use

Epoch AI model card (2026) · Has GPT-6-Astra Achieved Embodied Intelligence? (arXiv:2610.22471) · Early Evaluations on RoboDojo (arXiv:2609.24170).

World Models: Learning How the World Changes

A world model predicts the next state of the world given the current state and an action: $s_{t+1} \approx f(s_t, a_t)$

Generative video *predict pixels*	**Latent predictive** *predict embeddings (JEPA)*	**Simulators for evaluation & data** *use imagination as a test bed*
• Genie 3 (Aug 2025): real-time interactive worlds, minutes of consistency • NVIDIA Cosmos (2025–2026): world foundation models for physical AI	• V-JEPA 2 (Jun 2025): self-supervised on video; V-JEPA 2-AC plans robot actions zero-shot • ignore unpredictable pixel details	• Waymo World Model (Feb 2026): driving scenarios built on Genie 3 • Physical Intelligence (Jun 2026): evaluating robot policies in world models

Uses: imagine before acting (planning), train policies inside the dream, and test robots without breaking real ones.

Google DeepMind Genie 3 (2025) · NVIDIA Cosmos (2025) · Meta V-JEPA 2 (2025) · Waymo (2026) · Physical Intelligence (2026).

World-Action Models: Predict the Future and the Action Together

VLA	(observation, instruction) → action	*reacts; no explicit model of consequences*
World model	(observation, action) → next observation	*imagines, but does not choose actions*
World-action model	(observation, instruction) → future video + actions	*imagines the outcome and how to get there*

Two designs

- **Cascaded**: a video model imagines the future; an inverse-dynamics model reads off the actions
- **Joint**: one model learns video and actions together

2026 examples

- DreamZero (Feb): video-diffusion backbone; >2× generalization vs. VLAs
- GR00T N2 (GTC 2026); AW-0 and GAWM (Oct 2026)
- Why it helps: video pre-training supplies a physics prior

World Action Models are Zero-shot Policies (DreamZero, arXiv:2602.15922) · From World Models to World Action Models (arXiv:2607.00836) · arXiv:2610.30136, 2610.37091.

How Do We Improve Spatial Intelligence?

Today's MLLMs read images well but reason about space poorly: metric distance, mental rotation, counting in 3D, egocentric vs. allocentric views.

Data
3D-grounded QA from scans and simulators; egocentric video with camera poses

Representation
add geometry: 3D tokens, depth, features from 3D foundation models (e.g. VGGT)

Memory
keep an explicit 3D map of what was seen instead of a bag of frames

Objectives
predict depth, motion and the next view — not only the next word

Reasoning
think with images: sketch, crop, re-render views while reasoning

Embodiment
learn by acting: RL with physical outcomes as the reward

Our view: spatial intelligence will come from closing the loop — perceive → reconstruct → imagine → act → verify.

Yang et al. Thinking in Space (VSI-Bench, 2024) · Wang et al. VGGT (CVPR 2025) · Fei-Fei Li, From Words to Worlds (2025).

8 3D Understanding and Generation

Images are projections of a 3D world. Reconstructing, understanding and generating 3D is the foundation of spatial intelligence.

Why 3D? Images Are Only Projections

Representation	What it stores	Typical use
Point cloud	unordered xyz points	LiDAR, reconstruction
Mesh	vertices + faces	games, CAD, printing
NeRF	a network maps (x, y, z, view) → color	photorealistic views
3D Gaussian splatting	millions of soft 3D blobs	real-time rendering
Structured latents	voxel grid of learned features	3D generation

Two directions, one goal: reconstruct 3D from images (understanding) and create 3D from text or images (generation).

Robots, AR/VR, games, digital twins and spatial reasoning all need both.

Mildenhall et al. NeRF (2020) · Kerbl et al. 3D Gaussian Splatting (2023) · Xiang et al. TRELLIS (2024).

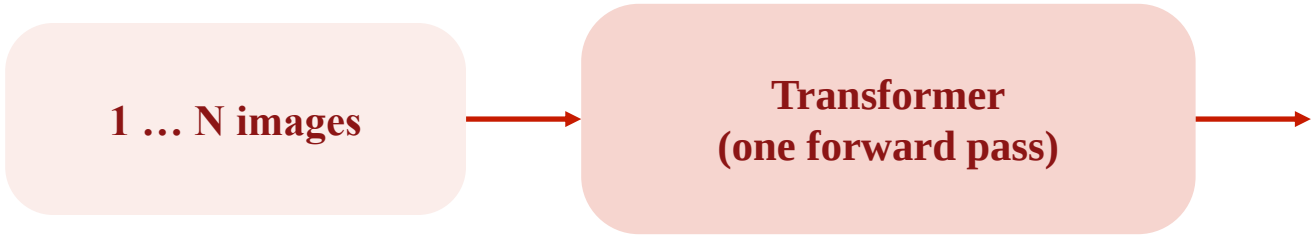
3D Reconstruction: From Hours of Optimization to One Forward Pass

Classic: Structure-from-Motion (e.g. COLMAP)



minutes to hours · fragile with few views, low texture or motion

Learned: a single network predicts the geometry



camera poses · depth maps · point maps · point tracks

in under a second

DUSt3R (CVPR 2024)
predicts aligned point maps from an image pair — no calibration needed

VGGT (CVPR 2025 best paper)
all 3D attributes from one to hundreds of views, feed-forward

Schönberger & Frahm COLMAP (2016) · Wang et al. DUSt3R (CVPR 2024) · Wang et al. VGGT (CVPR 2025).

3D Generation: Objects and Worlds

Objects

Multi-view then reconstruct

a 2D diffusion model draws consistent views; a reconstruction model lifts them to 3D

Native 3D latents

diffusion directly in a 3D latent space — TRELLIS (2024), Hunyuan3D 2.0 (2025)

From one photo

SAM 3D (Meta, Nov 2025): objects and human bodies from a single image

Worlds

- **World Labs Marble** (Nov 2025): persistent, navigable 3D worlds from text, images or video, as Gaussian splats; World API in Jan 2026
- **Genie 3** (Aug 2025): interactive worlds generated frame by frame, explorable in real time

The key distinction: explicit 3D (you can measure and simulate it) vs. pixels that only look 3D (consistency is not guaranteed).

For robots and engineering we need the former; for media the latter is often enough.

Xiang et al. TRELLIS (2024) · Tencent Hunyuan3D 2.0 (2025) · Meta SAM 3D (2025) · World Labs Marble (2025) · Google DeepMind Genie 3 (2025).

Language-Driven 3D: The LLM as a Scene Programmer

"A cozy reading room: two armchairs facing the window, a lamp between them" → **LLM plans a scene graph and layout (objects, sizes, relations)** → **retrieve or generate each 3D asset** → **solve constraints + physics check** → **simulator-ready 3D scene**

Intermediate representation (LLM output)

```
{"objects": [
  {"id": "chair_1", "type": "armchair"},
  {"id": "chair_2", "type": "armchair"},
  {"id": "lamp", "type": "floor_lamp"}],
 "relations": [
  ["chair_1", "faces", "window"],
  ["lamp", "between", "chair_1", "chair_2"]]}
```

Examples

- Holodeck (2024): LLM-generated 3D environments for training embodied agents
- 3D-LLM, ShapeLLM: language models that read point clouds
- Agentic scene builders (2026) write code in Blender or simulators

Yang et al. Holodeck (CVPR 2024) · Hong et al. 3D-LLM (2023) · Qi et al. ShapeLLM (2024).

Thinking Ahead: What Should Future Models Do in 3D?

1. **One world representation** — the same 3D-consistent latent for understanding, generation and action
2. **Executable worlds** — generated scenes must be simulatable — generation becomes a data engine for robots
3. **Language as interface, geometry as substrate** — talk in words, compute in 3D
4. **Close the loop** — perceive → reconstruct → imagine → act → verify, continuously
5. **Evaluate by doing** — the best test of a 3D world model is whether an agent trained in it succeeds in reality

Language models learned the world from text; the next models must learn it from space, time and action.

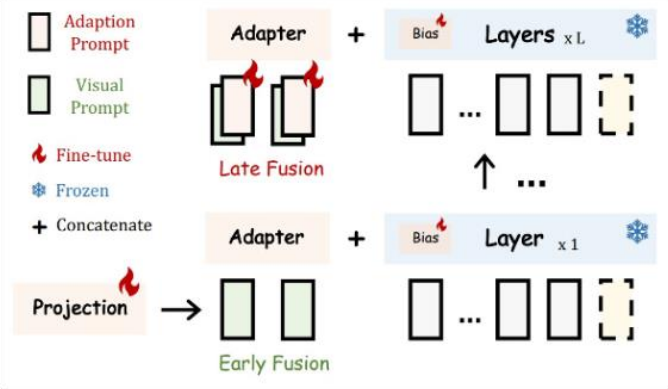
Lecture perspective · Fei-Fei Li, From Words to Worlds: Spatial Intelligence Is AI's Next Frontier (2025).

9 Research Directions and Outlook

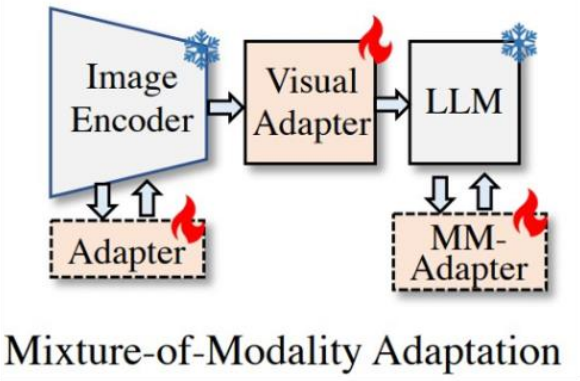
Where multimodal AI is going — and what is still open for you to work on.

Efficient multimodal MLLMs:

- **Projection** (see Fuyu) or **token reductions** [1]
- 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

1] Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael X. Guan, Benyou Wang. Less is More: A Simple yet Effective Token Reduction Method for Efficient Multi-modal LLMs. COLING 2025. https://aclanthology.org/2025.coling-main.508/

Generating multimodal outputs (vision, action, etc.)

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.

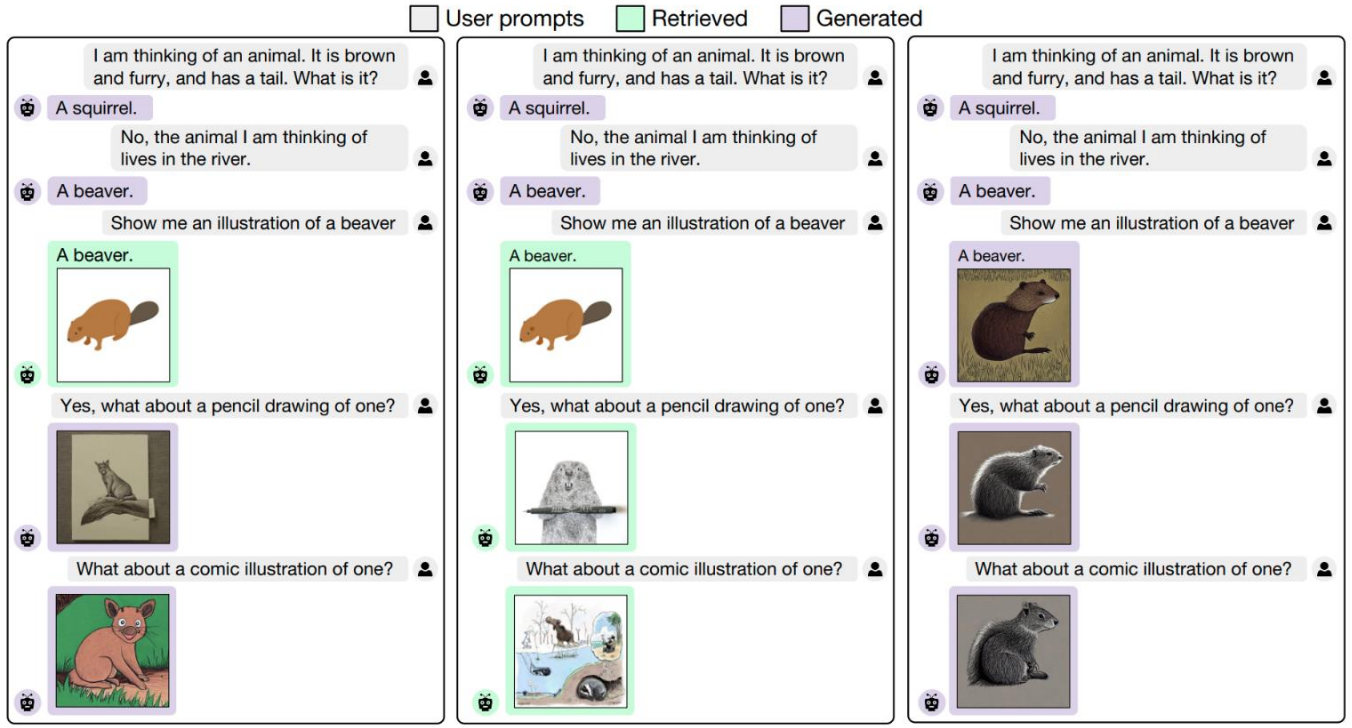
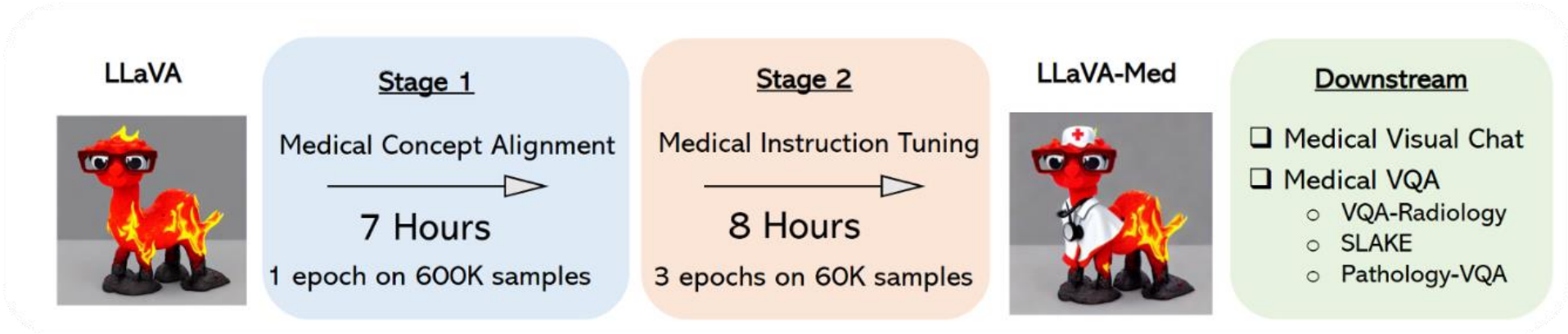


Figure 1: Our model is capable of generating text, retrieving images, generating novel images, and interleaving results into coherent multimodal dialogue.

Vertical Applications (e.g. medical 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



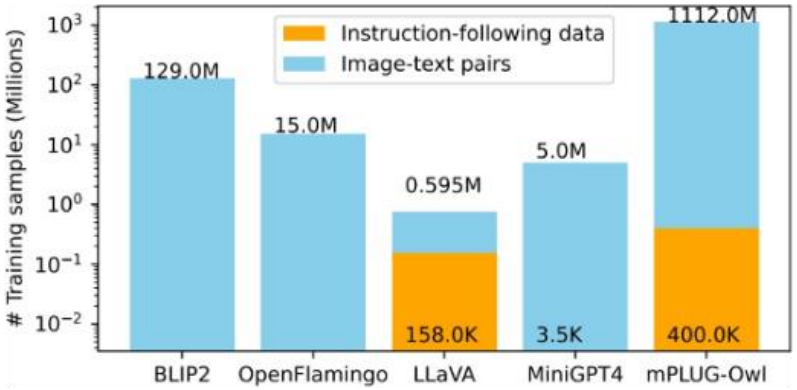


User	What is this image about?
LLaVA-Med	The image is a coronal view of a computed tomography (CT) scan of the abdomen and pelvis. It provides a detailed view of the internal organs and structures within the abdominal and pelvic cavities.
User	What is abnormal about this image?
LLaVA-Med	The abnormality 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?

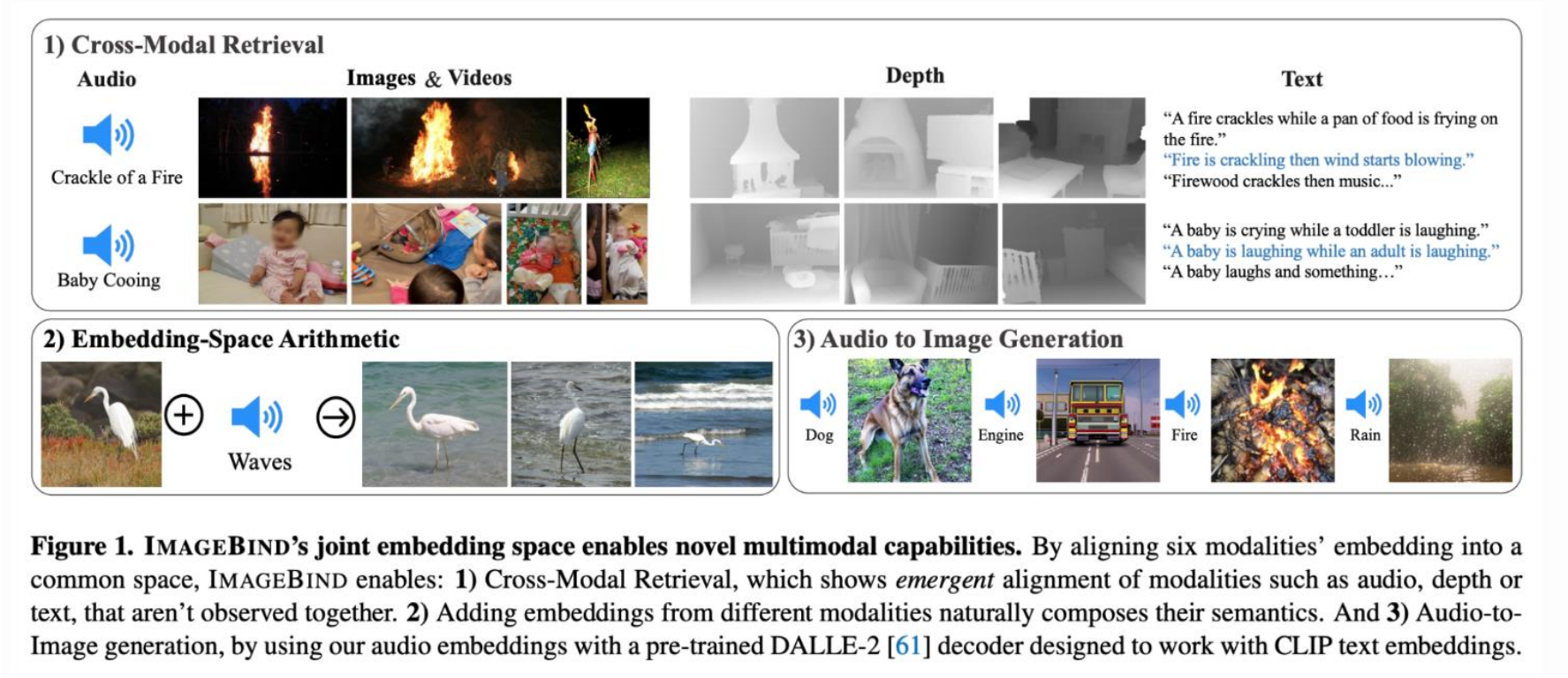


Figure 1. IMAGEBIND's joint embedding space enables novel multimodal capabilities. By aligning six modalities' embedding into a common space, IMAGEBIND enables: **1)** Cross-Modal Retrieval, which shows *emergent* alignment of modalities such as audio, depth or text, that aren't observed together. **2)** 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.

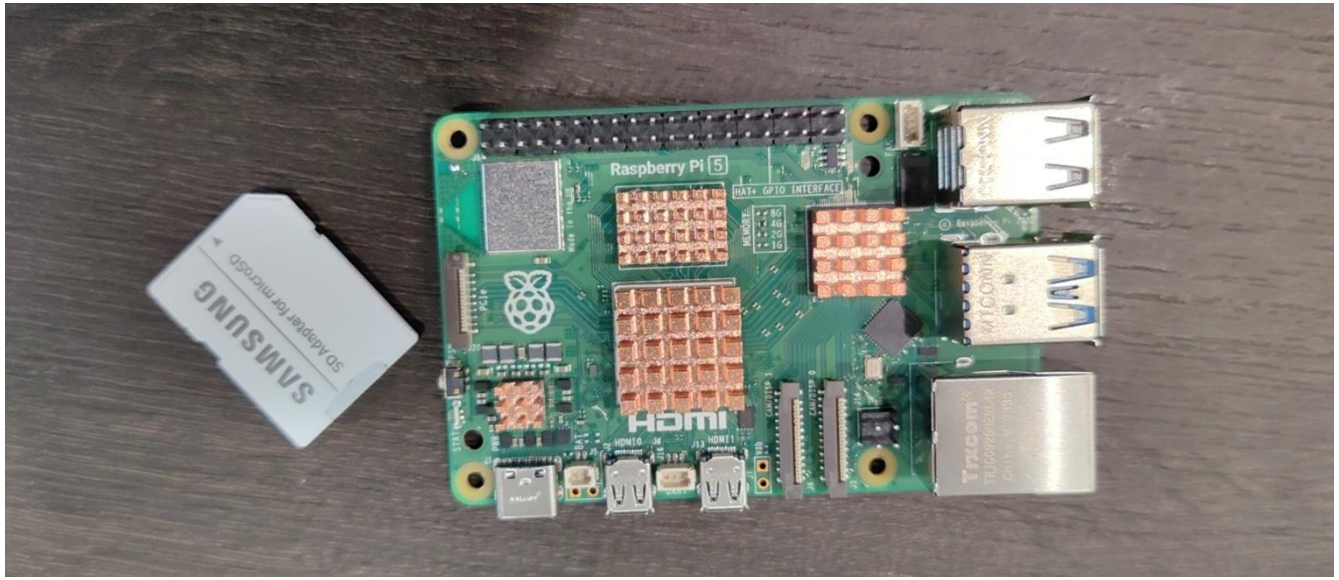
Video Understanding and Multimodal Reasoning

- **Video Understanding** at Scale
  - LongLLaVA: Hybrid architecture (Mamba + Transformer) for processing 1000+ images efficiently
  - VideoLLaVA / LLaVA-OneVision: Extending image MLLMs to video with temporal modeling
- **Multimodal Reasoning** with RL (Related: Lecture 8-1 Large Reasoning Models)
  - Video-R1: Reinforcement learning for video reasoning (ordered vs. shuffled frame experiments)
  - Visual Chain-of-Thought: Extending CoT prompting to visual inputs for step-by-step spatial reasoning
  - MathVista / MMMU: Benchmarks for multimodal reasoning (math, science, graduate-level QA)

Edge-device Multi-modal LLMs

MLLMs for Edge devices

Running on edge devices with streaming capability



A naïve demo



MLLMs running on edge devices

Multimodal AI in 2026: What Changed

Omni is the default	real-time speech + vision models (Qwen3.5-Omni, Apr 2026); full-duplex research systems
World models became products	Genie 3 → Waymo World Model; World Labs' World API
From VLA to world-action models	DreamZero, GR00T N2: predict the future, not only the action
General models go physical	GPT-6 Astra evaluated as a robot planner; Gemini Robotics 2
3D in one forward pass	feed-forward reconstruction and native 3D generation are now standard tools

Open problems for you: data for physical interaction · evaluation of interaction · efficient always-on models · safety of acting agents

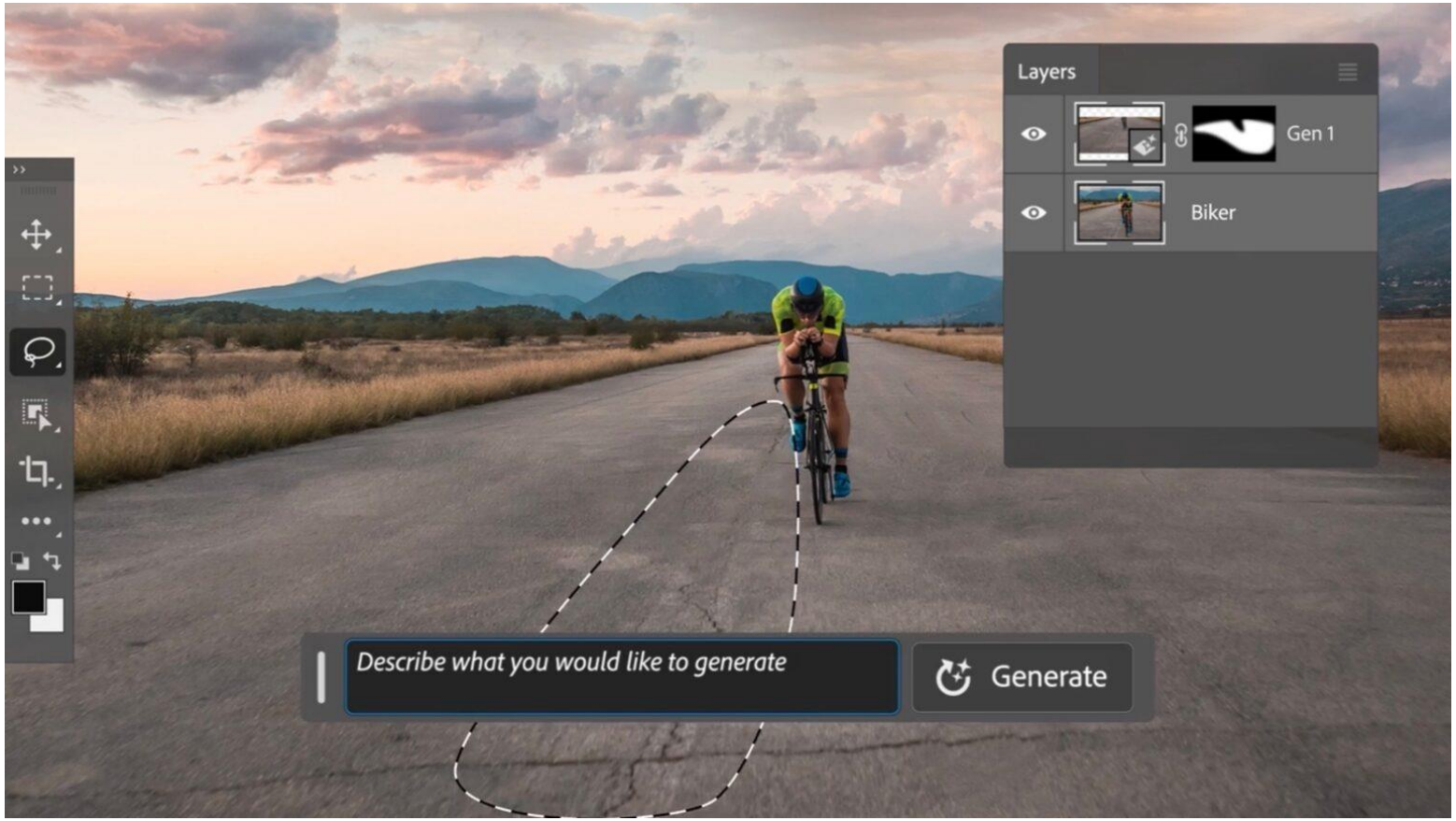
Qwen3.5-Omni (Apr 2026) · Waymo (Feb 2026) · World Labs (Jan 2026) · DreamZero (Feb 2026) · NVIDIA GTC 2026 · Google DeepMind (Jul 2026) · arXiv:2610.22471.

Final remarks from multi-modal interactions

Why We Need Multimodal LLMs

- Speech is more convenient for interaction;
  - See HuatuoGPT applications
- Video is language-agnostic
  - We could also enjoy the fun from TikTok in India without language barrier
- Data with other modality needs a **language interface** to interact;
  - Draw with language interface-based AI tools

Adobe Firefly turns Photoshop into effortless AI image generator



https://www.cultofmac.com/news/adobe-firefly-photoshop-ai-image-generator

Final remarks



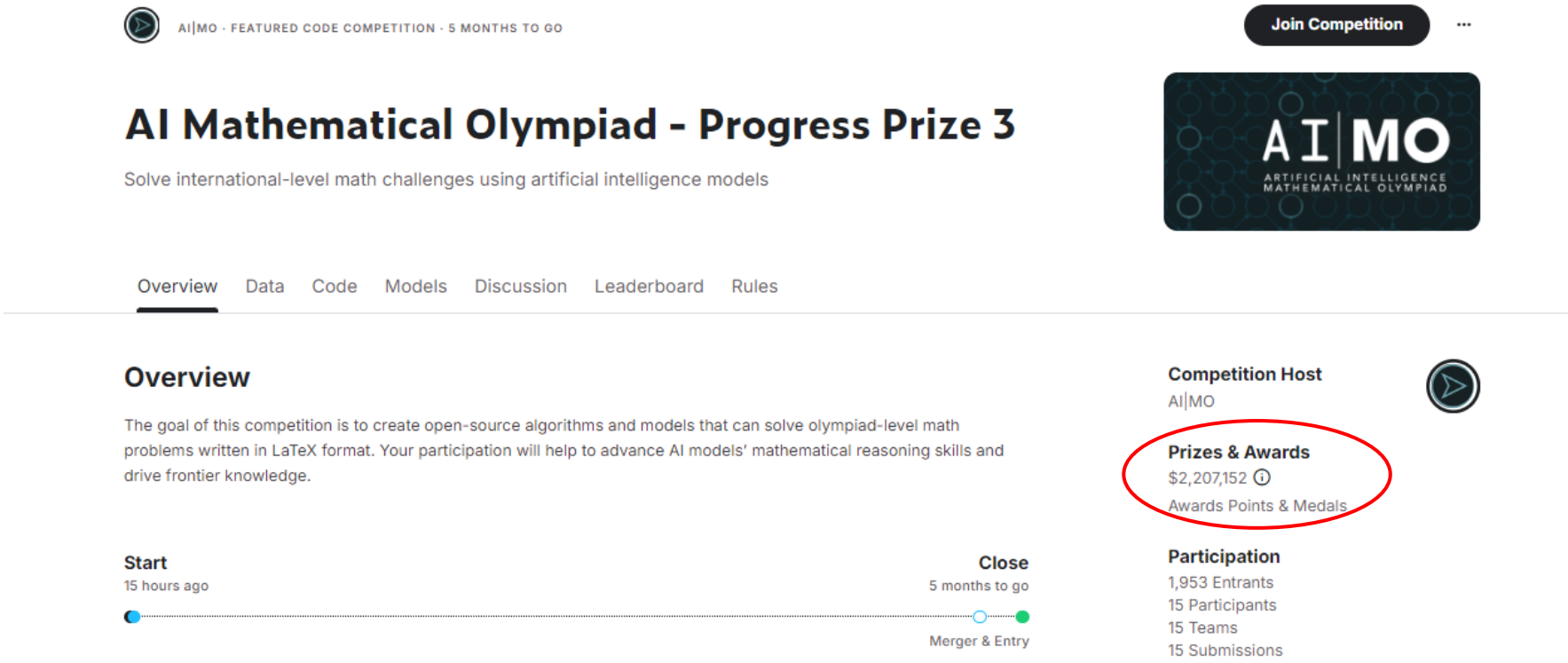
Hope it could finally pass Turing test in a video interaction !

Final projects

- **Ad Hoc Starbuck time every noon next week.**
  - **Tentative time: 12:30 -13:00 pm.**



AIMO 3 is coming



https://www.kaggle.com/competitions/ai-mathematical-olympiad-progress-prize-3

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/
- https://mllm2024.github.io/CVPR2024/