Lecture 4:
Cross-modal Alignment and Multimodal Scene Graph

Papers for Lecture 4: Cross-modal Alignment and Multimodal Scene Graph

P4-1: Cross-modal Alignment: (Presenter: Chai Zenghao; Reader: Stefan Putra Lionar)

(Must-Read) J Li, D Li, S Savarese & S Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ICML 2023.

(Must-Read): A Radford, JW Kim, C Hallacy, et al. Learning Transferable Visual Models from Natural Language Supervision. ICML 2021.

(To-Read): L Qu, M Liu, J Wu, Z Gao & L Nie. Dynamic Modality Interaction Modeling for Image-Text Retrieval. SIGIR 2021.

P4-2: Multimodal Scene Graph: (Presenter: Dibyadip Chatterjee; Reader: xx)

(Must-Read) J Yang, W Peng, X Li et al. Panoptic Video Scene Graph Generation. CVPR 2023.

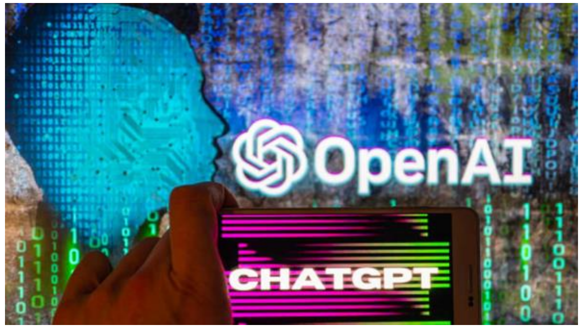
(To-Read) K Tang, Y Niu, J Huang et al. Unbiased Scene Graph Generation From Biased Training. CVPR 2020.

(First Dataset, Must-Read) R Krishna, Y Zhu, O Groth, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. IJCV 2017

- **Discussion of BNI (Brave New Idea) papers**

From LLMs to MFMs

- **Recently, Large Language Models (LLMs) have exhibited impressive abilities to handle various human-level tasks, aligned-well with human preference**

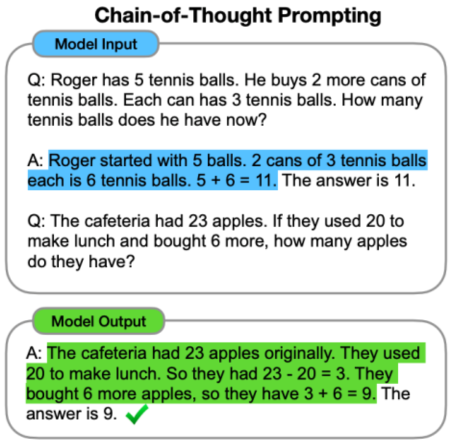


ChatGPT Passes Google Coding Interview for Level 3 Engineer With $183K Salary

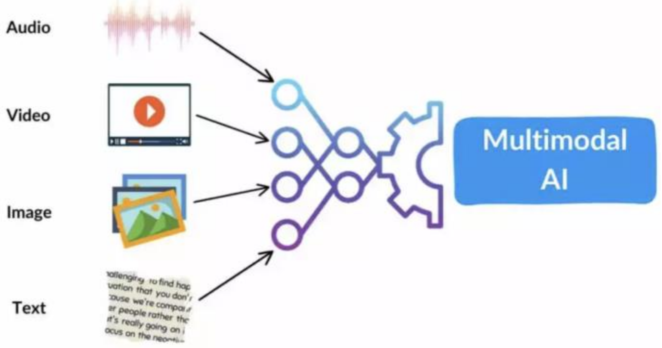
'Amazingly, ChatGPT gets hired at L3 when interviewed for a coding position,' reads a Google document, but ChatGPT itself says it can't replicate human creativity and problem-solving skills.

AI Passes U.S. Medical Licensing Exam

— Two papers show that large language models, including ChatGPT, can pass the USMLE



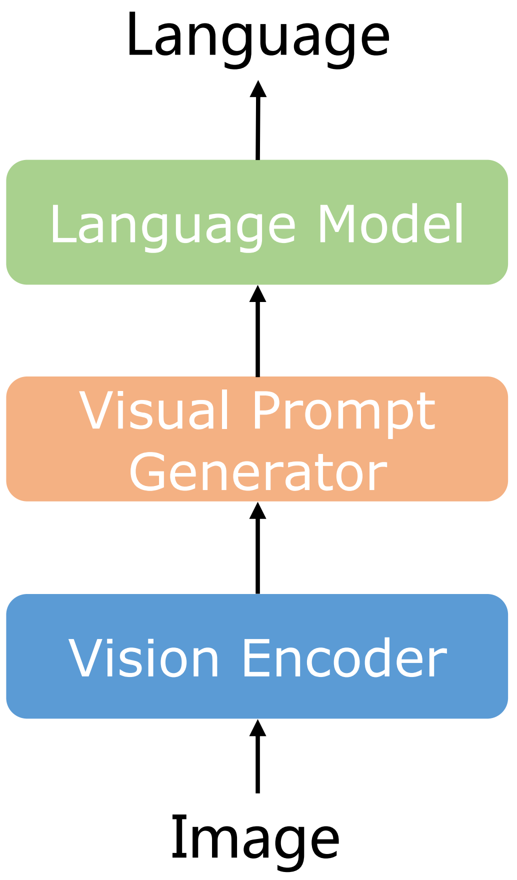
- **However, LLMs can only understand text-only data while multimodal tasks are more diverse in nature.**

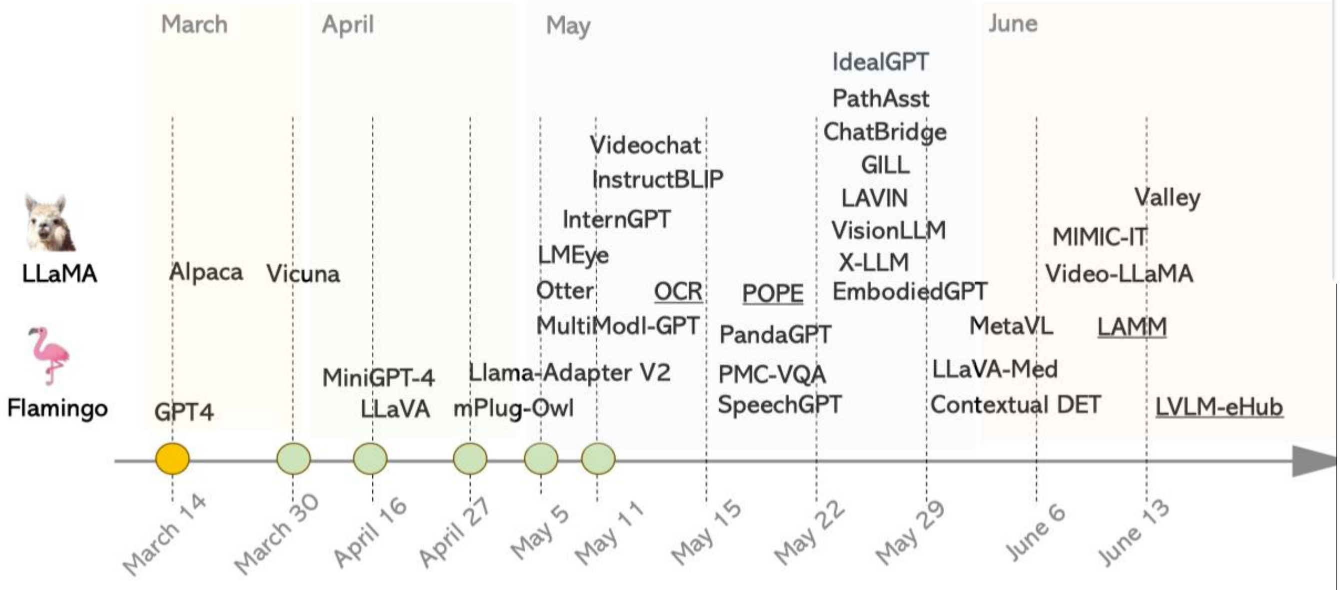


- **To achieve a longstanding aspiration of building unified Multimodal Foundation Models (MFMs)**

Multimodal Foundation Models

- **General idea:** adapt frozen instruction-tuned LLMs to understand visual inputs





GPT-4 visual input example, Chicken Nugget Map:

User: Can you explain this meme?

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

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

GPT-4 visual input example, Extreme Ironing:

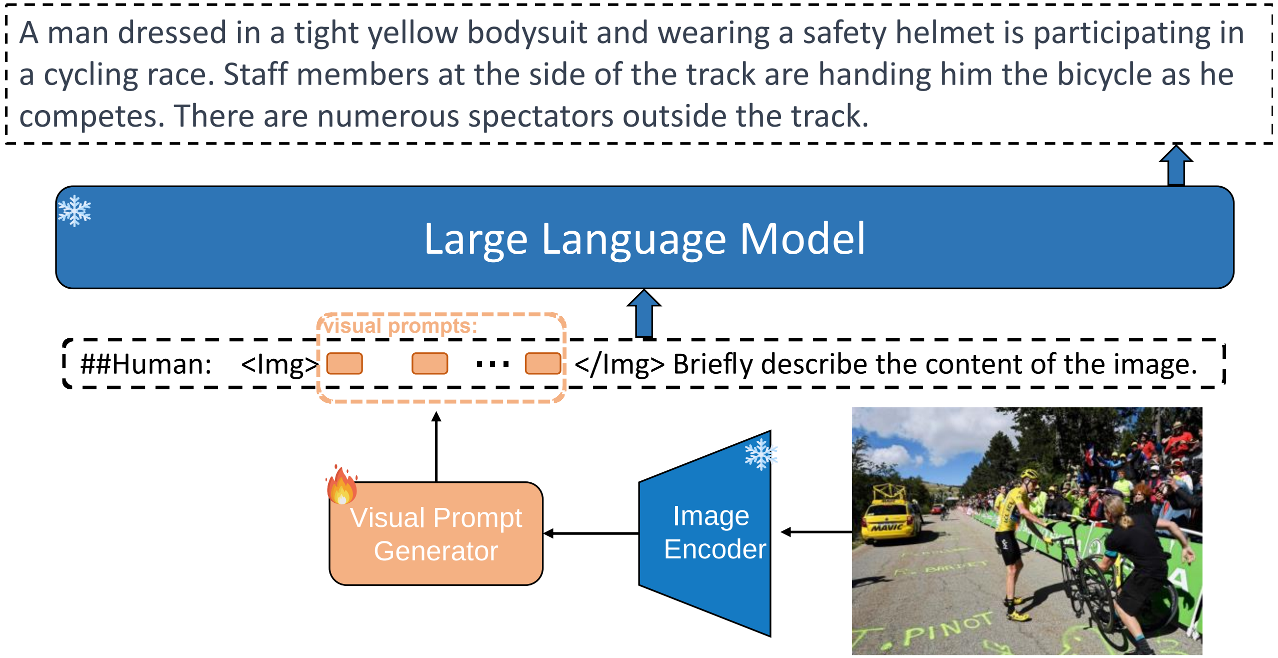
User: What is unusual about this image?

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

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

Multimodal Foundation Models

- Backbone modules:
  - Pre-trained large language models (e.g., Llama) and vision encoder (e.g., CLIP-ViT)
- Visual Prompt Generator (e.g. Linear Projection, Q-former, Perceiver):
  - Trainable modules to connect to two modalities
  - Translate visual features into tokens that LLMs can recognize.
- Training stages:
  - Multimodal alignment
  - Multimodal instruction tuning

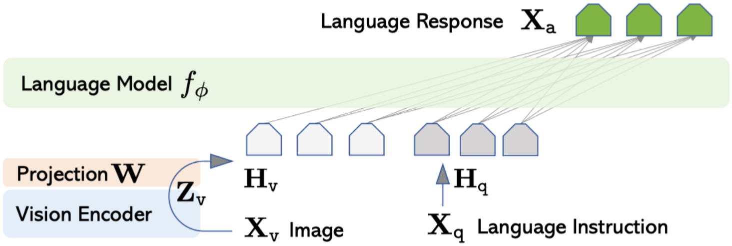


Multimodal Alignment

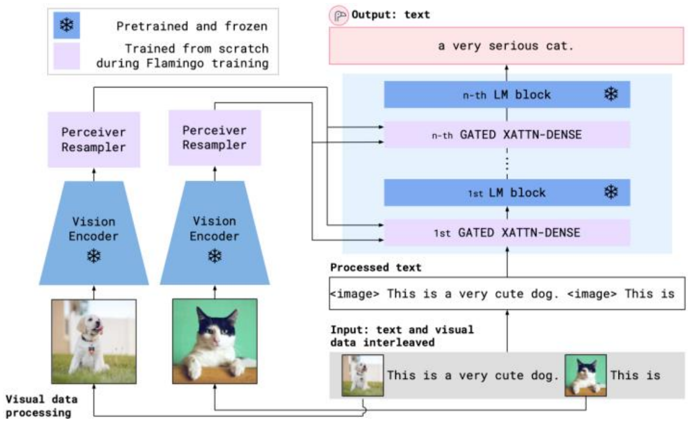
- Align different modalities with LLMs and enable LLMs to reason with multimodal information.

- Two categories of multimodal alignment methods:

- **Image tokens as prefixes:**
  - train a VPG to translates the visual features into tokens that LLMs can recognize
  - LLaVA, MiniGPT-4, …

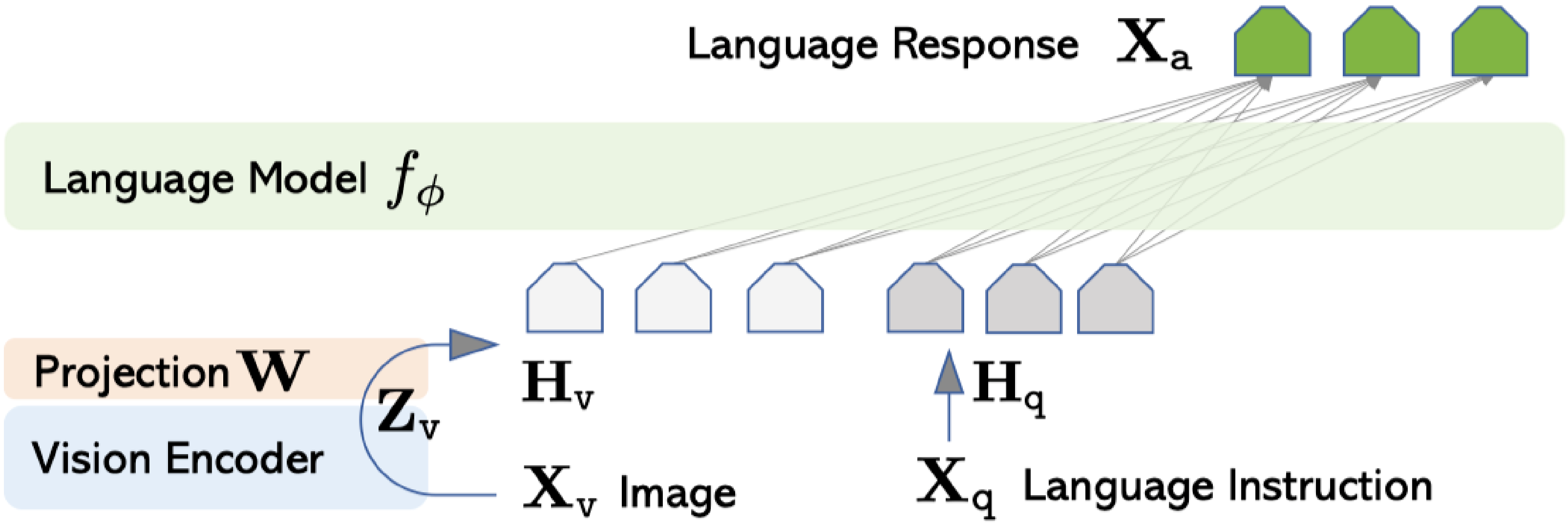


- **Cross-attention for feature fusion:**
  - adopt cross-attention to integrate the visual features into LLMs
  - Flamingo, LLaMA-Adapter, …



Multimodal Alignment: Image Tokens as Prefixes (LLaVA)

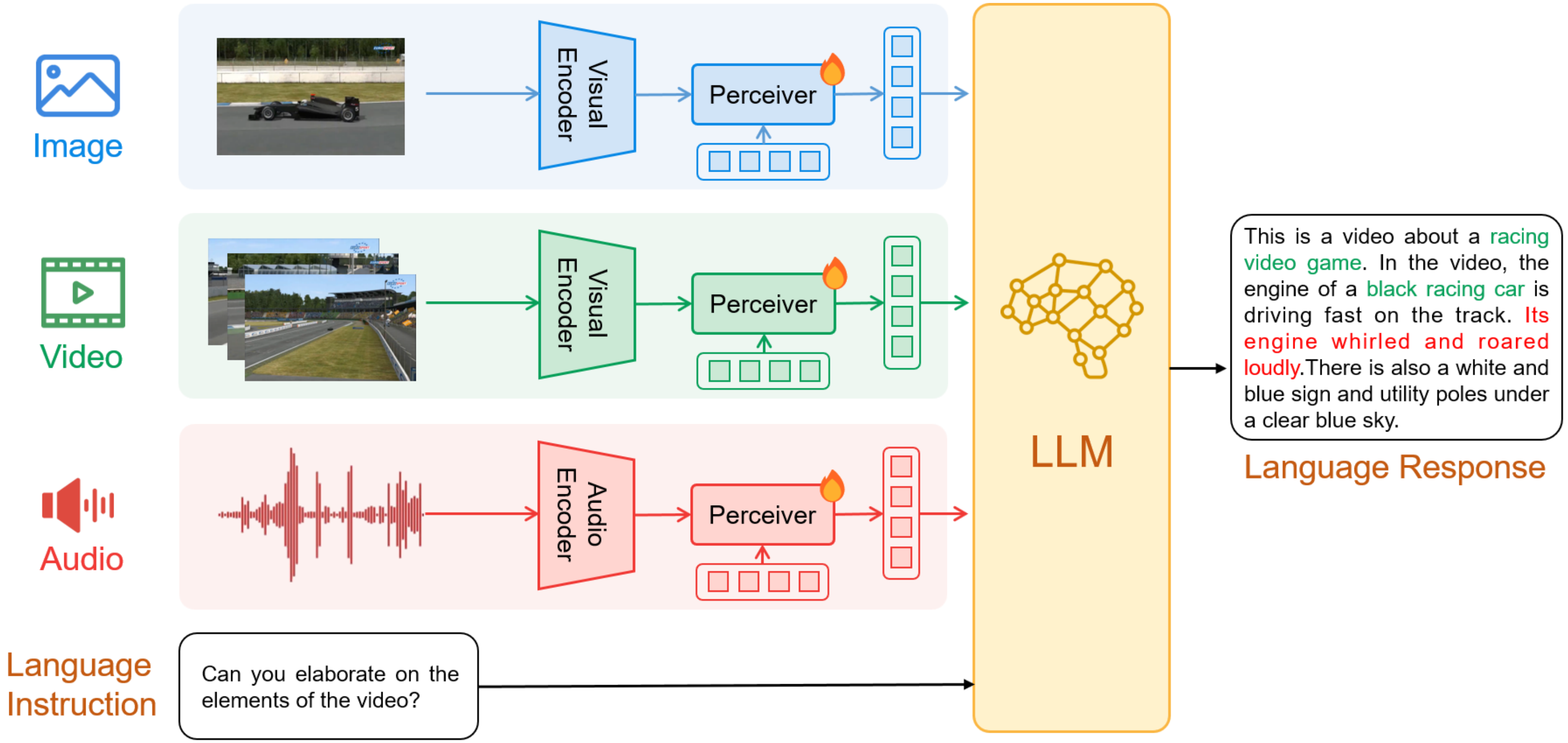
- Visual Prompt Generator (e.g. Linear Projection, Q-former, Perceiver):
  - trained on millions of image-caption pairs by requiring the frozen LLM to generate captions conditioned on the VPG-generated tokens of images.
  - training objective: auto-regressive loss on language output



Visual Instruction Tuning. Liu et al. NeurIPS 2023.

Align with More Modalities: ChatBridge

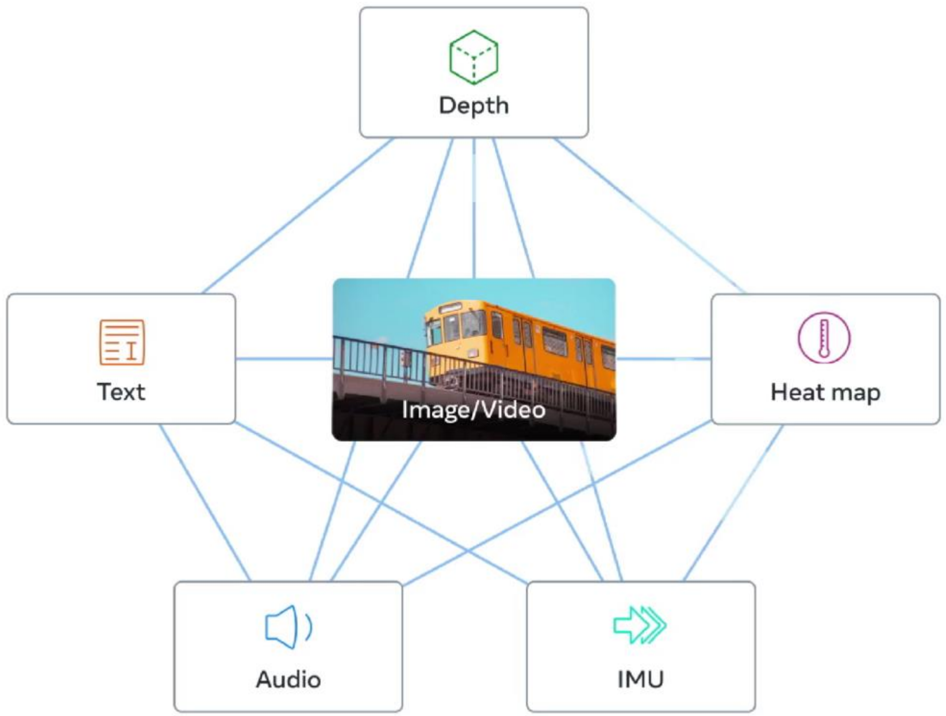
- Bridge various modalities with Large Language Model as a language catalyst
- **First stage:** aligns each modality with language, which brings emergent multimodal correlation and collaboration abilities
- **Second stage:** align it with user intent with our newly proposed multimodal instruction tuning dataset

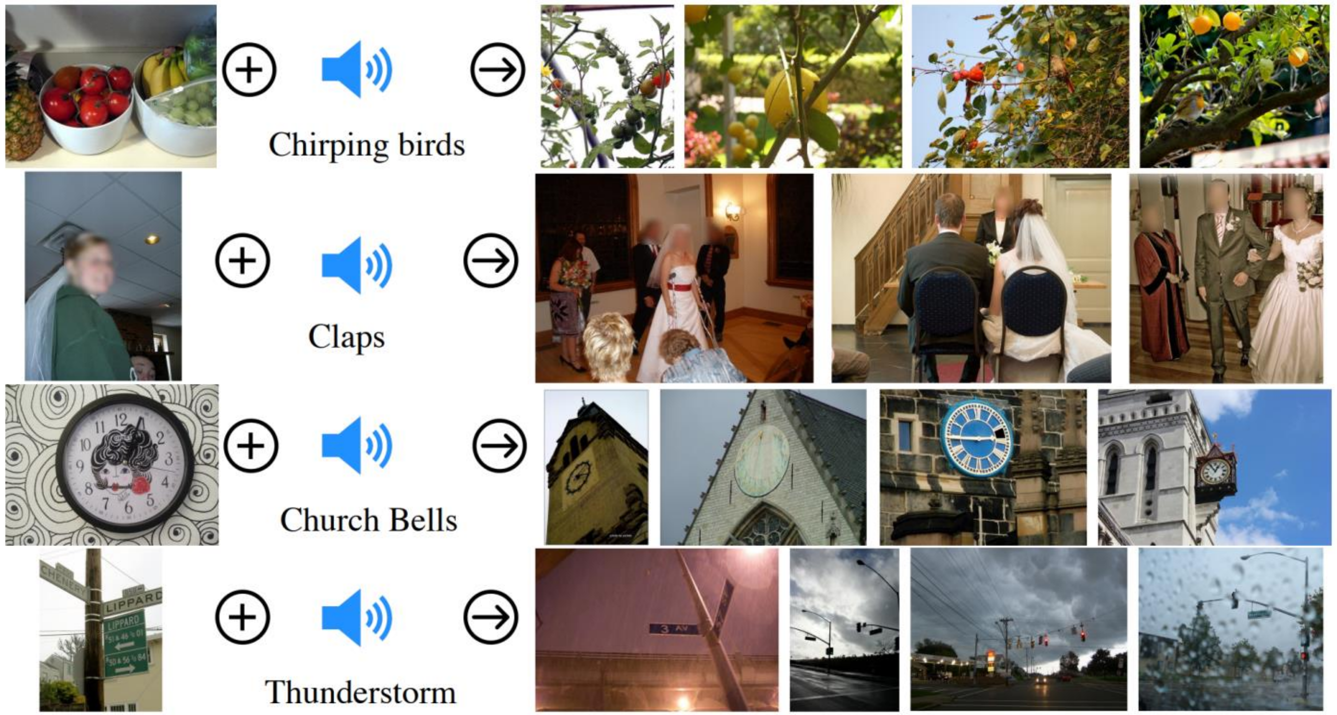


ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst. Zhao et al. Arxiv 2023.

Align with More Modalities: ImageBind

- Learn a joint embedding across 6 different modalities: images, text, audio, depth, thermal & IMU data
- Aligning each modality's embedding to image embeddings:
  - leads to an emergent alignment across all of the modalities
  - Why use image as bridge? There are lots of paired data of different modalities with images



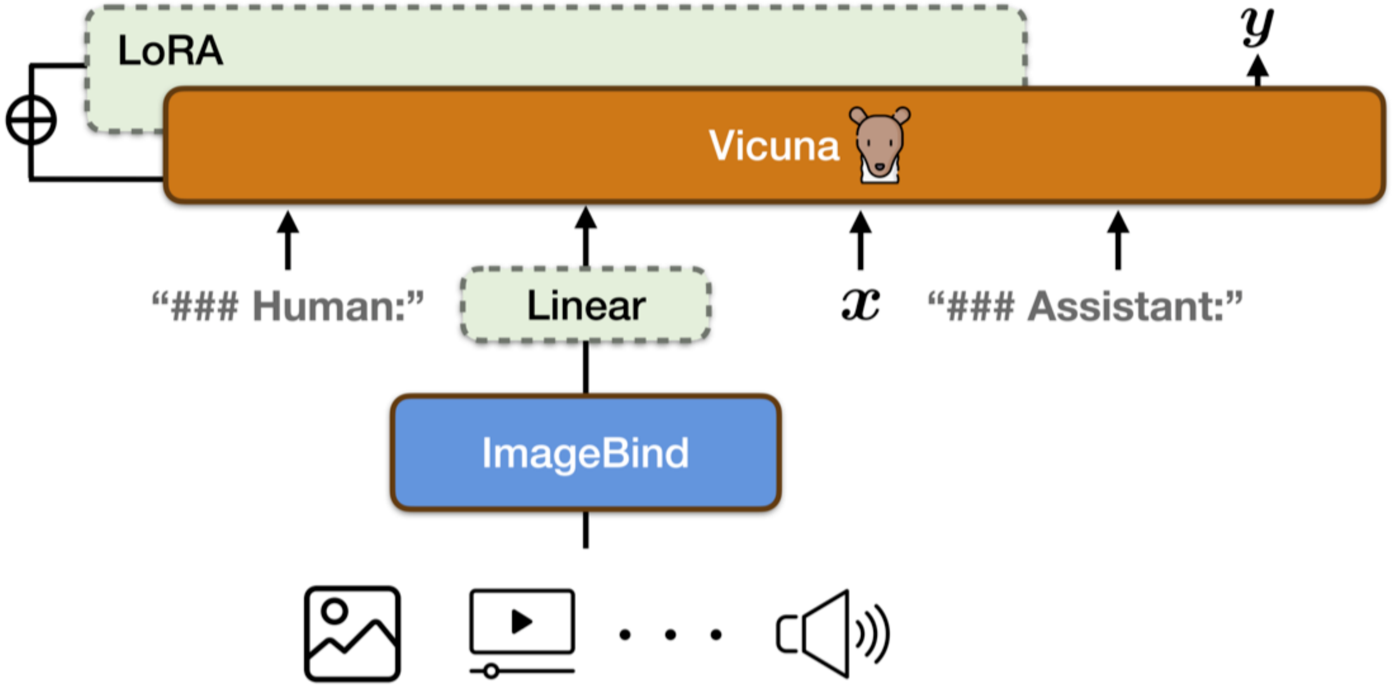


ImageBind: One Embedding Space to Bind Them All. Girdhar et al. Meta AI. CVPR 2023.

Align with More Modalities: PandaGPT

- This work aligns Vicuna with the multimodal encoder of ImageBind, but use only image-text pairs

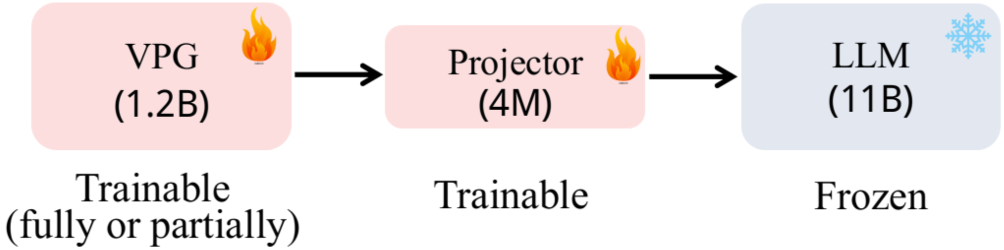
- Thanks to the strong capability of ImageBind in embedding data from different modalities into the same space, PandaGPT displays emergent (zero-shot) cross-modal behaviors for data other than image and text (e.g., video, audio, depth, thermal, and IMU)



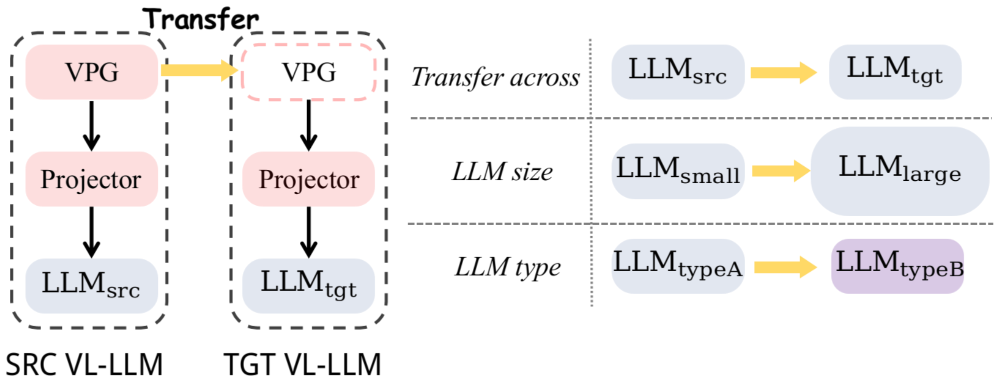
PandaGPT: One Model To Instruction-Follow Them All. Su et al. Arxiv 2305.16355.

Transfer Visual Prompt Generator Across LLMs VPGTrans

- **Motivation:** To build new MFMs with limited cost.
- **Key Idea:** Transfer existing visual perception modules to new LLMs (bigger size or new types) for building MLLMs.



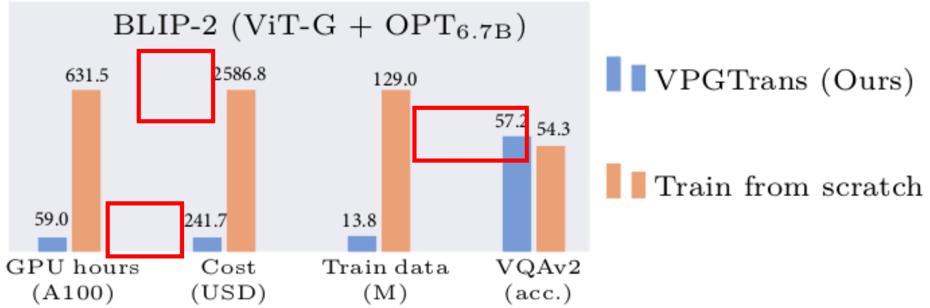
(a) Typical MLLM Architecture



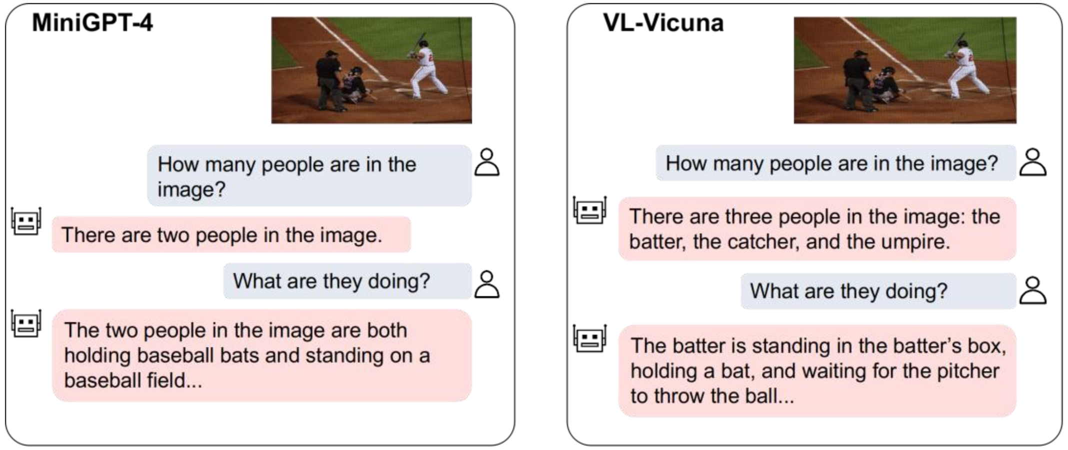
VPG transfer across LLMs

(b) Key Idea

- **Results:**
  - Cost reduction w/o performance drop.
  - New MLLMs, e.g., VL-Vicuna.



(c) Cost Reduction



(d) Demo

Zhang, et al. Transfer Visual Prompt Generator across LLMs. Arxiv.

Key Challenges

- Disparity between pre-trained visual encoder and language model in parameter scales (The parameter scale of widely-used visual encoders is usually less than 1 billion).
- The connection modules (VPGs) between visual encoder and language model are usually lightweight, which might be insufficient to capture the complicated cross-modal relations.
- As LLMs are large, fine-tuning the visual encoder or VPGs through LLMs is costly
- Both self-supervised visual encoder (DINO) and contrastive learning visual encoder have their shortcomings.
- Semantic misalignment between visual tokens and textual tokens: the textual tokens are a sequence of abstracted semantics with causal dependence, while the visual tokens might contain low-level visual details with a non-causal raster-scan order.

Requirements for Paper Presenters and Askers

- **Presenter:** The presentation of a sub-topic should cover (25 mins):
  - Objectives of papers
  - Clear literature reviews
  - Limitations, design/ implementation and results
  - Highlight key innovations, answer the how and why questions, such as How it works and Why it works
  - Future work.
  - Presenter Report: the presenter needs to submit a report within 2 weeks time (≤ 2 pages, Single-Spaced Times font 12)
- **Asker:**
  - You will need to pose 2-3 questions
  - Questions should have good depth and help to uncover insight of paper

Short Idea/ Opinion 1

- **Topic:**

  Can LFM (Large Foundation Model) use public data for training and content generation: what are the issues and guidelines?

- **Requirements for the Paper:**
  - The writeup should cover the background, issues, positions, analysis and insights. It must also contain a Section named **“Solution and Analysis”** that explains your proposed solution and your reasoning.
  - The article should be within **3 pages**, in ACM 2-column format (excluding references).
- **Grading Guidelines:**
  - I am looking for new angles into the issues, as well as innovative ideas, insights and solutions.
  - I will award a **B** if the paper covers most points above, and **A** for innovative ideas and insightful solution.
- **Deadlines:**
  - Article 1: 16 Feb @1700.

Requirements for Brave-New-Idea (BNI) papers

- **AIM of BNI Paper:**
  1) To propose a work that contain original ideas and research vision.
  2) The paper should offer: (i) novel, exploratory solutions with sufficient evidence of proof-of-concept; (ii) visions describing a new or open problem in multimedia research; and/or (iii) a novel perspective on existing multimedia research.
- **Guidelines:**
  - Must be in multimedia and is expected to have a high component of novelty
  - Should address an understudied, open problem in multimedia, while the ideas should be supported with sufficient scientific argumentation, experimentation and/or proof.
  - The paper should contain ideas not previously submitted nor published.
  - Should be within **5 pages**, excluding references, in ACM 2-column format.
- **Grading Criteria:**
  - Novelty; Conceptual leap; Depth of Impact; Breadth of impact
- **Deadlines:** 5 Apr (Fri) @1700

Papers for Lecture 5 (Few-Shot, Meta and Causal Learning)

P5-1: Few-show Learning: Presenter: Mehdi Yamini; Reader: Liu Nian

(Must-Read) X Liu, Y Zheng, Z Du, et al. GPT understands, too. AI Open 2023.

(Must-Read) O Vinyals, C Blundell, T. Lillicrap, K Kavukcouglu & D Wierstra. Matching Networks for One Shot Learning. NeurIPS 2016.

(To-Read) F Sung, Y Yang, L Zhang, T Xiang, P. Torr & T Hospedales. Learning to Compare: Relation Network for Few-Shot Learning. CVPR 2018.

P5-2: Meta Learning: Presenter: Bai Jinbin; Reader: Qin Hangyu

(Must-Read): J Snell, K Swersky & RS Zemel. Prototypical Networks for Few-Shot Learning. NeurIPS 2017.

(Must Read): C Finn, P Abbeel & S Levine: Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. ICML 2017.

P5-3: Causal Learning: Presenter: xx; Reader: Sui Yuan

(Must-Read) Y Niu, K Tang, H Zhang, et al. Counterfactual VQA: A Cause-Effect Look at Language Bias. CVPR 2021.

(To-Read) X Yang, H Zhang, G Qi & J Cai. Causal Attention for Vision-Language Tasks. CVPR 2021.