UCF Center for Research in Computer Vision

Visual-Language Models Introduction
Part-II
Image-Bind, Language-Bind, LLaVA
Lecture-6
CAP6412 Spring 2025

Mubarak Shah

shah@crcv.ucf.edu

UCF Center for Research in Computer Vision

Contents

- Lecture-4
  - CLIP
  - CoCa
  - PALI
- Lecture-5
  - FLAMINGO
  - FLAVA
  - Painter
  - BLIP-2
- Lecture-6
  - Image-Bind
  - Language-Bind
  - LLaVA
- Lecture-7
  - Video ChatGPT
  - Video LLaVA
  - PG-Video LLaVA

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

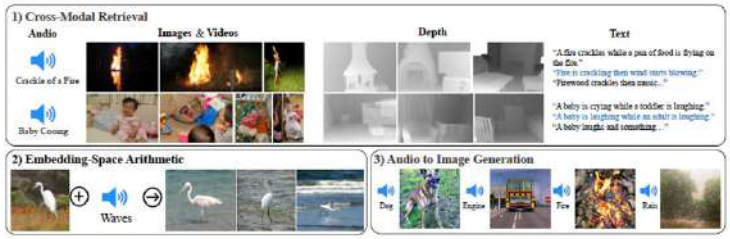


Figure 1. IMAGEBIND's joint embedding space enables novel multimodal capabilities. By aligning six modalities' embedding into a common space, IMAGEBIND enables: 1) Cross-Modal Retrieval, which shows *emergent* alignment of modalities such as audio, depth or text, that aren't observed together. 2) Adding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

A dog lying in grass

Input Text

$\vec{H}_t$

Text Representation

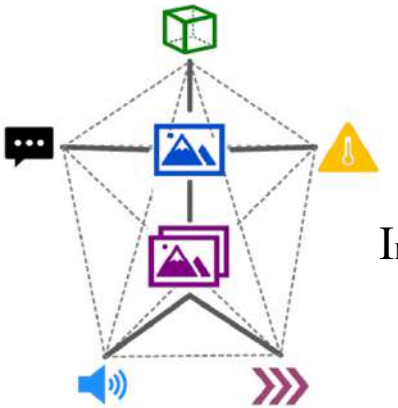
$$maximize(\frac{\vec{H}_i \cdot \vec{H}_t}{\|\vec{H}_i\| \times \|\vec{H}_t\|})$$

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

Image-Bind





ImageBind



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

Image-text Supervision

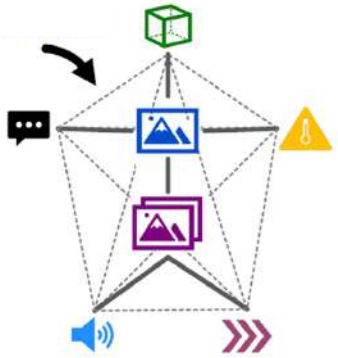
- OpenCLIP Models that are trained on billions of (image, text) pairs
  - ViT-H 630M params Image-Encoder
  - 302M params Text Encoder

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

Emerging Zero-Shot Classification Results

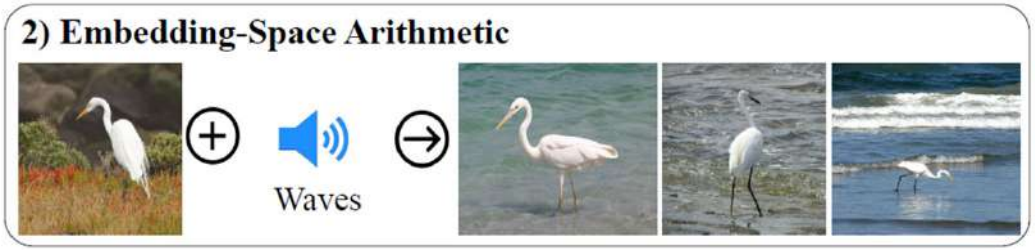
	IN1K	P365	K400	MSR-VTT	NYU-D	SUN-D	AS-A	VGGS	ESC	LLVIP	Ego4D
Random	0.1	0.27	0.25	0.1	10.0	5.26	0.62	0.32	2.75	50.0	0.9
IMAGEBIND	77.7	45.4	50.0	36.1	54.0	35.1	17.6	27.8	66.9	63.4	25.0
Text Paired	-	-	-	-	41.9*	25.4*	28.4† [27]	-	68.6† [27]	-	-
Absolute SOTA	91.0 [82]	60.7 [67]	89.9 [80]	57.7 [79]	76.7 [21]	64.9 [21]	49.6 [39]	52.5 [36]	97.0 [9]	-	-

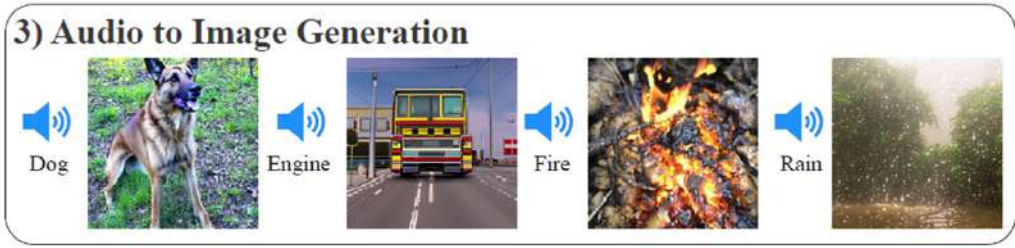


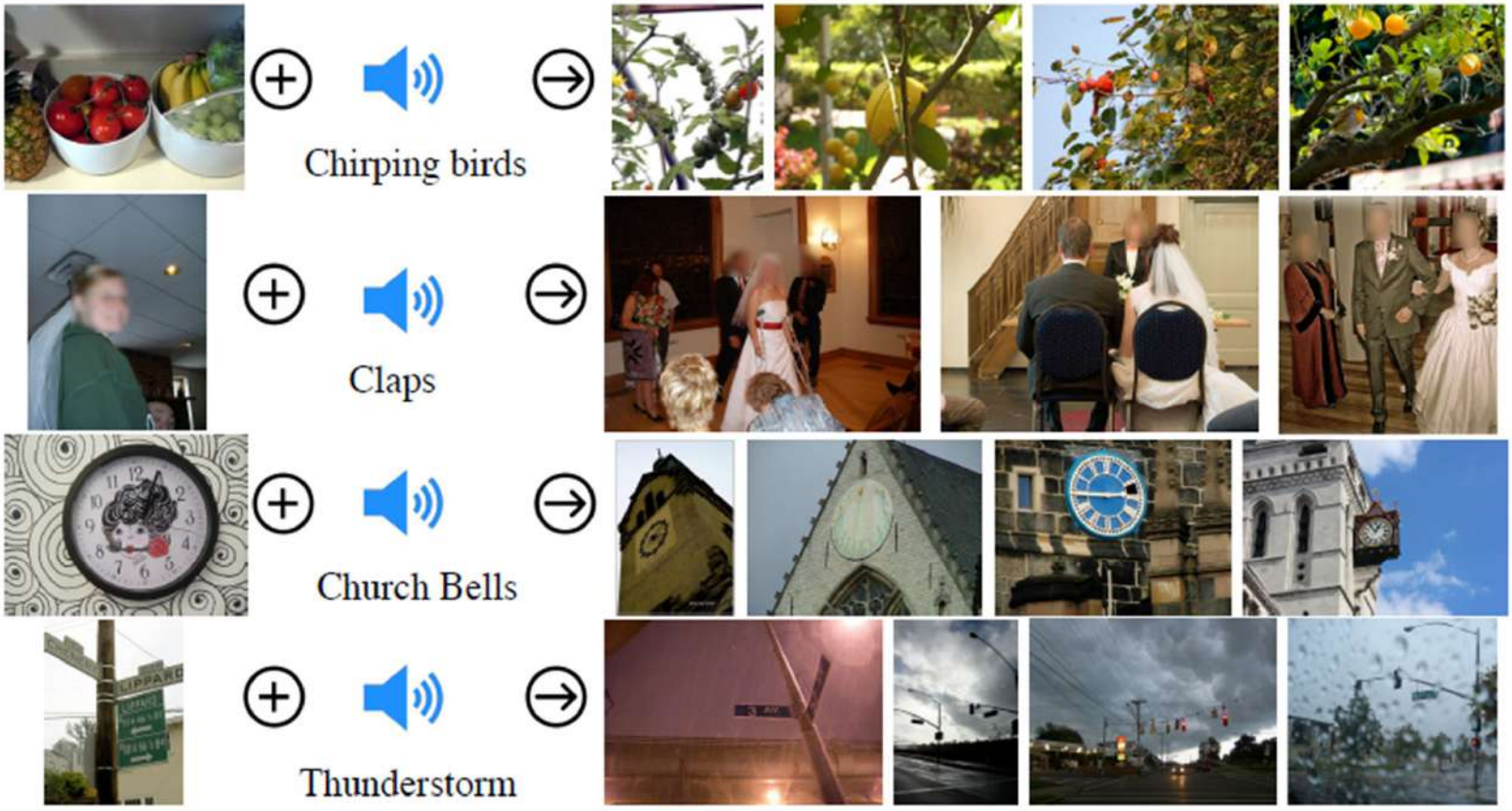
UCF Center for Research in Computer Vision

Image Bind

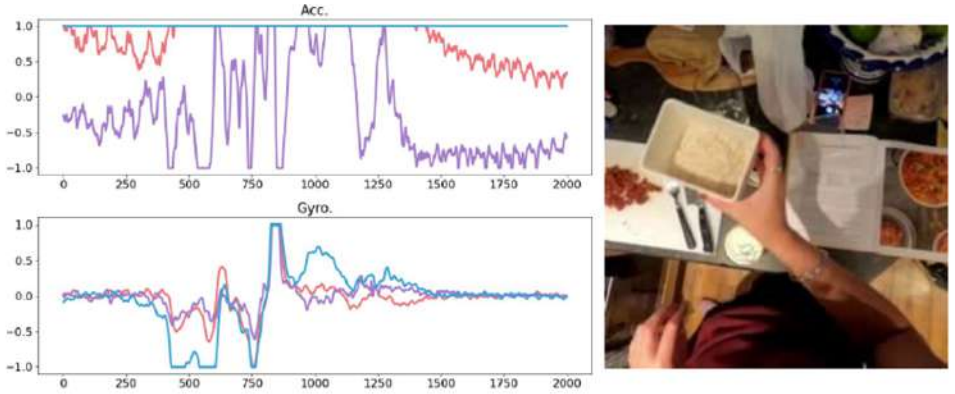




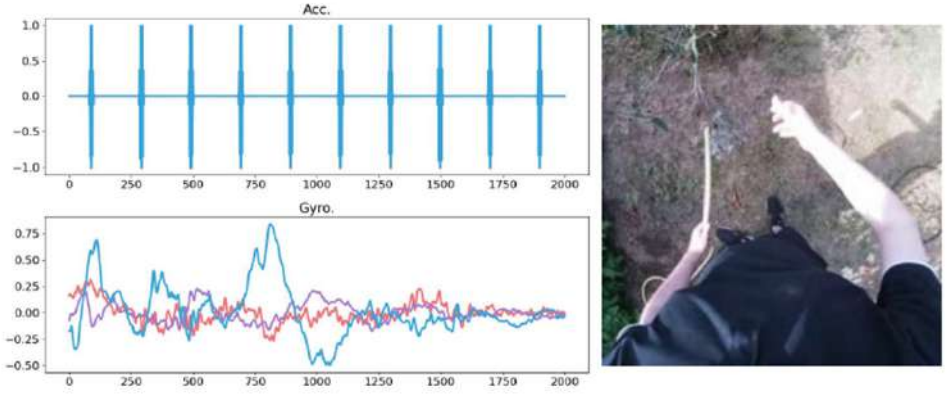



Chirping birds
Claps
Church Bells
Thunderstorm

Text query: *"Cooking a meal"*



Text query: *"A person doing gardening work outdoors"*



UCF CENTER FOR RESEARCH IN COMPUTER VISION

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

LANGUAGEBIND: EXTENDING VIDEO-LANGUAGE PRETRAINING TO N-MODALITY BY LANGUAGE-BASED SEMANTIC ALIGNMENT

Bin Zhu[1,2,*], Bin Lin[1,*], Munan Ning[3,4], Yang Yan[1], JiaXi Cui[1], Hongfa Wang[2], Yatian Pang[3], Wenhao Jiang[4], 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

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

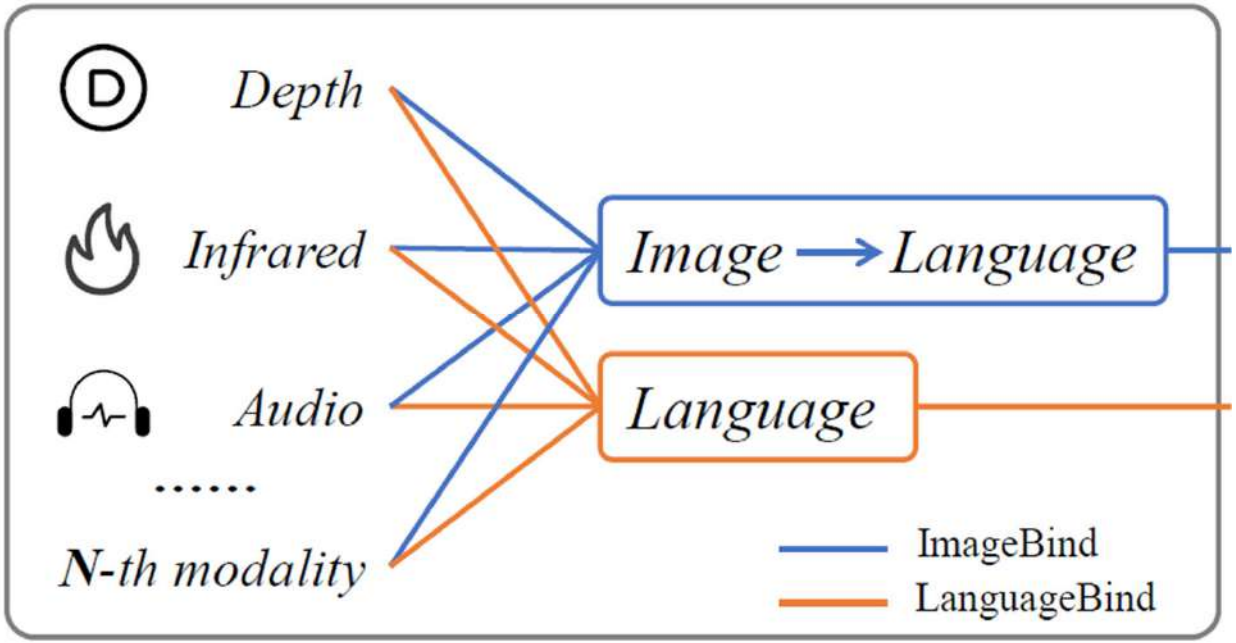
1 INTRODUCTION

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.

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

VIDAL-10M dataset

- 3 million pairs of video-language data,
- 3 million pairs of infrared- language data,
- 3 million pairs of depth-language data, and
- 1 million pairs of audio-language data.

VIDAL-10M construction

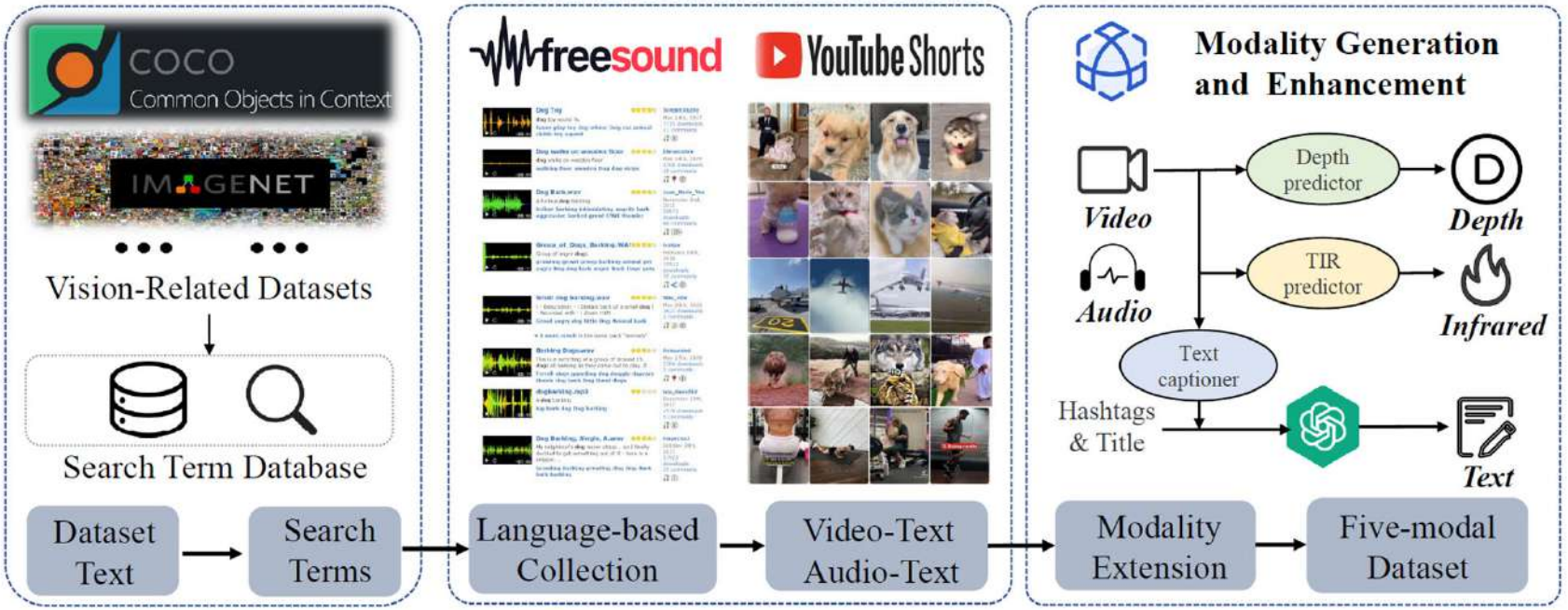
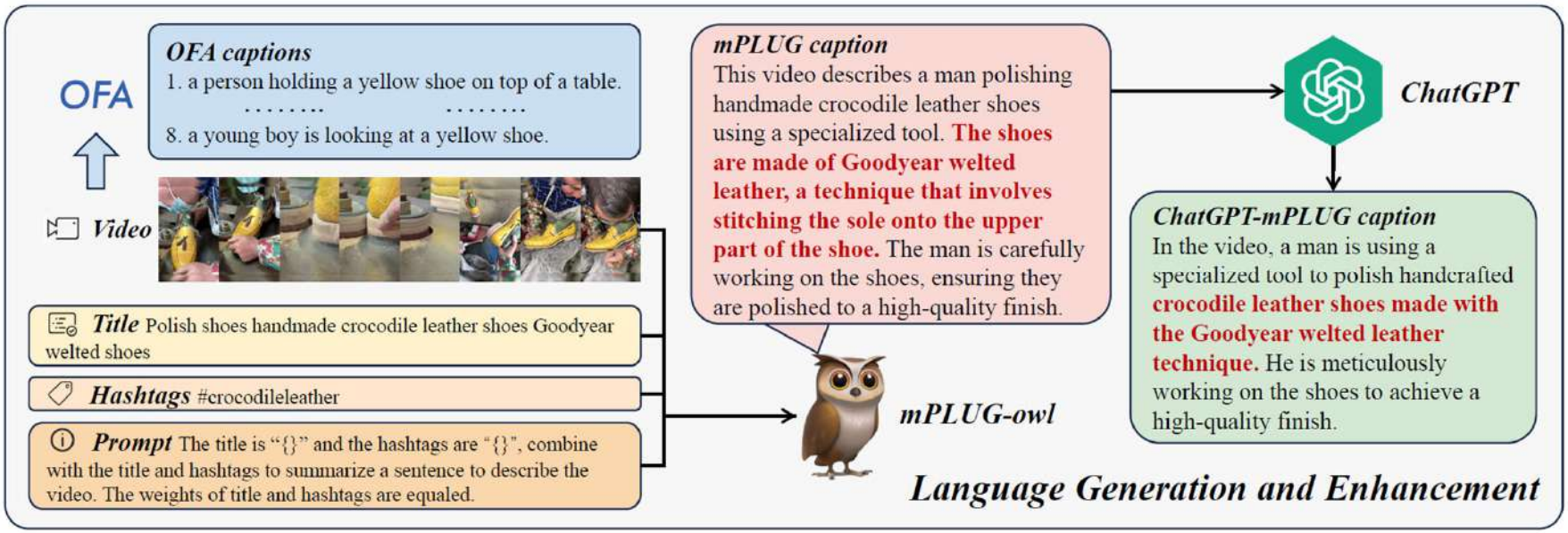


Table 9: Examples of textual descriptions from various datasets as search terms.

Dataset	Search terms
YouTube-8M	How to make a delicious chocolate cake. Learn to dance salsa in 10 easy steps. .....
Howto100M	How to play chess. How to make pizza. .....
ImageNet	lesser panda, red panda, panda, bear cat, cat bear, Ailurus fulgens, coon bear killer whale, killer, grampus, sea wolf, Orcinus orca, giant panda, panda, panda bear .....
COCO	A small boat floating on a body of water with a city skyline in the background. A man with a red helmet on a small moped on a dirt road. .....
Others	.....

Text generation and enhancement pipeline



Infrared and depth modality generation

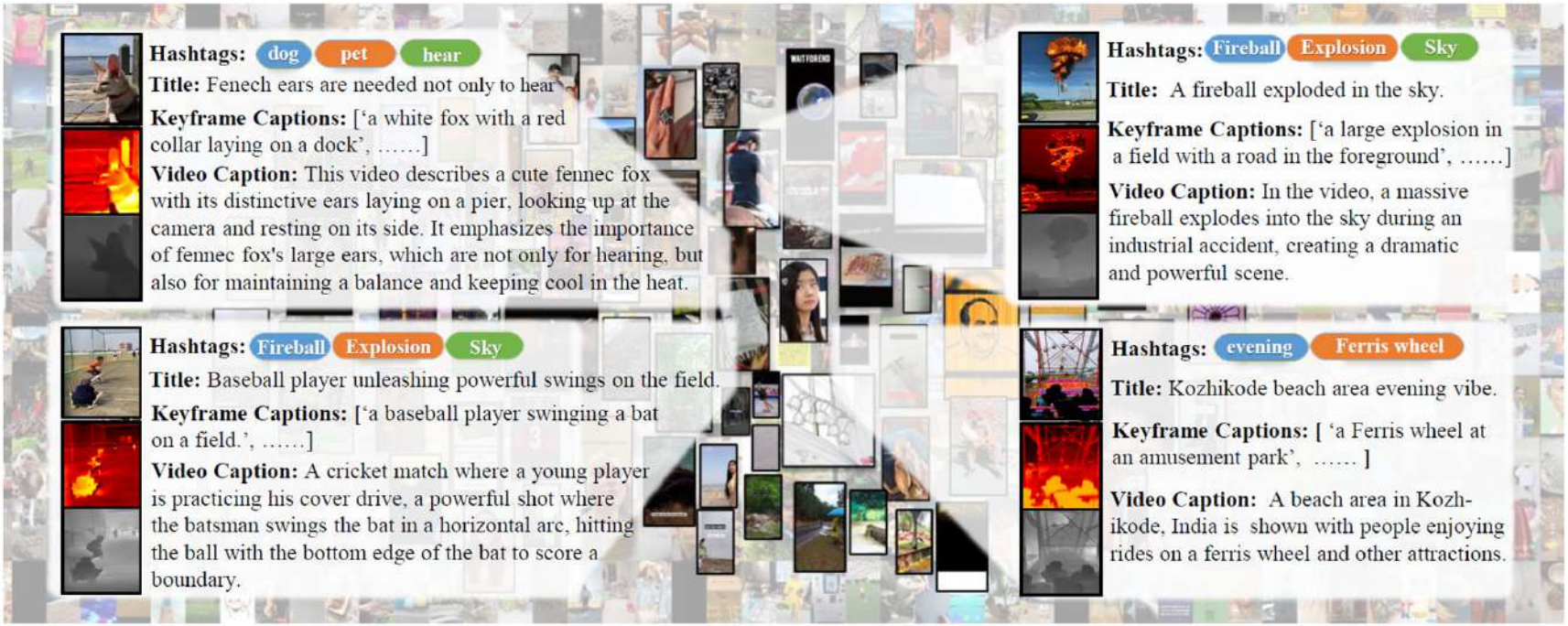
- Generative Models are used to Infrared and Depth modalities
  - The sRGB-TIR model is used for infrared modality generation
  - GLPN model for depth modality generation

UCF Center for Research in Computer Vision

video-audio-text-depth-infrared pairs

Pets & Animals

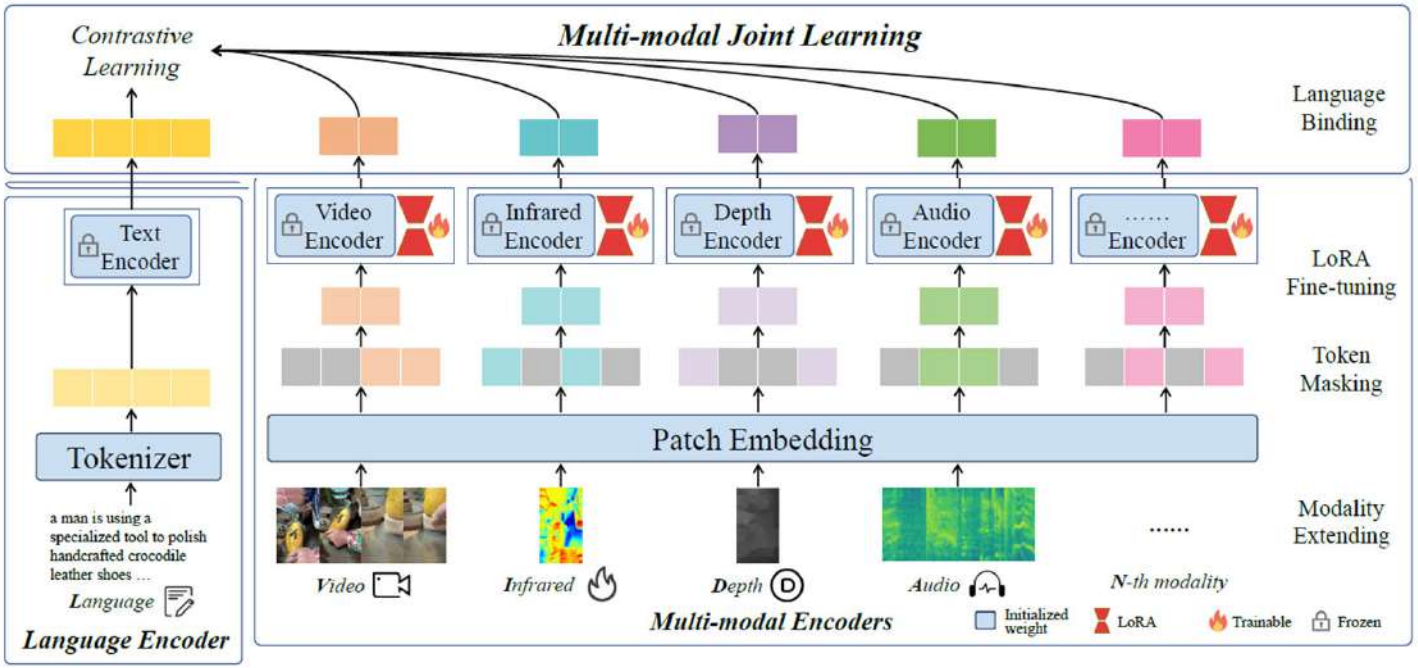
News & Politics



Sports

Education

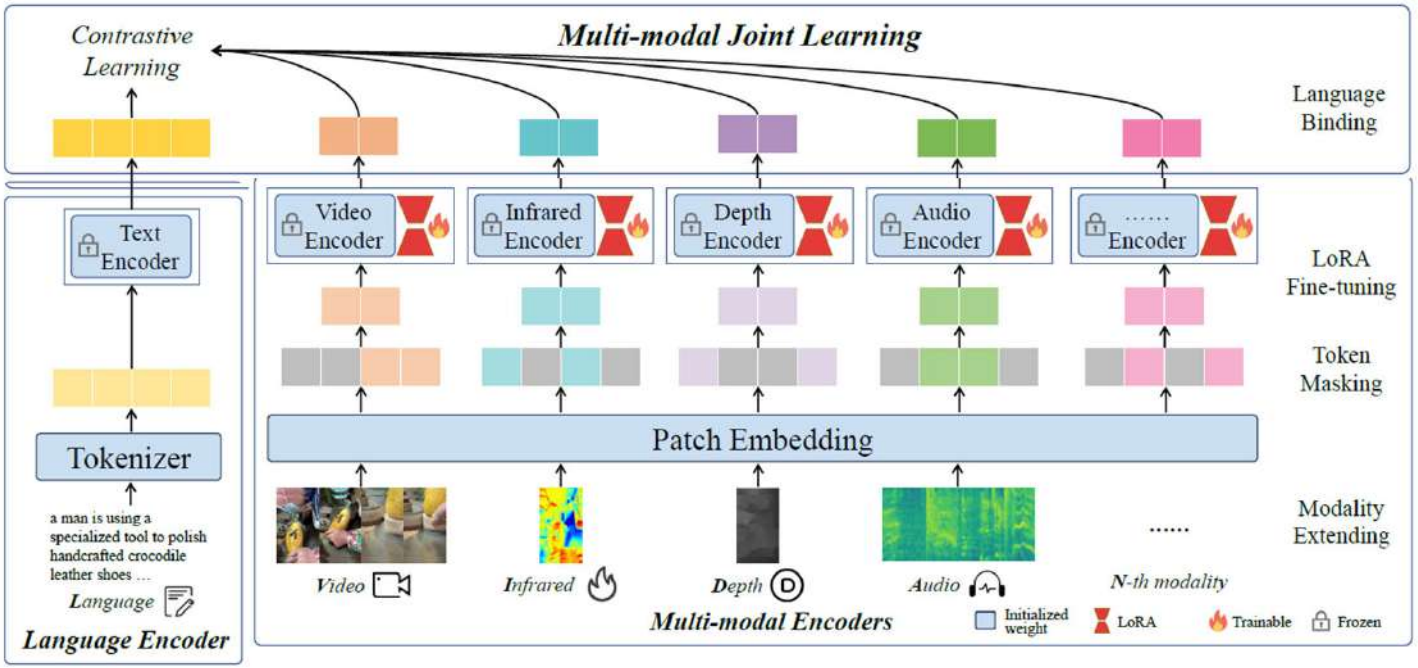
Language-Bind



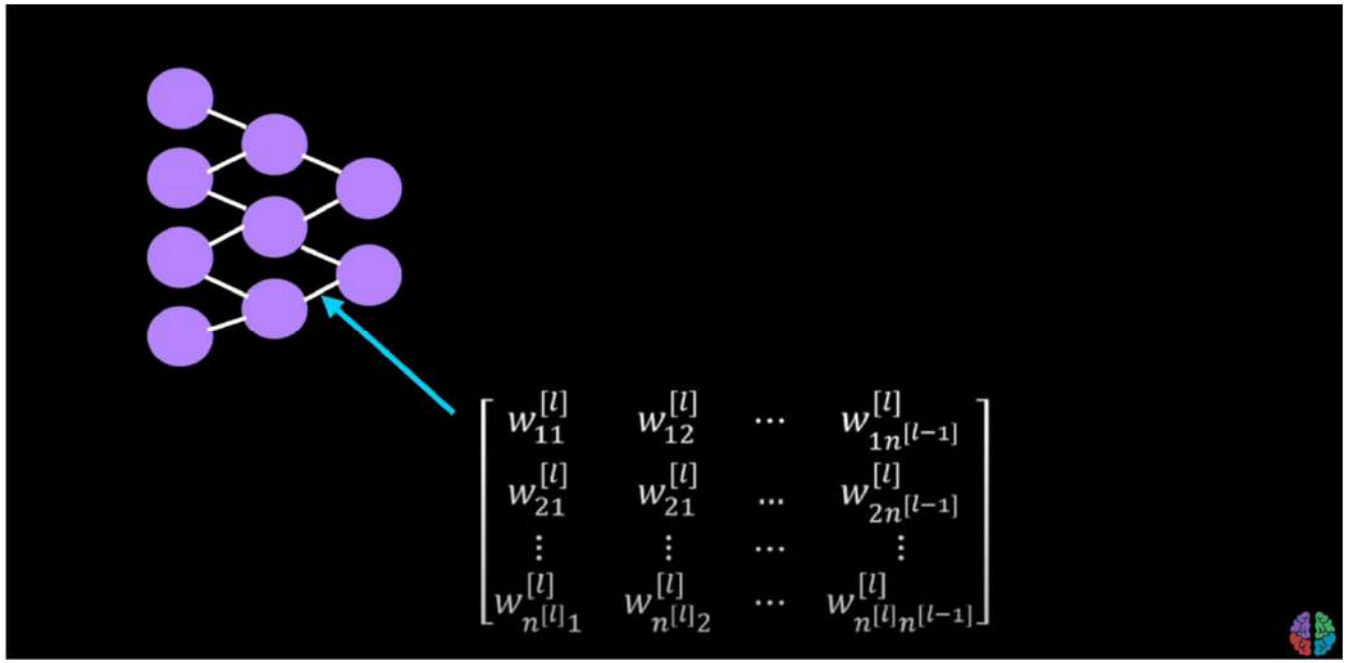
Multi-Modal Encoders

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

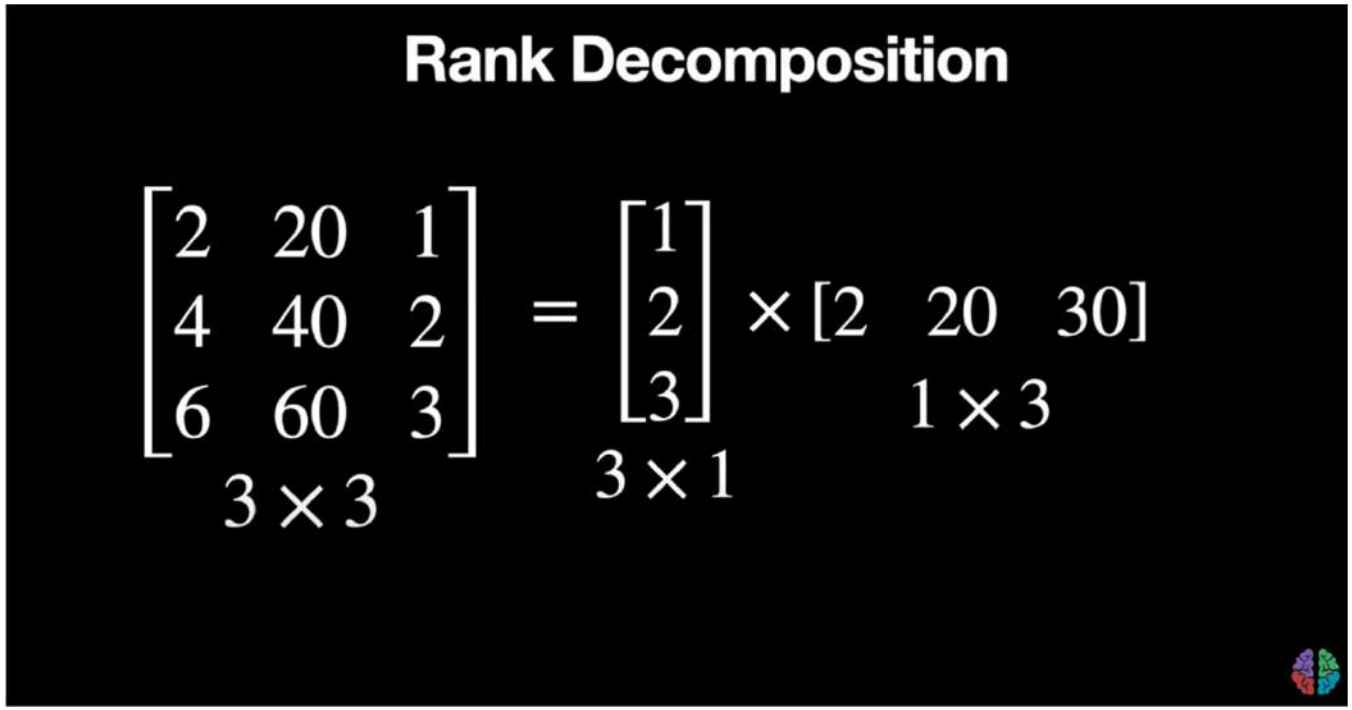
Language-Bind



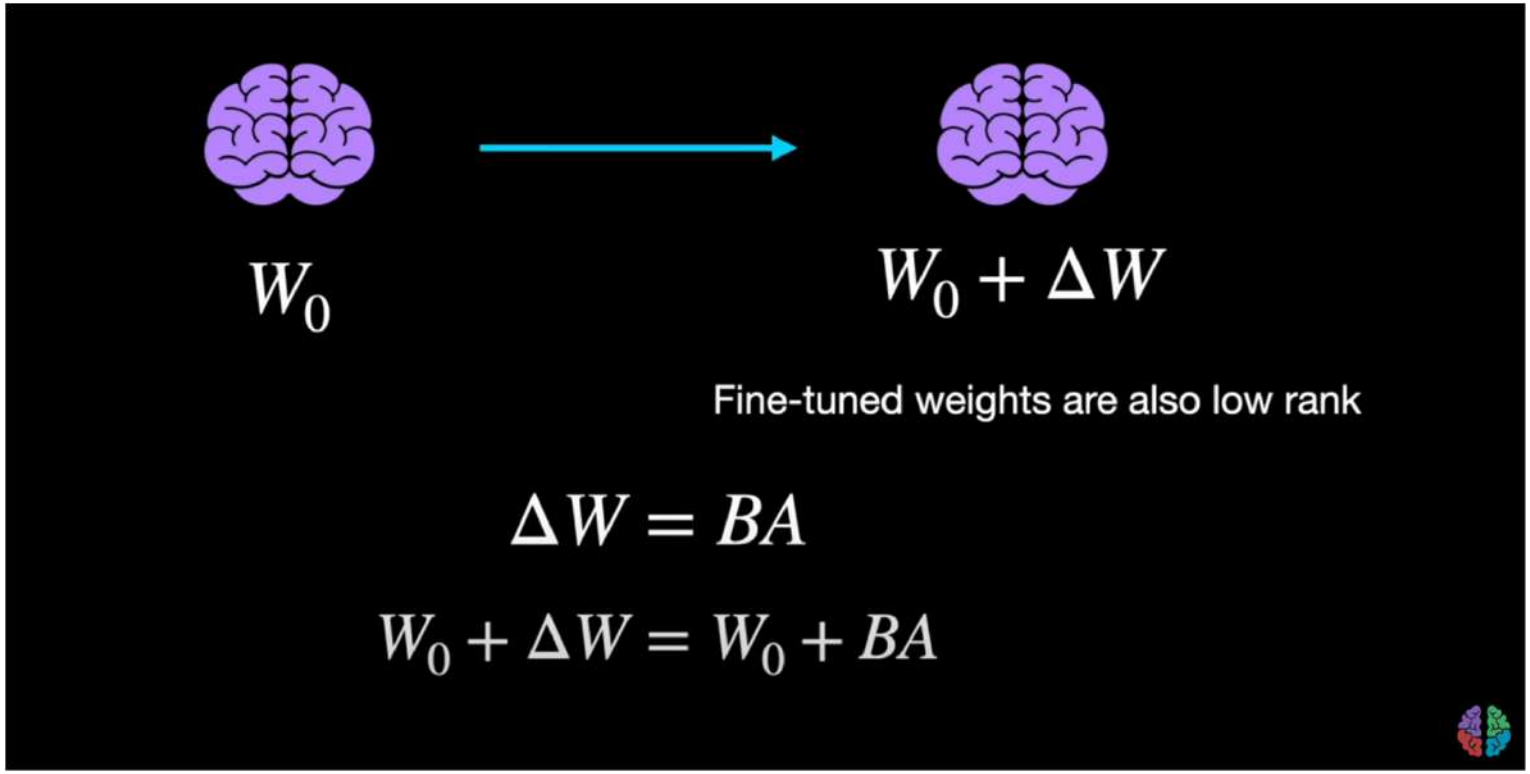
Weight Matrix



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

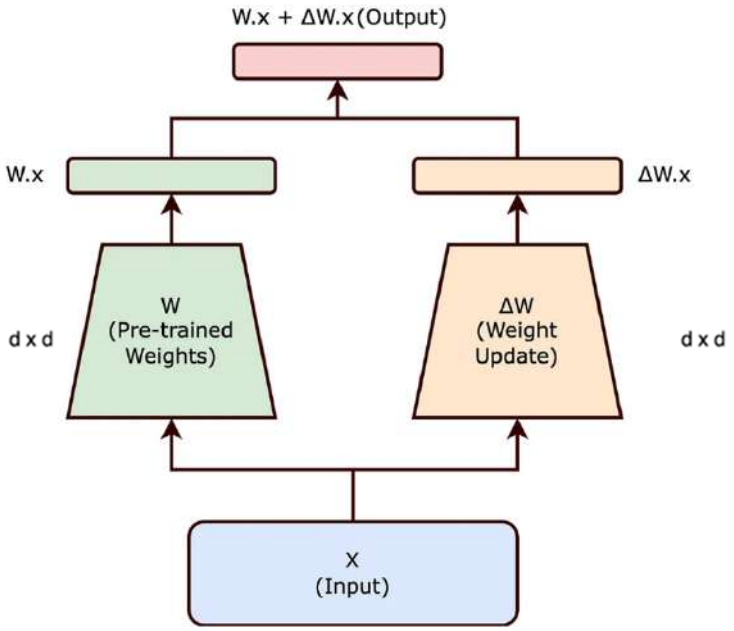


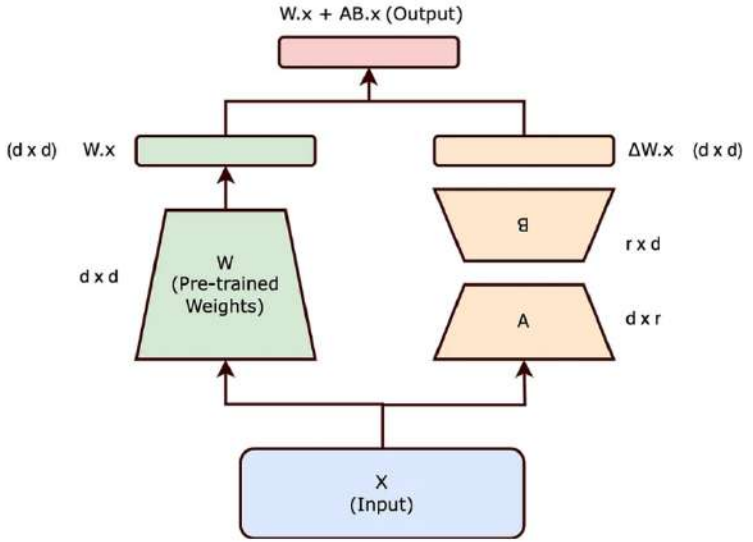
https://www.youtube.com/watch?v=X4VvO3G6_vw



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

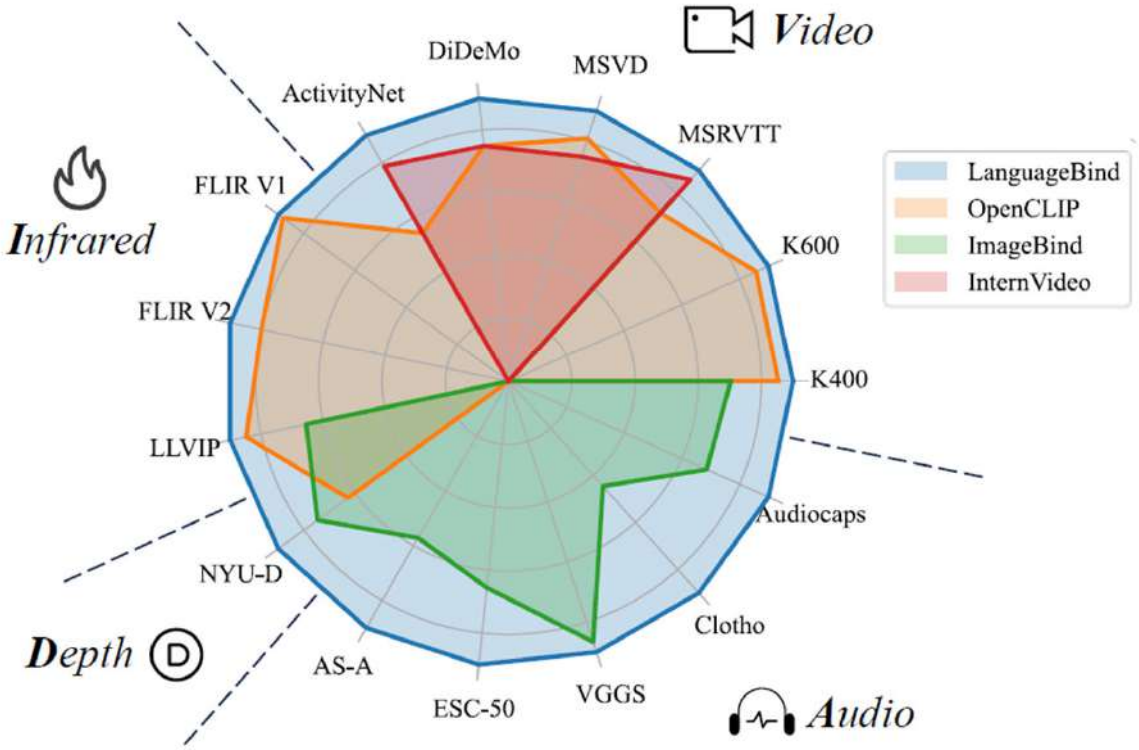
LORA Finetuning





https://towardsdatascience.com/understanding-lora-low-rank-adaptation-for-finetuning-large-models-936bce1a07c6

Results


Video
DiDeMo
MSVD
ActivityNet
MSRVTT
FLIR V1
K600
Infrared
FLIR V2
K400
LLVIP
Audiocaps
NYU-D
Clotho
Depth
AS-A
VGGS
ESC-50
Audio
LanguageBind
OpenCLIP
ImageBind
InternVideo

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, 28], segmentation [23, 59, 54] and captioning [46, 23], 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

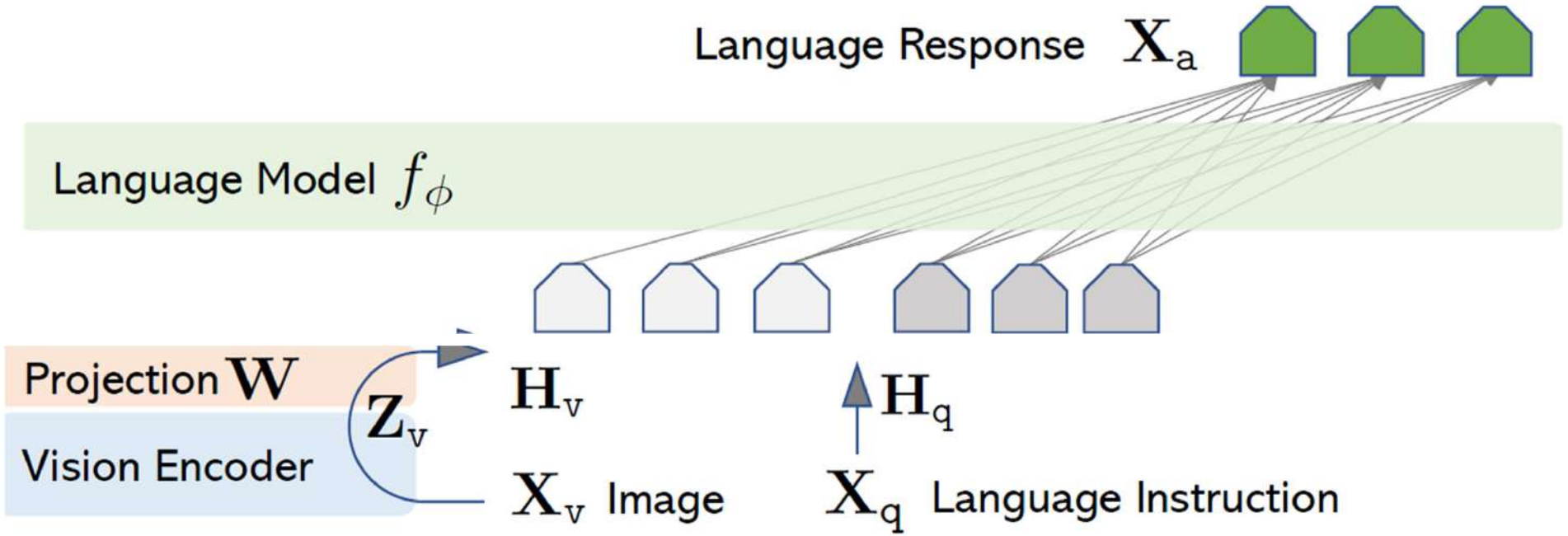
Preprint. Work in progress

660 citations: 2024
4749 citations: 2025

LLaVA: Large Language and Vision Assistant

- 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

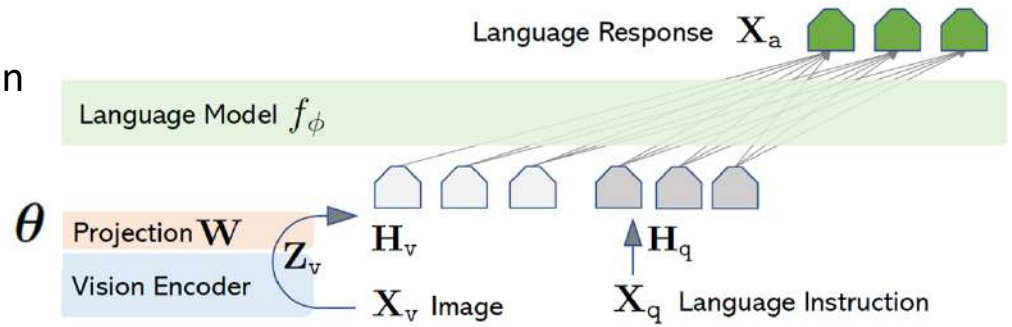


$$\mathbf{H}_v = \mathbf{W} \cdot \mathbf{Z}_v, \quad \text{with } \mathbf{Z}_v = g(\mathbf{X}_v)$$

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 $\theta = \mathbf{W}$ (the projection matrix)



Stage 2: Fine-tuning End-to-End

$$\theta \doteq \{\mathbf{W}, \phi\}$$

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

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 $\cdots$

$\mathbf{X}_{\texttt{system-message}}$ = 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 $\mathbf{X}_\texttt{v}$

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

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

Instruction Following Data

- 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

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"

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

Follow up work on LLaVA

- LLaVA-1.5
- LLaVA-Next
- LLaVA-01
- Video-LLaVA

LLaVA 1.5

arXiv:2310.03744v2 [cs.CV] 15 May 2024

Improved Baselines with Visual Instruction Tuning

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

Abstract

Large multimodal models (LMM) have recently shown encouraging progress with visual instruction tuning. In this paper, we present the first systematic study to investigate the design choices of LMMs in a controlled setting under the LLaVA framework. We show that the fully-connected vision-language connector in LLaVA is surprisingly powerful and data-efficient. With simple modifications to LLaVA, namely, using CLIP-ViT-L-336px with an MLP projection and adding academic-task-oriented VQA data with response formatting prompts, we establish stronger baselines that achieve state-of-the-art across 11 benchmarks. Our final 13B checkpoint uses merely 1.2M publicly available data, and finishes full training in ∼1 day on a single 8-A100 node. Furthermore, we present some early exploration of open problems in LMMs, including scaling to higher resolution inputs, compositional capabilities, and model hallucination, etc. We hope this makes state-of-the-art LMM research more accessible. Code and model will be publicly available.

1. Introduction

Large multimodal models (LMMs) have become increasingly popular in the research community, as they are the key building blocks towards general-purpose assistants [2, 30, 43]. Recent studies on LMMs are converging on a central concept known as visual instruction tuning [36]. The results are promising, *e.g.* LLaVA [36] and MiniGPT-4 [62] demonstrate impressive results on natural instruction-following and visual reasoning capabilities. To better understand the capability of LMMs, multiple benchmarks [17, 27, 34, 37, 55] have been proposed. Recent works further demonstrate improved performance by scaling up the pretraining data [3, 14, 54], instruction-following data [14, 18, 29, 58], visual encoders [3], or language models [39], respectively. The LLaVA architecture is also leveraged in different downstream tasks and domains, including region-level [8, 56] and pixel-level [26, 50] understanding, biomedical assistants [31], image generation [5], adversarial studies [6, 59].

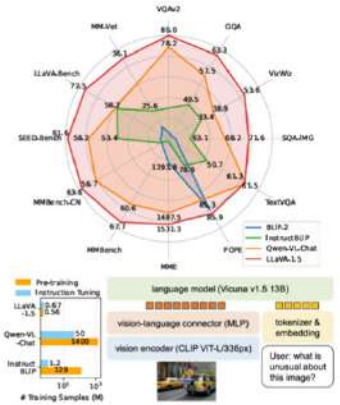


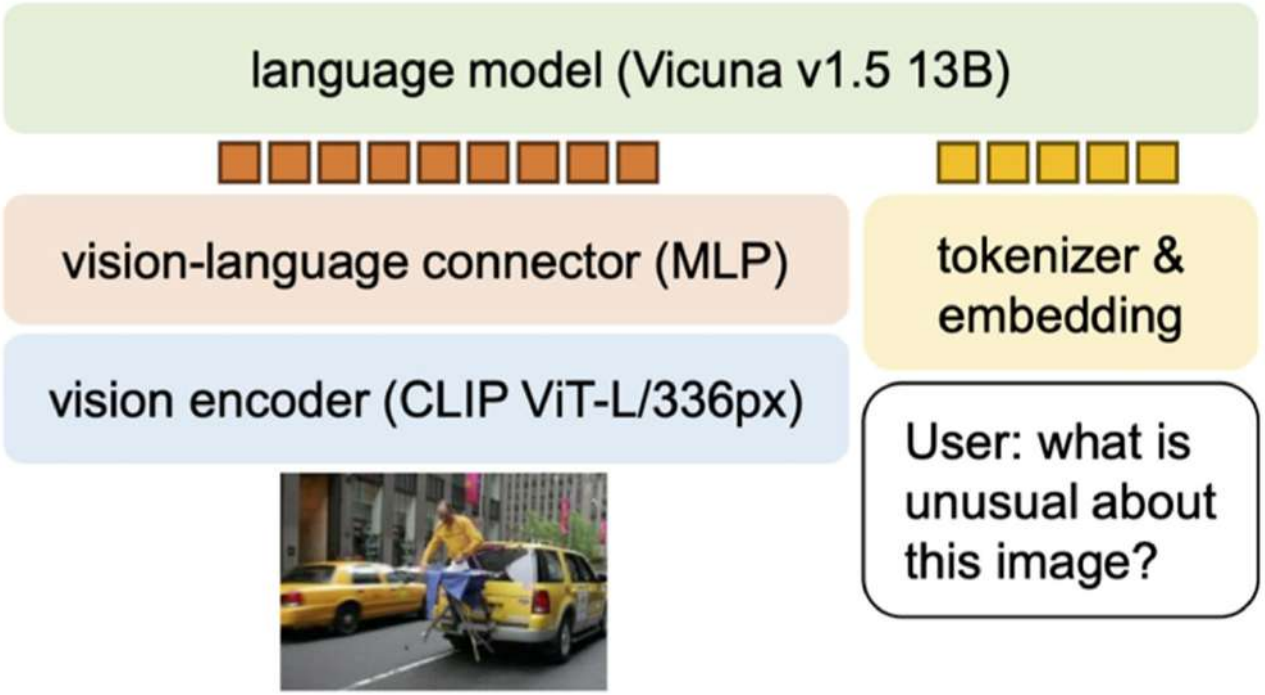
Figure 1. **LLaVA-1.5** achieves SoTA on a broad range of 11 tasks (Top), with high training sample efficiency (Left) and simple modifications to LLaVA (Right): an MLP connector and including academic-task-oriented data with response formatting prompts.

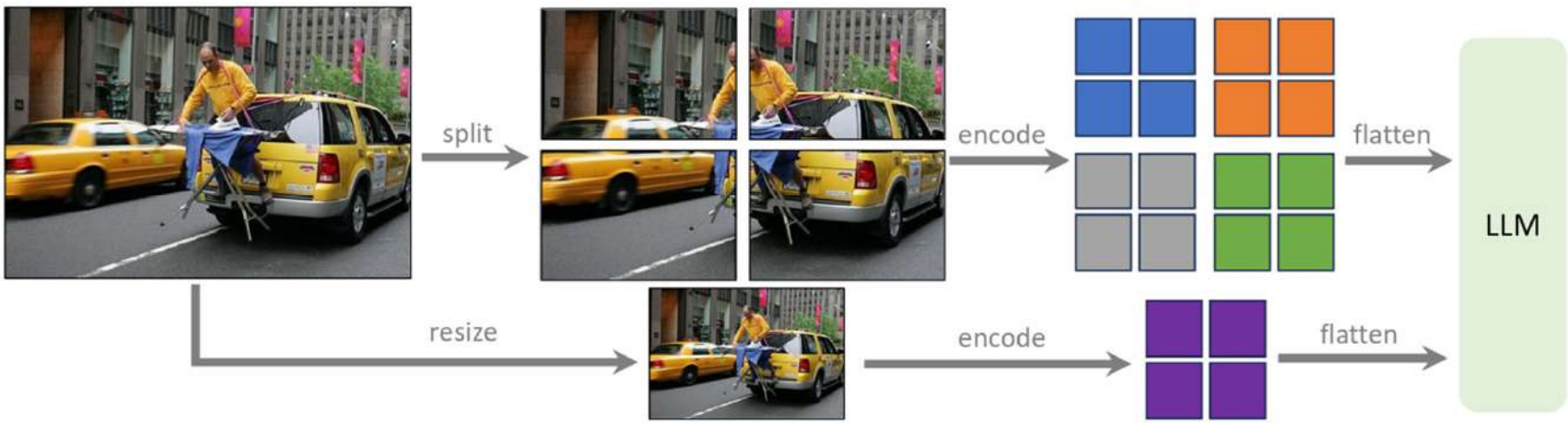
However, despite many benchmarks and developments, it still remains unclear what the best recipe is to train LMMs towards the goal of general-purpose assistants. For example, LLaVA [36] excels in conversational-style visual reasoning and even outperforms later approaches like InstructBLIP [14] on such benchmarks [55], while InstructBLIP excels in traditional VQA benchmarks that demands single-word or short answers. Given the significant differences in the model architecture and training data between them, the root cause of the disparity in their capabilities remains elusive, despite conjectures [37, 55]: the amount of training data, the usage of resamplers like Qformer [32], *etc.* To this

Main Points

- MLP Instead of Linear layer
- More Datasets
- Higher Resolution Images
- Prompt Formatting

LLaVA1.5 Architecture




split
encode
flatten
LLM
resize
encode
flatten

Prompt Formatting

Visual input example, Multitask Balancing Problem:	
	
User	Is this unusual? Please explain in detail.
InstructBLIP	yes

(a) Example of InstructBLIP [14] (Vicuna-13B) having difficulty balancing between short- and long-form answers.

Visual input example, Different Format Prompts:	
Normal prompt	What is the color of the shirt that the man is wearing?
Response	The man is wearing a yellow shirt.
Ambiguous prompt	Q: What is the color of the shirt that the man is wearing? A:
Response	The man is wearing a yellow shirt.
Formatting prompt	What is the color of the shirt that the man is wearing? **Answer the question using a single word or phrase.**
Response	Yellow.

	Method	LLM	Res.	GQA	MME	MM-Vet
	InstructBLIP	14B	224	49.5	1212.8	25.6
	Only using a subset of InstructBLIP training data					
0	**LLaVA**	7B	224	–	809.6	25.5
1	+VQA-v2	7B	224	47.0	1197.0	27.7
2	+Format prompt	7B	224	46.8	1323.8	26.3
3	+MLP VL connector	7B	224	47.3	1355.2	27.8
4	+OKVQA/OCR	7B	224	50.0	1377.6	29.6
	Additional scaling					
5	+Region-level VQA	7B	224	50.3	1426.5	30.8
6	+Scale up resolution	7B	336	51.4	1450	30.3
7	+GQA	7B	336	62.0*	1469.2	30.7
8	+ShareGPT	7B	336	62.0*	1510.7	31.1
9	+Scale up LLM	13B	336	**63.3***	**1531.3**	**36.1**

Visual input example, Constrained JSON Output:	
	
User	Please read the text in this image and return the information in the following JSON format (note xxx is placeholder, if the information is not available in the image, put "N/A" instead). {"class": xxx, "DLN": xxx, "DOB": xxx, "Name": xxx, "Address": xxx, "EXP": xxx, "ISS": xxx, "SEX": xxx, "HGT": xxx, "WGT": xxx, "EYES": xxx, "HAIR": xxx, "DONOR": xxx}
GPT-4V	{"class": "C", "DLN": "1234568", "DOB": "08/31/1977", "Name": "Ima Cardholder", "Address": "2570 24th Street, Anytown, CA 95818", "EXP": "08/31/2014", "ISS": "08/31/2011", "SEX": "F", "HGT": "5-05", "WGT": "125 lbs", "EYES": "BRN", "HAIR": "BRN", "DONOR": "VETERAN"}
LLaVA	{"class": "Driver License", "DLN": "CA123 456 789", "DOB": "01/01/1990", "Name": "Amy", "Address": "Anytown, CA", "EXP": "08/31/2023", "ISS": "Veteran", "SEX": "F", "HGT": "5'5", "WGT": "130", "EYES": "Brown", "HAIR": "Brown", "DONOR": "Unknown"}
LLaVA-1.5	{"class": "C", "DLN": "1132456789", "DOB": "08/23/1971", "Name": "Ima Cardholder", "Address": "Anytown, State, Zip", "EXP": "08/23/2014", "ISS": "California", "SEX": "F", "HGT": "5'5", "WGT": "123", "EYES": "Brown", "HAIR": "Brown", "DONOR": "N/A"}

LLaVA-NeXT: Improved reasoning, OCR, and world knowledge

1. **Increasing the input image resolution** to 4x more pixels: up to 672x672, 336x1344, 1344x336 resolution.

2. **Better visual reasoning and OCR capability** with an improved visual instruction tuning data mixture.

3. **Better visual conversation for more scenarios**, covering different applications. Better world knowledge and logical reasoning.

4. **Efficient deployment and inference**

UCF Center for Research in Computer Vision

arXiv:2411.10440v1 [cs.CV] 15 Nov 2024

LLaVA-o1: Let Vision Language Models Reason Step-by-Step

Guowei Xu[2] Peng Jin[1,3,4] Li Hao[1,3,4] Yibing Song[5] Lichao Sun[6] Li Yuan[1,3,4*]
[1]School of Electronic and Computer Engineering, Peking University
[2] Institute for Interdisciplinary Information Sciences, Tsinghua University [3]Peng Cheng Laboratory
[4]AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School
[5]Alibaba DAMO Academy [6]Computer Science and Engineering, Lehigh University

*Correspondence: yuanli-ece@pku.edu.cn

Abstract

Large language models have demonstrated substantial advancements in reasoning capabilities, particularly through inference-time scaling, as illustrated by models such as OpenAI's o1. However, current Vision-Language Models (VLMs) often struggle to perform systematic and structured reasoning, especially when handling complex visual question-answering tasks. In this work, we introduce LLaVA-o1[1], a novel VLM designed to conduct autonomous multistage reasoning. Unlike chain-of-thought prompting, LLaVA-o1 independently engages in sequential stages of summarization, visual interpretation, logical reasoning, and conclusion generation. This structured approach enables LLaVA-o1 to achieve marked improvements in precision on reasoning-intensive tasks. To accomplish this, we compile the LLaVA-o1-100k dataset, integrating samples from various visual question answering sources and providing structured reasoning annotations. Besides, we propose an inference-time stage-level beam search method, which enables effective inference-time scaling. Remarkably, with only 100k training samples and a simple yet effective inference time scaling method, LLaVA-o1 not only outperforms its base model by 8.9% on a wide range of multimodal reasoning benchmarks, but also surpasses the performance of larger and even closed-source models, such as Gemini-1.5-pro, GPT-4o-mini, and Llama-3.2-90B-Vision-Instruct.

1. Introduction

Large language models, represented by OpenAI o1 [63], demonstrate strong capabilities for systematic and in-depth reasoning, validating the efficacy of inference-time scaling for language models [47]. However, vision is equally important for enabling models to fully understand the world and extend their cognitive abilities [6]. Therefore, developing a multimodal model that integrates both language and vision while facilitating effective, systematic, and deep reasoning holds substantial significance.

Early open-source vision language models (VLMs) mainly employ a direct prediction approach [21, 30, 32], generating brief answers immediately in response to a question. The main limitation of this direct-response paradigm is its lack of a structured reasoning process, making it less effective for tasks demanding logical reasoning [62]. Recent studies have shown that incorporating Chain-of-Thought (CoT) reasoning encourages the model to reason step by

[1]There are similar names of recent VLM works. To clarify, LLaVA-o1 is built upon Llama-3.2-Vision model [40], rather than LLaVA [32].

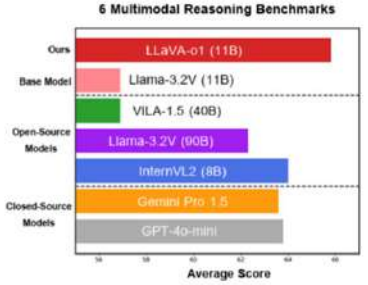


Figure 1. **Performance of LLaVA-o1 and other models across six multimodal reasoning benchmarks.** Although LLaVA-o1 is fine-tuned from the Llama-3.2-11B-Vision-Instruct [40] model (which has the lowest average score), it outperforms many larger open-source models and even some closed-source models. Detailed benchmark results are shown in Table 7.

UCF Center for Research in Computer Vision

Example 1: Reasoning Problems	***Answer from Llama-3.2-11B-Vision-Instruct:***	***Answer from LLaVA-o1:***
 ***Question:*** Subtract all tiny shiny balls. Subtract all purple objects. How many objects are left? ***Options:*** A. 4, B. 8, C. 2, D. 6	To determine the correct answer, we need to follow the instructions provided. The first instruction is to subtract all tiny shiny balls. Since there are **three tiny shiny balls** in the image, we would be left with no shiny balls after performing this action. The second instruction is to subtract all purple objects. However, upon closer inspection, it becomes apparent that there are **no shiny purple objects present in the image**. Therefore, we cannot subtract any purple objects. Given that we have already removed all the tiny shiny balls and there were no purple objects to begin with, the correct answer is C. 2, as initially, **there were 2 purple objects, but they are not to be subtracted** according to the instruction.	***Summary stage: What's the problem? What should I do?*** <SUMMARY> I will examine the image to identify and count all the objects, **focusing on the tiny shiny balls and purple objects**. Then, I will subtract these objects from the total count and **determin the number of remaining objects**. </SUMMARY> ***Caption stage: What can I know from the image?*** <CAPTION> The image shows various 3D geometric shapes on a flat surface, including spheres, cubes, and cylinders in different colors. Notably, there are **small shiny green spheres and a purple cylinder**. </CAPTION> ***Reasoning stage: How to solve the problem step-by-step?*** <REASONING> First, I will count all the objects in the image: there are **10 objects in total**. Next, I will identify and subtract the tiny shiny balls. There is **one small shiny green sphere**. Then, I will identify and subtract the purple objects. There is **one purple cylinder**. By subtracting the two objects (the shiny ball and the purple cylinder) from the total, I find that **8 objects remain**. </REASONING> ***Conclusion stage: What is the final answer?*** <CONCLUSION> B </CONCLUSION>