UCF CENTER FOR RESEARCH IN COMPUTER VISION

Visual-Language Models Introduction Part-III Image-Bind, Language-Bind, LLaVA Lecture-6 CAP6412 Spring 2024

Mubarak Shah

shah@crcv.ucf.edu

UCF Center for Research in Computer Vision

Contents

- CoCa
- PALI
- FLAMINGO
- FLAVA
- Painter
- BLIP-2
- Image-Bind
- Language-Bind
- LLaVA
- Video ChatGPT

UCF CENTER FOR RESEARCH IN COMPUTER VISION

IMAGEBIND: One Embedding Space To Bind Them All

Rohit Girdhar* Alaaeldin El-Nouby* Zhuang Liu Mannat Singh
Kalyan Vasudev Alwala Armand Joulin Ishan Misra*
FAIR, Meta AI
https://facebookresearch.github.io/ImageBind

arXiv:2305.05665v2 [cs.CV] 31 May 2023

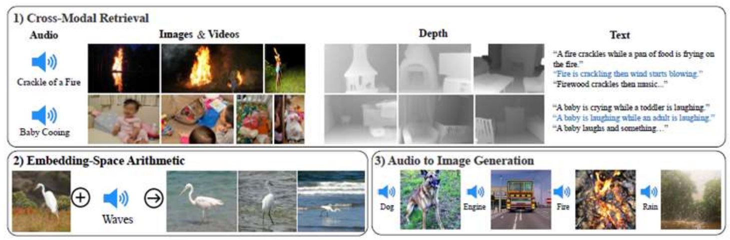


Figure 1. IMAGEBIND's joint embedding space enables novel multimodal capabilities. By aligning six modalities' embedding into a common space, IMAGEBIND enables: 1) Cross-Modal Retrieval, which shows *emergent* alignment of modalities such as audio, depth or text, that aren't observed together. 2) Adding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

Abstract

We present IMAGEBIND, *an approach to learn a joint embedding across six different modalities - images, text, audio, depth, thermal, and IMU data. We show that all combinations of paired data are not necessary to train such a joint embedding, and only image-paired data is sufficient to bind the modalities together.* IMAGEBIND *can leverage recent large scale vision-language models, and extends their zero-shot capabilities to new modalities just by using their natural pairing with images. It enables novel emergent applications 'out-of-the-box' including cross-modal retrieval, composing modalities with arithmetic, cross-modal detection and generation. The emergent capabilities improve with the strength of the image encoder and we set a new state-of-the-art on emergent zero-shot recognition tasks across modalities, outperforming specialist supervised models. Finally, we show strong few-shot recognition results outperforming prior work, and that* IMAGEBIND *serves as a new way to evaluate vision models for visual and non-visual tasks.*

*Equal technical contribution.

1. Introduction

A single image can bind together many experiences – an image of a beach can remind us of the sound of waves, the texture of the sand, a breeze, or even inspire a poem. This 'binding' property of images offers many sources of supervision to learn visual features, by aligning them with any of the sensory experiences associated with images. Ideally, for a single joint embedding space, visual features should be learned by aligning to all of these sensors. However, this requires acquiring all types and combinations of paired data with the same set of images, which is infeasible.

Recently, many methods learn image features aligned with text [1, 31, 46, 60, 64, 65, 82, 83], audio [3, 4, 50, 55, 56, 70] *etc.* These methods use a single pair of modalities or, at best, a few visual modalities. However, the final embeddings are limited to the pairs of modalities used for training. Thus, video-audio embeddings cannot directly be used for image-text tasks and vice versa. A major obstacle in learning a true joint embedding is the absence of large quantities of multimodal data where all modalities are present together.

Contrastive Learning Objective - similar (image, text) pair

Input Image

$$\vec{H}_i$$

Image Representation

$$maximize(\frac{\vec{H}_i \cdot \vec{H}_t}{\|\vec{H}_i\| \times \|\vec{H}_t\|})$$

A dog lying in grass

Input Text

$$\vec{H}_t$$

Text Representation

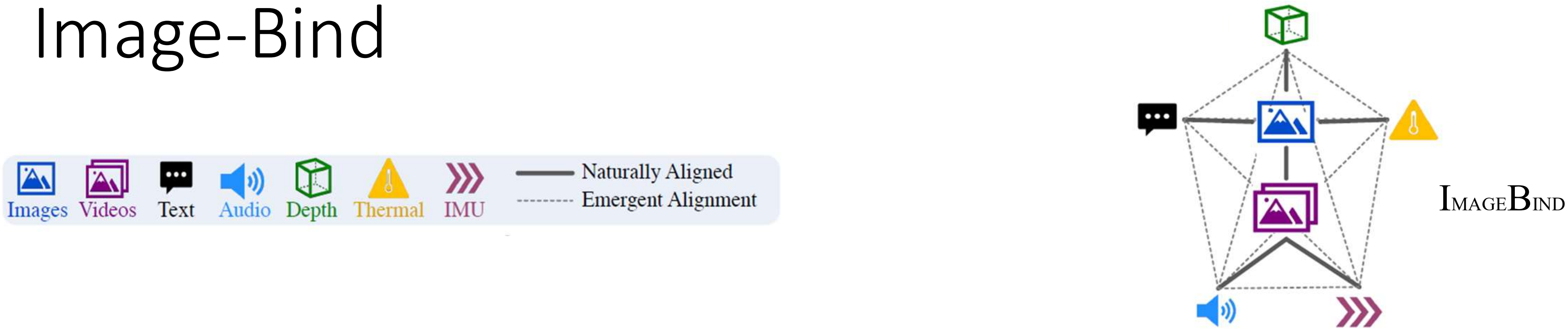
UCF CENTER FOR RESEARCH IN COMPUTER VISION

Image-Bind

- Learn a joint embedding across six different modalities
  - Image
  - Text
  - Audio
  - Depth
  - Thermal
  - IMU
- Only image-paired data is sufficient to bind the modalities together

CENTER FOR RESEARCH IN COMPUTER VISION

Image-Bind



Naturally Paired Datasets

- Video, audio pairs from the Audioset dataset
- Image, depth pairs from the SUN RGB-D dataset
- Image, thermal pairs from the LLVIP dataset
- Video IMU pairs from the Ego4D dataset

UCF Center for Research in Computer Vision

Image-text Supervision

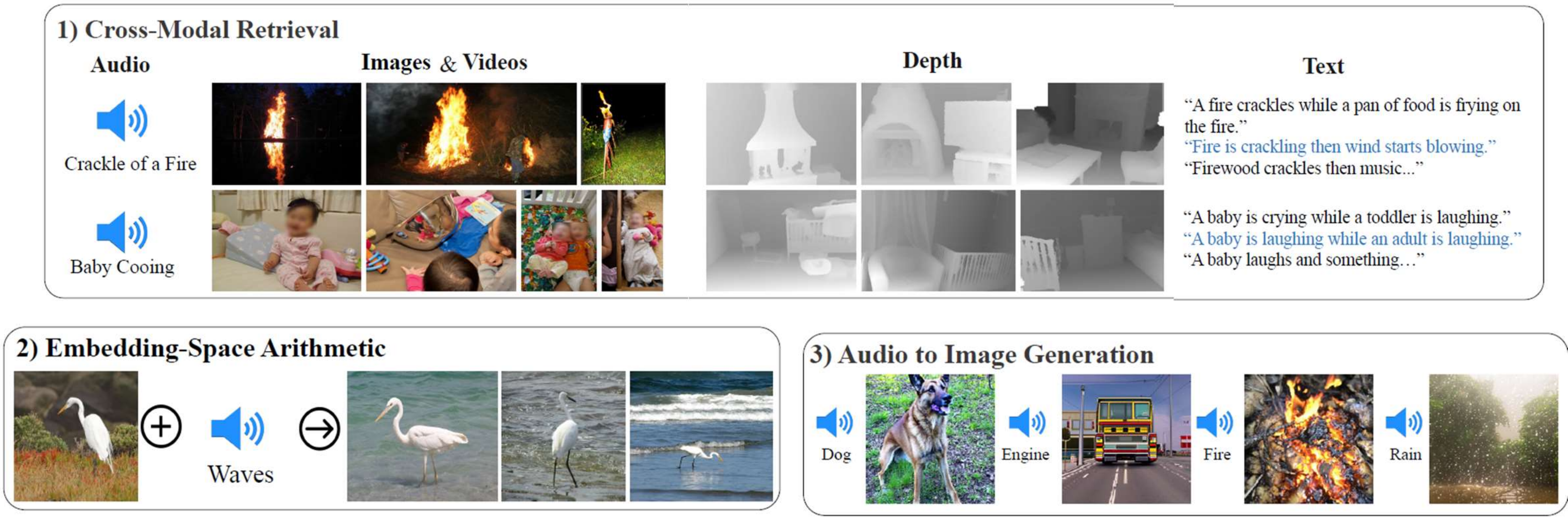
- OpenCLIP Models that are trained on billions of (image, text) pairs
  - ViT-H 630M params Image-Encoder
  - 302M params Text Encoder

UCF CENTER FOR RESEARCH IN COMPUTER VISION

Emerging Zero-Shot Classification Datasets

Dataset	Task	#cls	Metric	#test
Audioset Audio-only (AS-A) [19]	Audio cls.	527	mAP	19048
ESC 5-folds (ESC) [59]	Audio cls.	50	Acc	400
Clotho (Clotho) [17]	Retrieval	-	Recall	1045
AudioCaps (AudioCaps) [37]	Retrieval	-	Recall	796
VGGSound (VGGS) [8]	Audio cls.	309	Acc	14073
SUN Depth-only (SUN-D) [69]	Scene cls.	19	Acc	4660
NYU-v2 Depth-only (NYU-D) [66]	Scene cls.	10	Acc	653
LLVIP (LLVIP) [32]	Person cls.	2	Acc	15809
Ego4D (Ego4D) [23]	Scenario cls.	108	Acc	68865

Image Bind


1) Cross-Modal Retrieval
Audio
Images & Videos
Depth
Text
Crackle of a Fire
"A fire crackles while a pan of food is frying on the fire."
"Fire is crackling then wind starts blowing."
"Firewood crackles then music..."
Baby Cooing
"A baby is crying while a toddler is laughing."
"A baby is laughing while an adult is laughing."
"A baby laughs and something…"
2) Embedding-Space Arithmetic
Waves
3) Audio to Image Generation
Dog
Engine
Fire
Rain

LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment

Bin Zhu[1,2,*], Bin Lin[1,*], Munan Ning[1,4], Yang Yan[1], JiaXi Cui[1], Hongfa Wang[2], Yatian Pang[3], Wenhao Jiang[6], Junwu Zhang[1], Zongwei Li[2], Wancai Zhang[5], Zhifeng Li[2], Wei Liu[2], Li Yuan[1,4,†]
[1]Peking University, [2]Tencent Data Platform, [3]National University of Singapore, [4]Pengcheng Lab, [5]Nari Technology Development Limited Company, [6]Guangming Laboratory

arXiv:2310.01852v6 [cs.CV] 27 Nov 2023

Abstract

The video-language (VL) pretraining has achieved remarkable improvement in multiple downstream tasks. However, the current VL pretraining framework is hard to extend to multiple modalities (N modalities, $N \geq 3$) beyond vision and language. We thus propose ***LanguageBind***, taking the language as the bind across different modalities because the language modality is well-explored and contains rich semantics. Specifically, we freeze the language encoder acquired by VL pretraining and then train encoders for other modalities with contrastive learning. As a result, all modalities are mapped to a shared feature space, implementing multi-modal semantic alignment. While LanguageBind ensures that we can extend VL modalities to N modalities, we also need a high-quality dataset with alignment data pairs centered on language. We thus propose ***VIDAL-10M*** with 10 ***M***illion data with ***V***ideo, ***I***nfrared, ***D***epth, ***A***udio and their corresponding ***L***anguage. In our VIDAL-10M, all videos are from short video platforms with complete semantics rather than truncated segments from long videos, and all the video, depth, infrared, and audio modalities are aligned to their textual descriptions. After pretraining on our dataset, we outperform CLIP4Clip by **3.7%** R@1 on the MSR-VTT dataset with only **26% of data** in the zero-shot video-text retrieval. Beyond this, our LanguageBind achieves superior performances on a broad range of **15 benchmarks** across video, audio, depth and infrared. Code address: https://github.com/PKU-YuanGroup/LanguageBind

1 Introduction

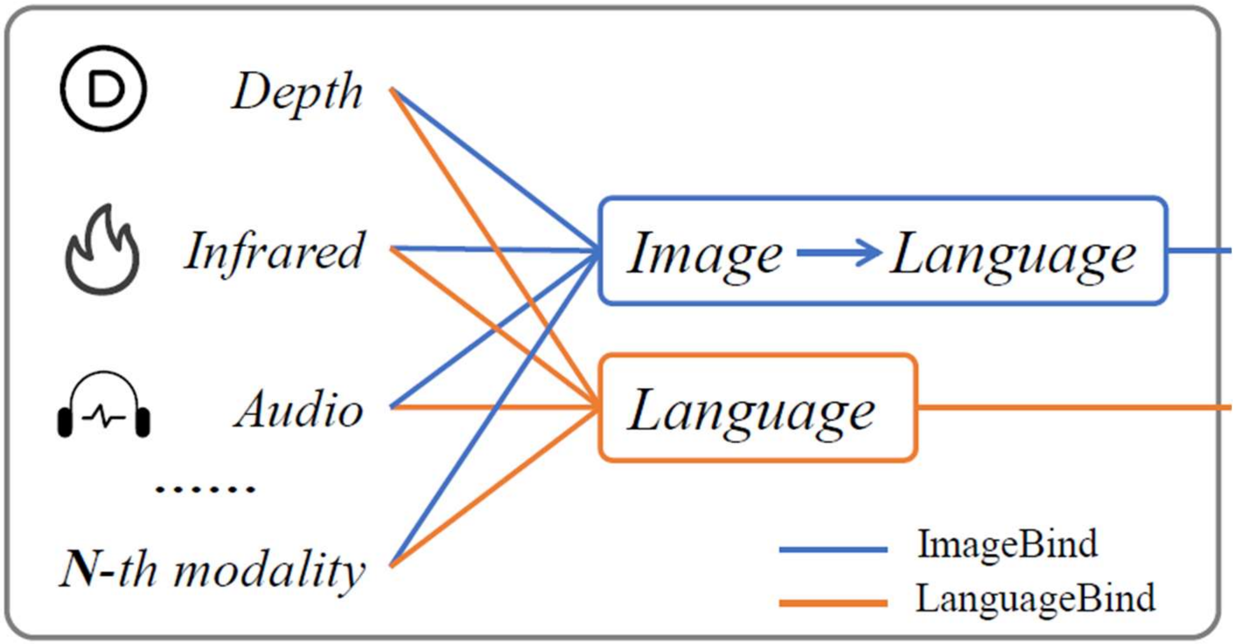
With the development of the Internet and smartphones, there has been a proliferation of video websites and apps (*e.g.*, Youtube and TikTok), leading to a substantial increase number of videos (Xue et al., 2022). Consequently, a set of video tasks have emerged, such as video search (Smith & Chang, 1997), video recommendation (Deldjoo et al., 2016), and video editing (Casares et al., 2002; Bonneel et al., 2014). To solve video understanding tasks, video-language pretraining has been employed by training foundation models by combining computer vision (He et al., 2016; Dosovitskiy et al., 2020) and natural language processing (Vaswani et al., 2017). These models can capture video semantics and solve downstream tasks (Karpathy et al., 2014; Mithun et al., 2018).

However, current VL pretraining frameworks are often limited to vision and language modalities. The ImageBind (Girdhar et al., 2023) introduces an indirect alignment method for multi-modal pretraining. It aligns other modalities to images, facilitating a comprehensive understanding of various modalities such as infrared (Jia et al., 2021), depth (Kim et al., 2022), audio (Piczak, 2015), and IMU (Grauman et al., 2022). In practical tasks such as zero-shot retrieval and classification as shown in Figure 1, the alignment with language modality is predominantly required for various modalities. While the indirect alignment of ImageBind may result in performance degradation,

*Equal contribution.
†Corresponding author.

UCF Center for Research in Computer Vision

Image-Bind &Language-Bind

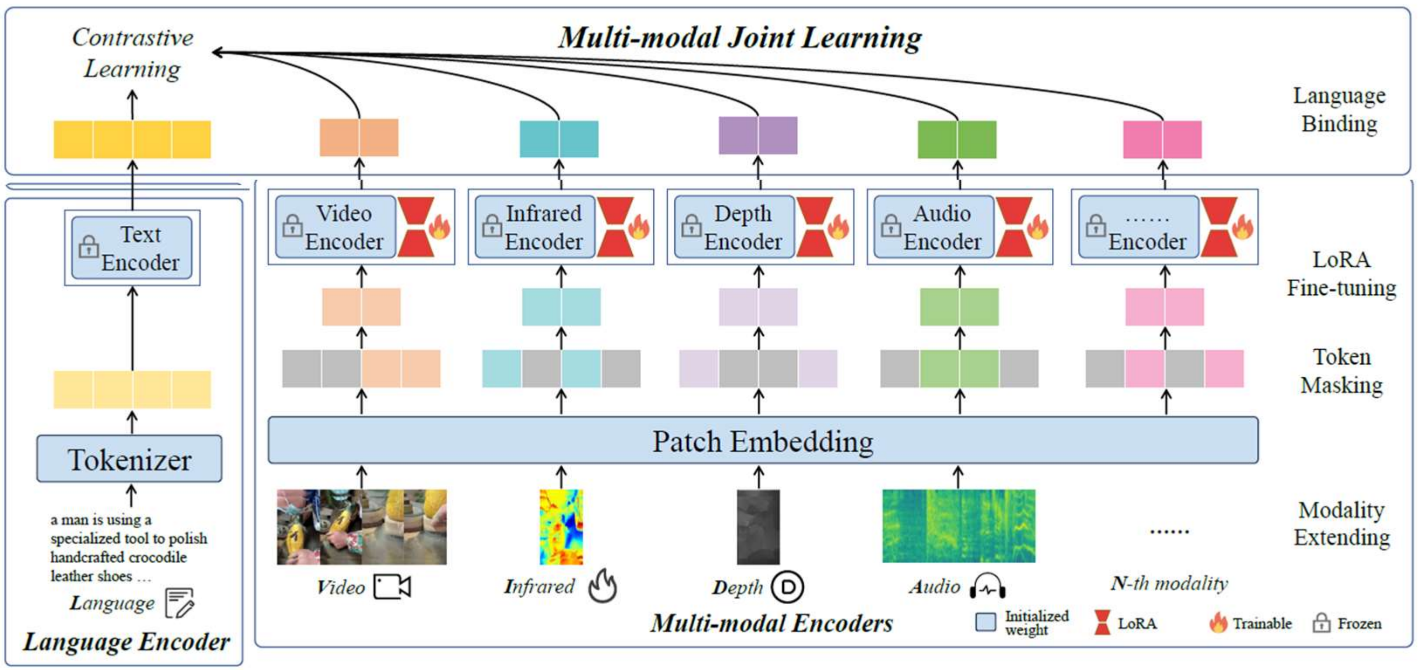


Comparison of VIDAL-10M Dataset with other Datasets

Datasets	**Samples**	**Modality**	**Year**
HMDB-51	7K	V	2011
UCF-101	13K	V	2012
ActivityNet-200	20K	VT	2015
WebVid-10M	10.7M	VT	2021
HD-VILA-100M	100M	VT	2022
HowTo-100M	136M	VT	2019
LLVIP	15k	VI	2021
FLIR V1	10k	VI	2015
FLIR V2	12k	VI	2015
NYU-D	1.4k	VD	2012
YouTube-8M	6.1M	VAT	2016
AVA	58K	VAT	2017
VIDAL-10M (Ours)	10M	VIDAL	2023

UCF CENTER FOR RESEARCH IN COMPUTER VISION

Language-Bind

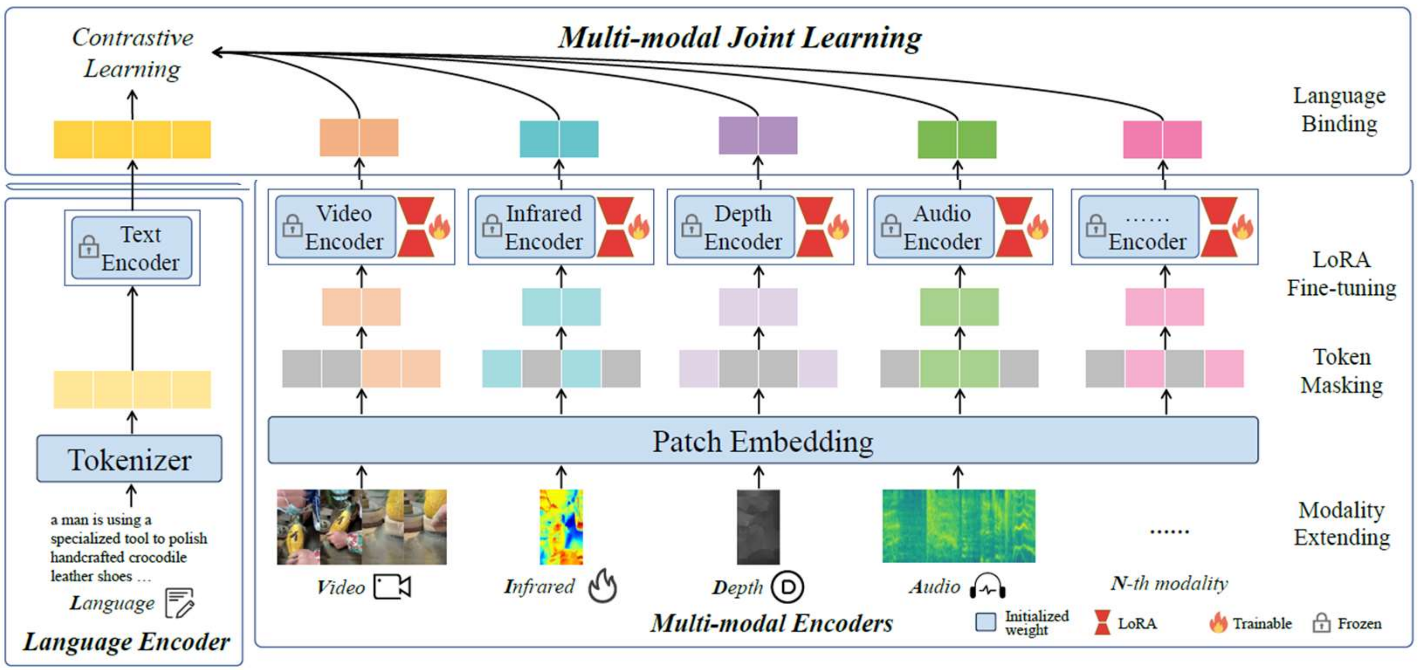


Multi-Modal Encoders

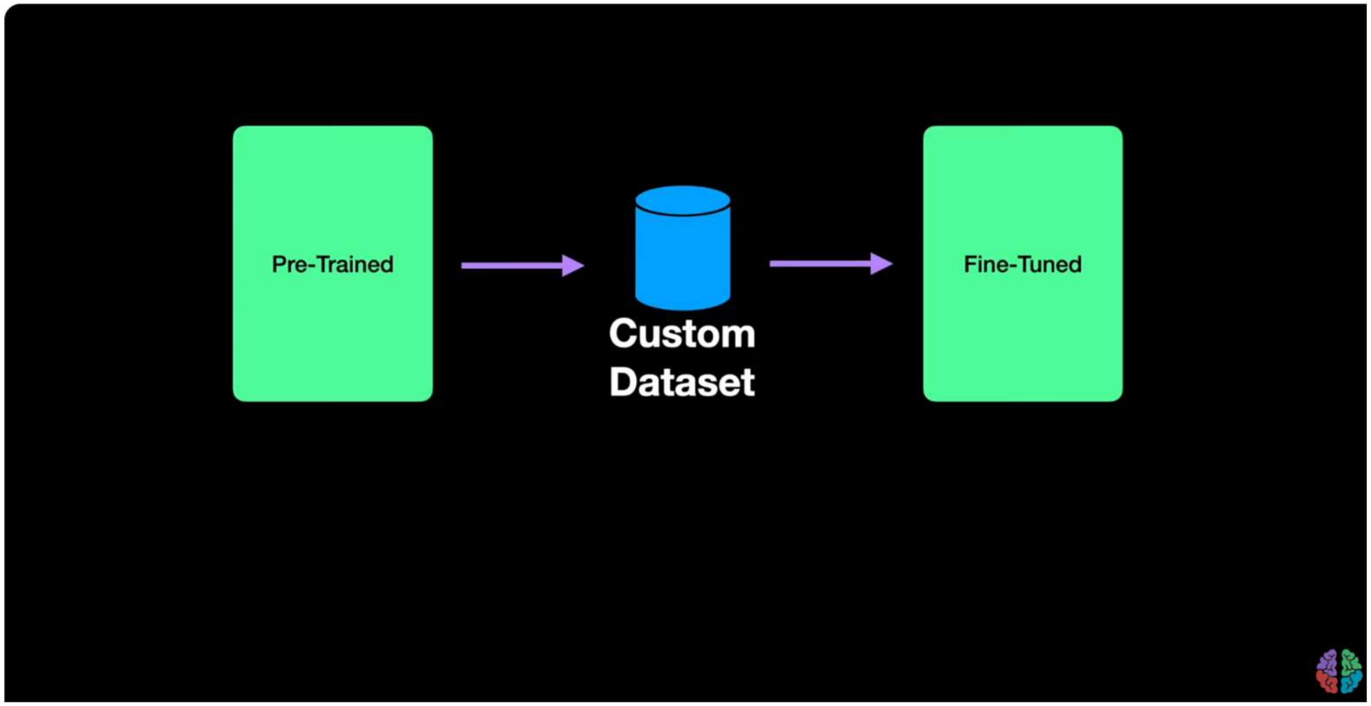
- 24-layer, 1024-dimensional VIT with a patch size of 14
- Initialize with OpenCLIP-large
- Low Rank Adpatation (LORA)

UCF Center for Research in Computer Vision

Language-Bind

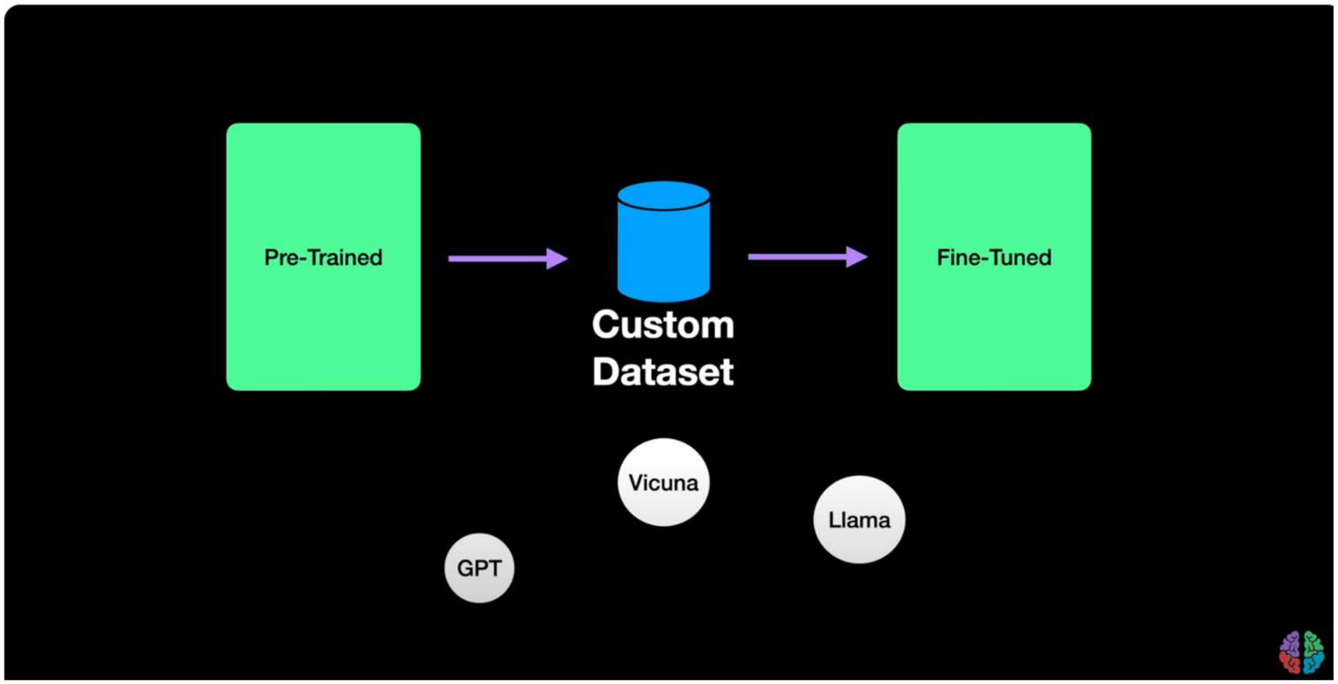


Finetuning



https://www.youtube.com/watch?v=X4VvO3G6_vw

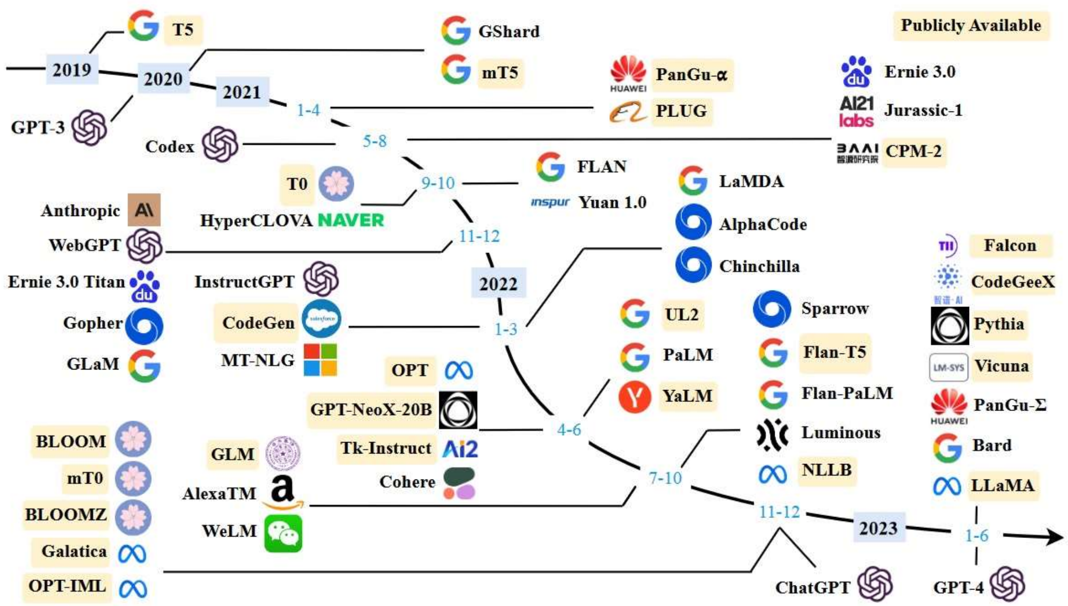
Finetuning



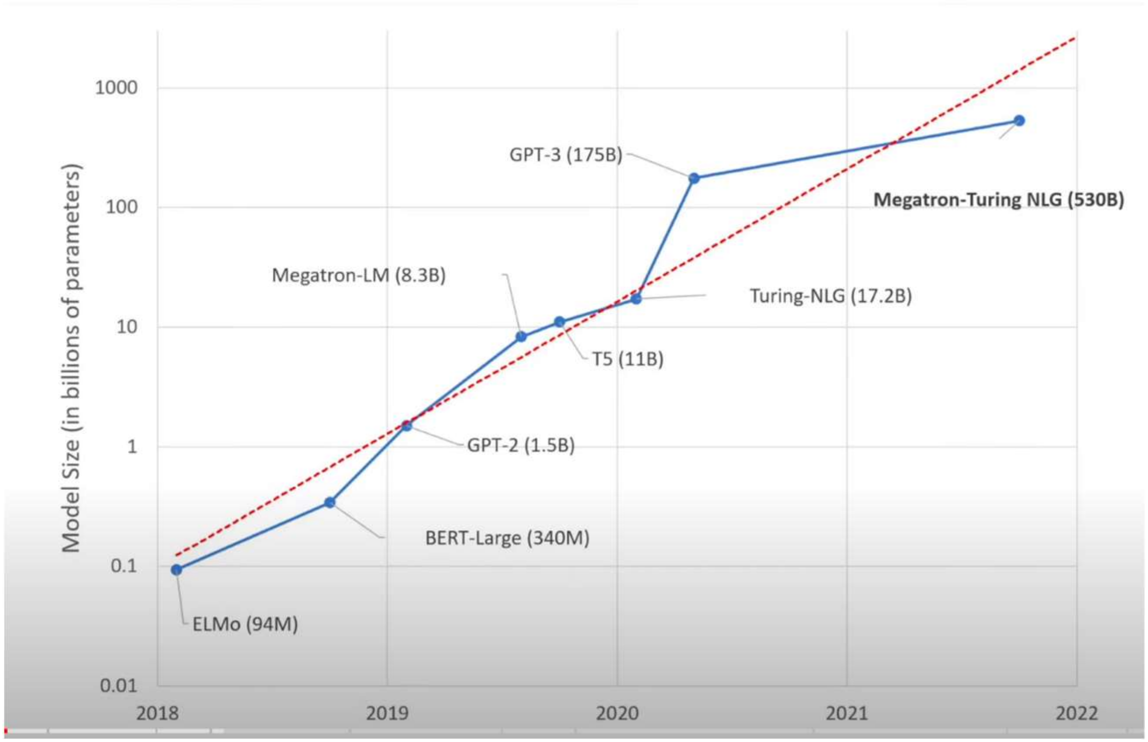
https://www.youtube.com/watch?v=X4VvO3G6_vw

UCF CENTER FOR RESEARCH IN COMPUTER VISION

Recent Advances in Large Language Models (LLMs)

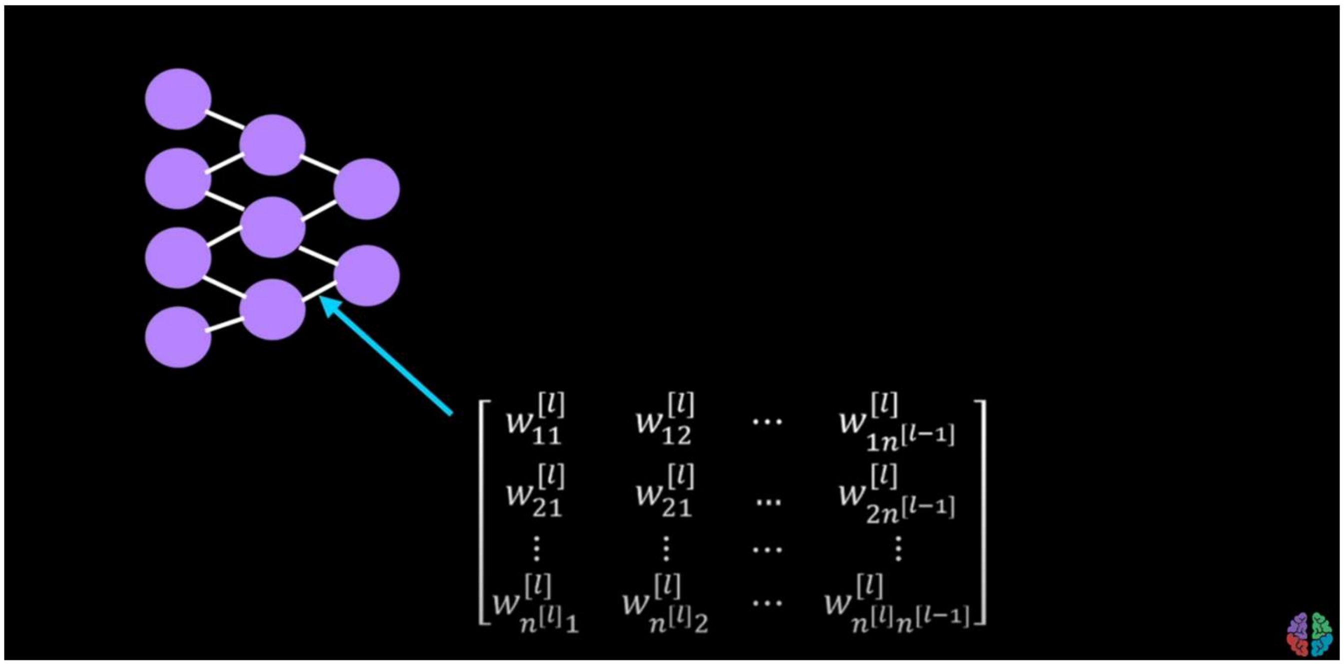


Zhao et al., A Survey of Large Language Models, Arxiv 2023.



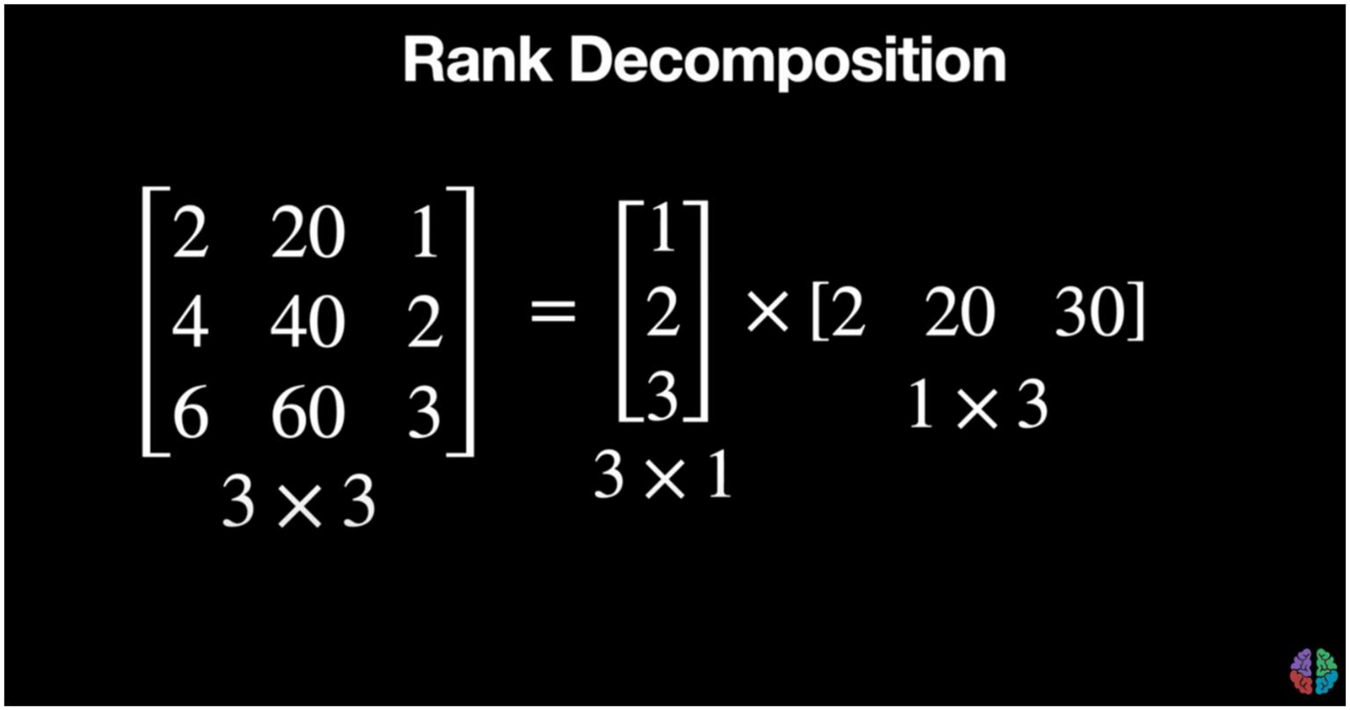
https://www.youtube.com/watch?v=X4VvO3G6_vw

Weight Matrix

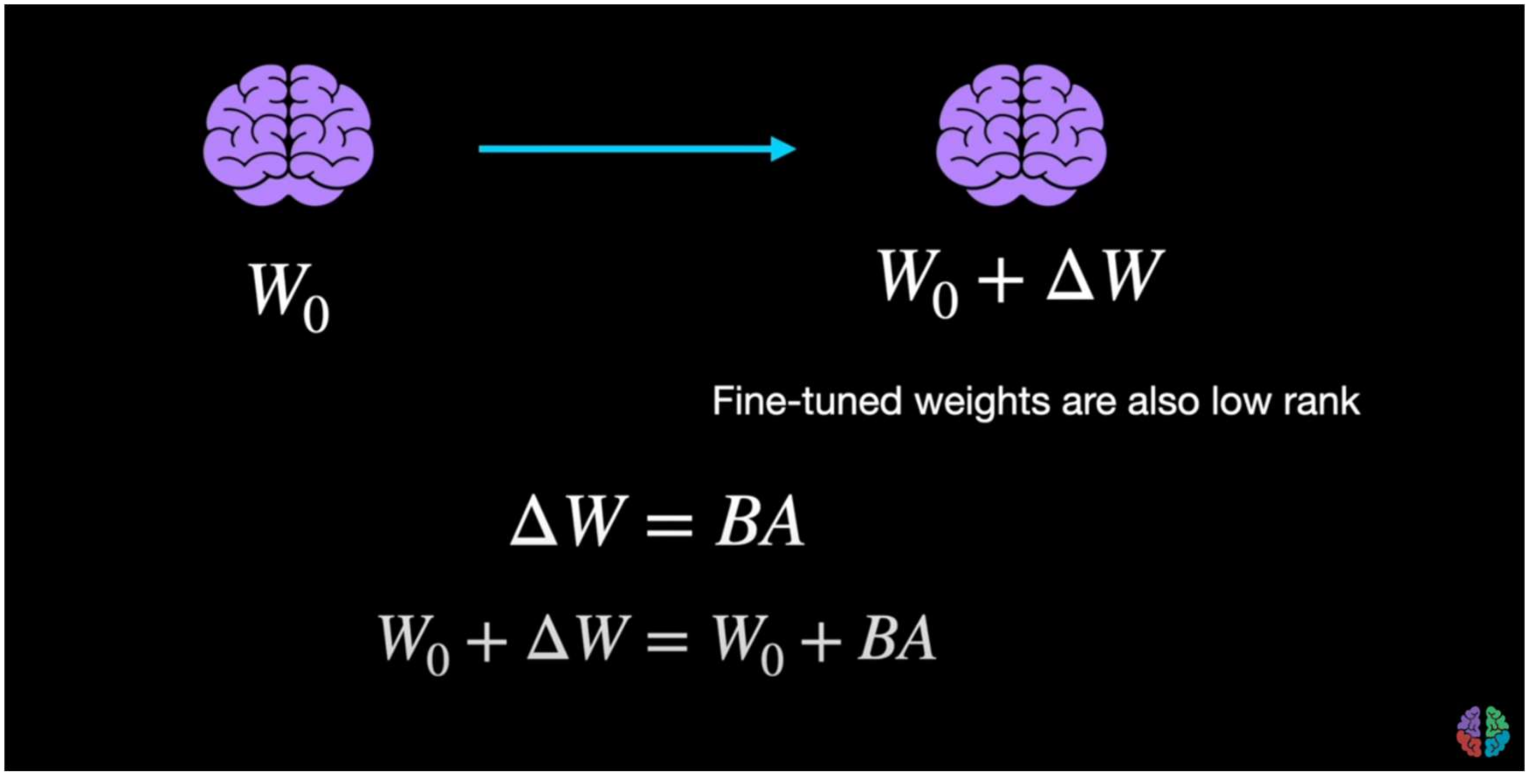


https://www.youtube.com/watch?v=X4VvO3G6_vw

UCF Center for Research in Computer Vision



https://www.youtube.com/watch?v=X4VvO3G6_vw



https://www.youtube.com/watch?v=X4VvO3G6_vw

LORA: Low Rank Adaptation

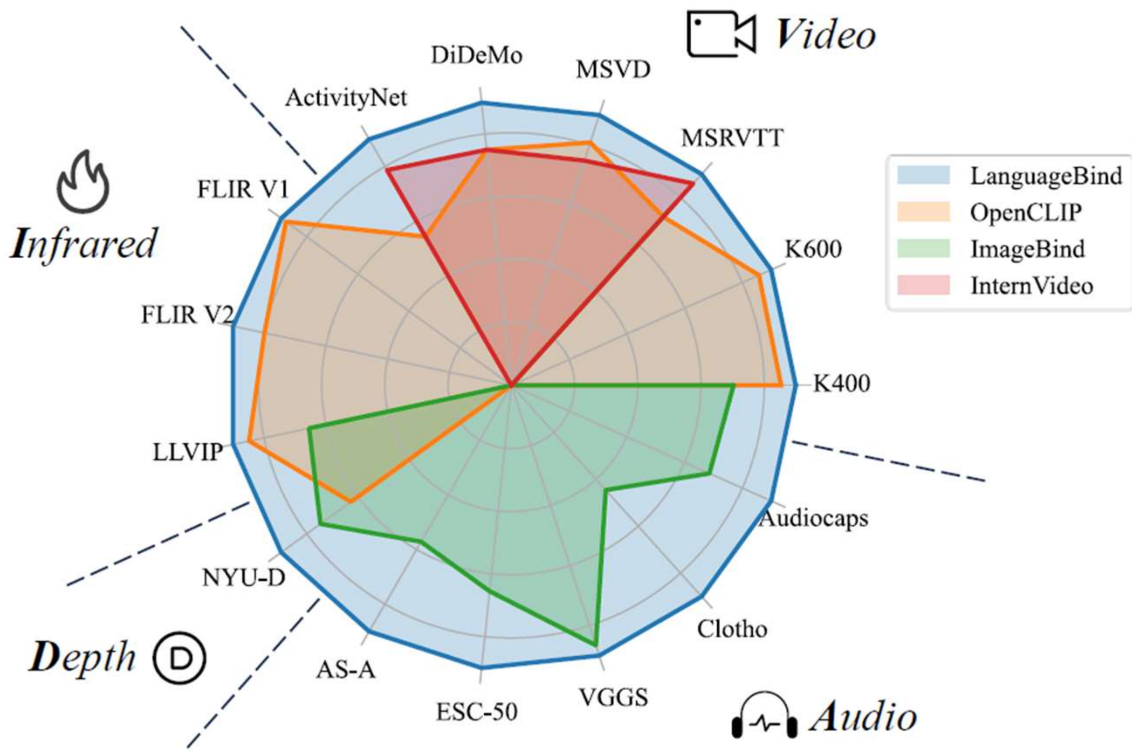
$$W_0 \in \mathbb{R}^{d \times k}$$

$$h(\boldsymbol{x}) = W_0\boldsymbol{x} + BA\boldsymbol{x}$$

$$B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k},$$

UCF Center for Research in Computer Vision

Results



UCF CENTER FOR RESEARCH IN COMPUTER VISION

Visual Instruction Tuning

Haotian Liu[1]*, Chunyuan Li[2]*, Qingyang Wu[3], Yong Jae Lee[1]
[1]University of Wisconsin–Madison [2]Microsoft Research [3]Columbia University
https://llava-vl.github.io

arXiv:2304.08485v1 [cs.CV] 17 Apr 2023

Abstract

Instruction tuning large language models (LLMs) using machine-generated instruction-following data has improved zero-shot capabilities on new tasks, but the idea is less explored in the multimodal field. In this paper, we present the first attempt to use language-only GPT-4 to generate multimodal language-image instruction-following data. By instruction tuning on such generated data, we introduce LLaVA: Large Language and Vision Assistant, an end-to-end trained large multimodal model that connects a vision encoder and LLM for general-purpose visual and language understanding. Our early experiments show that LLaVA demonstrates impressive multimodel chat abilities, sometimes exhibiting the behaviors of multimodal GPT-4 on unseen images/instructions, and yields a 85.1% relative score compared with GPT-4 on a synthetic multimodal instruction-following dataset. When fine-tuned on Science QA, the synergy of LLaVA and GPT-4 achieves a new state-of-the-art accuracy of 92.53%. We make GPT-4 generated visual instruction tuning data, our model and code base publicly available.

1 Introduction

Humans interact with the world through many channels such as vision and language, as each individual channel has a unique advantage in representing and communicating certain world concepts, and thus facilitates a better understanding of the world. One of the core aspirations in artificial intelligence is to develop a general-purpose assistant that can effectively follow multi-modal vision-and-language instructions, aligned with human intent to complete various real-world tasks in the wild [4, 24].

To this end, the community has witnessed an emergent interest in developing language-augmented foundation vision models [24, 14], with strong capabilities in open-world visual understanding such as classification [36, 18, 53, 50, 35], detection [26, 58, 29], segmentation [23, 59, 54] and captioning [46, 25], as well as visual generation and editing [38, 39, 52, 13, 40, 27]. We refer readers to the *Computer Vision in the Wild* reading list for a more up-to-date literature compilation [11]. In this line of work, each task is solved independently by one single large vision model, with the task instruction implicitly considered in the model design. Further, language is only utilized to describe the image content. While this allows language to play an important role in mapping visual signals to language semantics—a common channel for human communication, it leads to models that usually have a fixed interface with limited interactivity and adaptability to the user's instructions.

Large language models (LLM), on the other hand, have shown that language can play a wider role: a universal interface for a general-purpose assistant, where various task instructions can be explicitly represented in language and guide the end-to-end trained neural assistant to switch to the task of interest to solve it. For example, the recent success of ChatGPT [31] and GPT-4 [32] have demonstrated the power of aligned LLMs in following human instructions, and have stimulated tremendous interest in developing open-source LLMs. Among them, LLaMA [44] is an open-source LLM that matches the performance of GPT-3. Alpaca [43], Vicuna [45], GPT-4-LLM [34]

*Equal contribution

Preprint. Work in progress

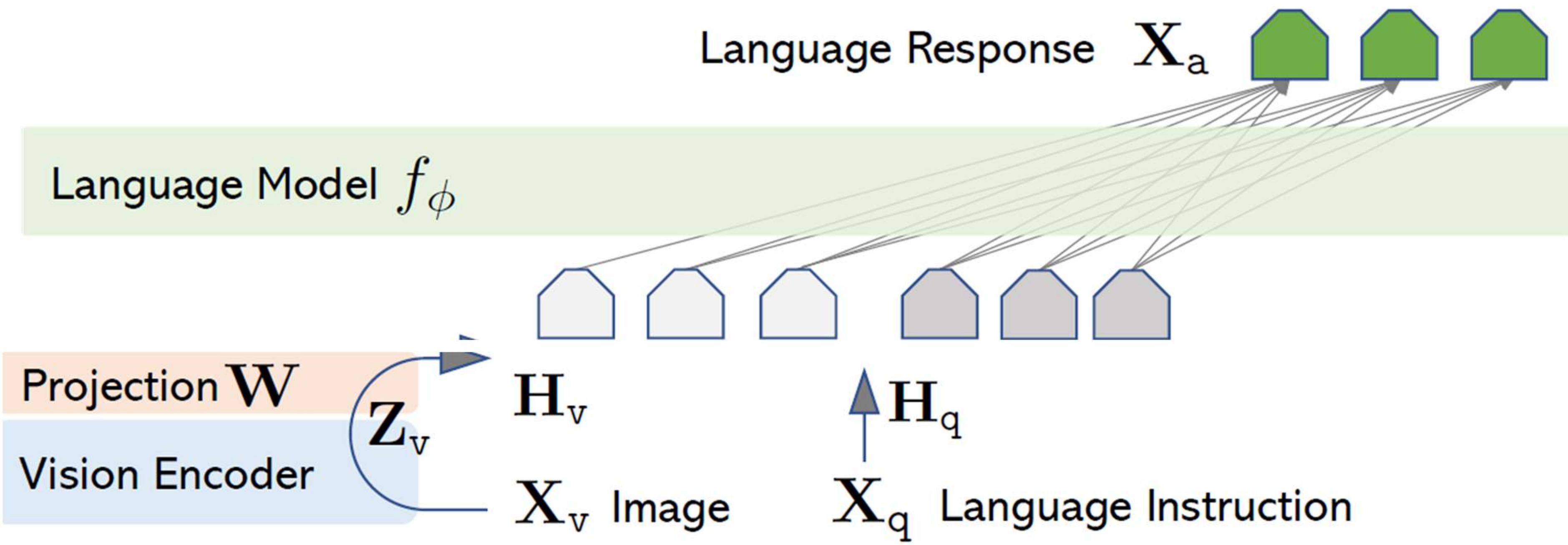
660 citations

UCF CENTER FOR RESEARCH IN COMPUTER VISION

LLaVA: **L**arge **L**anguage and **V**ision **A**ssistant

- First successful Image-text conversation model
- Instruction Tunning
  - Instruction, Question, Answer
- Instruction Following Data
  - Convert Image-Text pairs to instruction Data
- Achieves 92% accuracy on Science QA
  - 21k multimodal multiple-choice questions
  - Across 3 subjects,
  - 26 topics,
  - 127 categories, and
  - 379 skills

LLaVA Architecture



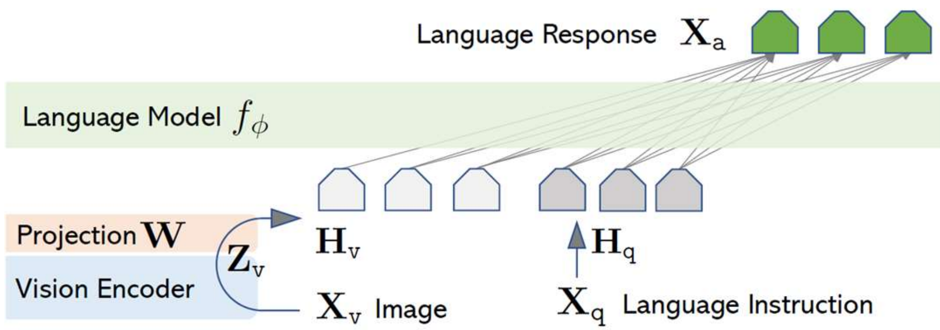
$$\mathbf{H}_v = \mathbf{W} \cdot \mathbf{Z}_v, \text{ with } \mathbf{Z}_v = g(\mathbf{X}_v)$$

UCF Center for Research in Computer Vision

Training

Stage 1: Pre-training for Feature Alignment.

- Filter CC3M to 595K image-text pairs
- Keep both the visual encoder and LLM weights frozen
- Train for $\boldsymbol{\theta} = \mathbf{W}$ (the projection matrix)



Stage 2: Fine-tuning End-to-End

$$\boldsymbol{\theta} = \{\mathbf{W}, \phi\}$$

$$\mathbf{H}_\mathrm{v} = \mathbf{W} \cdot \mathbf{Z}_\mathrm{v}, \text{ with } \mathbf{Z}_\mathrm{v} = g(\mathbf{X}_\mathrm{v})$$

UCF CENTER FOR RESEARCH IN COMPUTER VISION

Instruction Tunning

$\mathbf{X}_{\texttt{system-message}}$ <STOP> \n
Human : $\mathbf{X}^1_{\texttt{instruct}}$ <STOP> \n Assistant: $\mathbf{X}^1_{\texttt{a}}$ <STOP> \n
Human : $\mathbf{X}^2_{\texttt{instruct}}$ <STOP> \n Assistant: $\mathbf{X}^2_{\texttt{a}}$ <STOP> \n $\cdots$

$X_{\text{system-message}}$ = A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions. and <STOP> = ###.

For each image $\mathbf{X}_{\text{v}}$

$$(\mathbf{X}^1_{\texttt{q}}, \mathbf{X}^1_{\texttt{a}}, \cdots, \mathbf{X}^T_{\texttt{q}}, \mathbf{X}^T_{\texttt{a}}). \quad \mathbf{X}^t_{\texttt{instruct}} = \begin{cases} \text{Random choose } [\mathbf{X}^1_{\texttt{q}}, \mathbf{X}_{\texttt{v}}] \text{ or } [\mathbf{X}_{\texttt{v}}, \mathbf{X}^1_{\texttt{q}}], & \text{the first turn } t = 1 \\ \mathbf{X}^t_{\texttt{q}}, & \text{the remaining turns } t > 1 \end{cases}$$

$$p(\mathbf{X}_{\texttt{a}} | \mathbf{X}_{\texttt{v}}, \mathbf{X}_{\texttt{instruct}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i | \mathbf{X}_{\texttt{v}}, \mathbf{X}_{\texttt{instruct},<i}, \mathbf{X}_{\texttt{a},<i})$$

UCF CENTER FOR RESEARCH IN COMPUTER VISION

LLaVA

- Use COCO images to generate three types of instruction-following data
  - Conversation
  - Detailed Description
  - Complex Reasoning

- Captions and boxes used to prompt GPT to generate Q&A pairs
- 158K unique language-image instruction-following samples
  - 58K in conversations
  - 23K in detailed description
  - 77k in complex reasoning

UCF CENTER FOR RESEARCH IN COMPUTER VISION

Instruction-following data

- Captions and boxes are used to prompt GPT
- No visual image is not used to prompt GPT

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]

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

Instructions for brief image description

- "Describe the image concisely."
- "Provide a brief description of the given image."
- "Offer a succinct explanation of the picture presented."
- "Summarize the visual content of the image."
- "Give a short and clear explanation of the subsequent image."
- "Share a concise interpretation of the image provided."
- "Present a compact description of the photo’s key features."
- "Relay a brief, clear account of the picture shown."
- "Render a clear and concise summary of the photo."
- "Write a terse but informative summary of the picture."
- "Create a compact narrative representing the image presented."

The list of instructions for detailed image description

- "Describe the following image in detail"
- "Provide a detailed description of the given image"
- "Give an elaborate explanation of the image you see"
- "Share a comprehensive rundown of the presented image"
- "Offer a thorough analysis of the image"
- "Explain the various aspects of the image before you"
- "Clarify the contents of the displayed image with great detail"
- "Characterize the image using a well-detailed description"
- "Break down the elements of the image in a detailed manner"
- "Walk through the important details of the image"
- "Portray the image with a rich, descriptive narrative"
- "Narrate the contents of the image with precision"
- "Analyze the image in a comprehensive and detailed manner"
- "Illustrate the image through a descriptive explanation"
- "Examine the image closely and share its details"
- "Write an exhaustive depiction of the given image"

Few-Shot In context Prompt

```
messages = [ {"role":"system", "content": f"""You are an AI visual assistant, and you are seeing a single image. What you see are provided with five sentences, describing the same image you are looking at. Answer all questions as you are seeing the image.

Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the image, including the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Only include questions that have definite answers:
(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently.

Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary."""}
]
for sample in fewshot_samples:
    messages.append({"role":"user", "content":sample['context']})
    messages.append({"role":"assistant", "content":sample['response']} )
messages.append({"role":"user", "content":'\n'.join(query)})
```

UCF CENTER FOR RESEARCH IN COMPUTER VISION

In-context Learning

Captions
There is a movie theater that displays the show times above the doors.
A red fire hydrant is deep in the snow.
The fire hydrant is in the snow near a recently plowed sidewalk.
This city has had a very hard winter with snow.
A hotel for dogs in the snow in winter.

Conversation
Question:
What color is the fire hydrant in the image?
===
Answer:
The fire hydrant in the image is red.
===
Question:
Is the fire hydrant fully covered by snow or partially visible?
===
Answer:
The fire hydrant is partially visible, with a significant portion of it covered by the deep snow.
===
Question:
What is the condition of the sidewalk near the fire hydrant?
===
Answer:
The sidewalk near the fire hydrant has been recently plowed, so it is relatively clear of snow compared to the surrounding area.
===
Question:
What does the image suggest about the winter season in the city?
===
Answer:
The image suggests that the winter season in the city has been quite harsh and challenging. The significant snow accumulation seen in the image indicates that the city has experienced heavy snowfall, which can lead to various difficulties and disruptions for its residents.

For instance, the deep snow covering the fire hydrant could potentially hinder its accessibility in case of emergencies, which poses safety concerns. The recently plowed sidewalk implies that the city's maintenance crews have been actively working to keep the walkways clear and safe for pedestrians, but the sheer amount of snowfall might make it difficult to keep up with the continuous snow removal efforts.

Furthermore, such extreme winter conditions can impact transportation, leading to traffic issues, delays in public transit, and increased risks of accidents due to slippery roads. It can also cause problems for businesses and schools, as people might face challenges commuting to work or attending classes. Additionally, the heavy snow can put extra strain on infrastructure, such as roofs and power lines, increasing the likelihood of structural damage or power outages.

In conclusion, the image of the red fire hydrant deep in the snow and the recently plowed sidewalk suggest that the city has faced a particularly severe winter season, with substantial snowfall that has likely caused various challenges and disruptions for its residents and infrastructure.

Visual input example, Extreme Ironing:

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

User: What is unusual about this image?

LLaVA: The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

[*Start a new conversation, and clear the history*]

User: What's happening in the scene?

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

Visual input example, Chicken Nugget Map:

User: Can you explain this meme in detail?

LLaVA: The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.

UCF CENTER FOR RESEARCH IN COMPUTER VISION

What's the ending of this movie?

In the image, a young woman is leaning on a man in front of a beautiful sunset, with both of them standing on a boat. They appear to be enjoying a romantic moment together. The scene suggests that they could be depicted as the characters Jack and Rose from the popular movie Titanic. In the movie, their romance blossoms aboard the fictional ocean liner Titanic, with the pair ultimately sharing a passionate kiss as the ship sinks into the ocean. The ending of the movie is known for its emotional and heartfelt scenes, where the characters are forced to say goodbye to each other as they are separated by the sinking ship. The image evokes a sense of romance and nostalgia, paying tribute to the iconic love story portrayed in the movie Titanic.