Multimodal Models: Part 2

Large Language Models: Introduction and Recent Advances

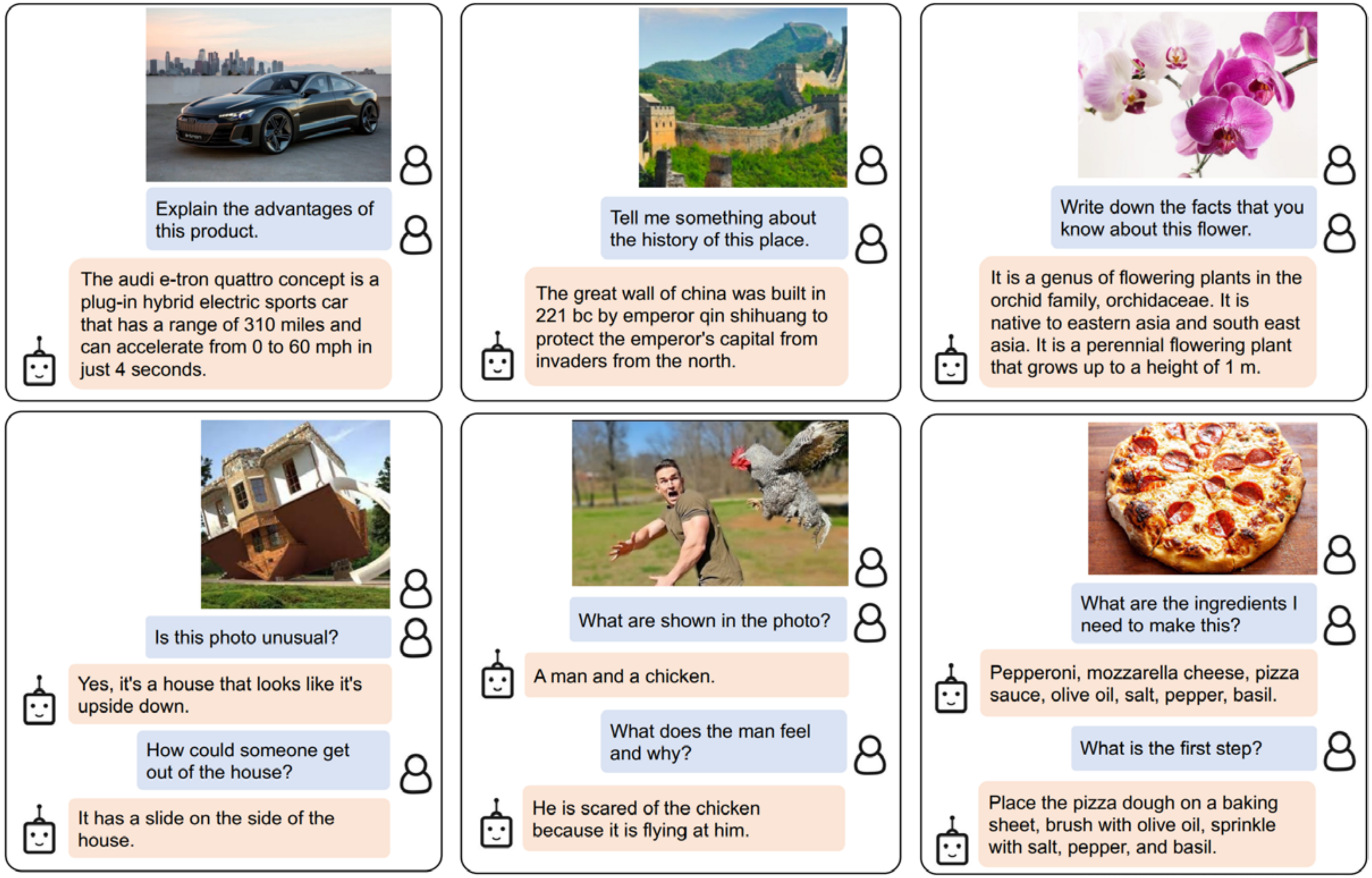
ELL881 · AIL821

Manish Gupta
Principal Applied Scientist, Microsoft
https://sites.google.com/view/manishg/

Agenda

- Multimodal Encoders
- **Text Generation with multimodal inputs**

Visual conversation



Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." arXiv:2301.12597 (2023).

Multimodal ChatBot

What are the things I should be cautious about when I visit here?

When visiting this peaceful location with a pier extending over a serene lake, surrounded by mountains, there are several things to be cautious about. First, pay attention to the weather conditions, as sudden changes in weather, especially strong winds or storms, can cause the water to become rough and pose risks to the stability of the pier. Second, be mindful of the structural integrity of the pier; avoid overloading it with too many people or excessive weight, as this could potentially damage the pier or cause it to collapse. Additionally, be aware of the surrounding wildlife and respect their space; do not harm or disturb them in any way. Finally, if you decide to go fishing, ensure that you adhere to any local fishing regulations and use appropriate fishing techniques to keep the lake ecosystem healthy and maintain the balance of the food chain.

Compose an engaging travel blog post about my trip to this place, highlighting the beautiful sceneries, joyful experiences, and include possible anecdotes.

Traveling to this picturesque location with a serene lake and mountains in the background was a refreshing and rejuvenating experience. The pier extending over the calm water provided an excellent vantage point from which to appreciate the breathtaking scenery. As I stood on the pier, I couldn't help but feel a deep connection with nature, taking a moment to soak in the peaceful atmosphere.

One of the highlights of my trip was undoubtedly the fishing experience. The area is teeming with life, and I was thrilled to land a sizeable catch, despite the modest size of the lake. The joy of fishing in such a beautiful setting was truly rewarding, and it allowed me to take a break from my busy life and truly immerse myself in the tranquility of nature.

While exploring the surrounding area, I came across a family of birds that had built a nest close to the pier, adding to the charm of the location. Interacting with the wildlife and observing their behavior was an enriching experience that I'll cherish for years to come.

In conclusion, this idyllic destination provided me with an opportunity to unwind and recharge, surrounded by mesmerizing natural beauty and the gentle sounds of water and wildlife. If you're looking for a serene escape from the hustle and bustle of everyday life, this pier overlooking the lake is the perfect place to find solace and reconnect with nature.

Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual Instruction Tuning." arXiv:2304.08485 (2023).

Scene Understanding

What could have happened based on the current scene?

Based on the current scene in the image, it is possible that a hurricane or severe weather event caused significant damage to the buildings and infrastructure in the area. The person walking through the debris-covered area suggests that they are inspecting the damage caused by the disaster. The presence of palm trees indicates that the location may be in a tropical or sub-tropical region, which can be more susceptible to hurricanes and other natural disasters.

Knowledge-grounded image description

Introduce me this painting in detail.

The painting depicts a young girl with a pearl earring, also known as "The Girl with the Pearl Earring" or "Girl with a Pearl Earring". She is wearing a blue dress and has a pearl earring on her left ear. Her hair is pulled back into a bun, and she is looking directly at the viewer. This painting is a famous work of art by Dutch painter Johannes Vermeer, created in 1665. It is considered one of his masterpieces and is widely recognized for its intricate details and realistic portrayal of the young girl's facial expression.

[Wenliang Dai and Junnan Li and Dongxu Li and Anthony Meng Huat Tiong and Junqi Zhao and Weisheng Wang and Boyang Li and Pascale Fung and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv 2305.06500 (2023)]

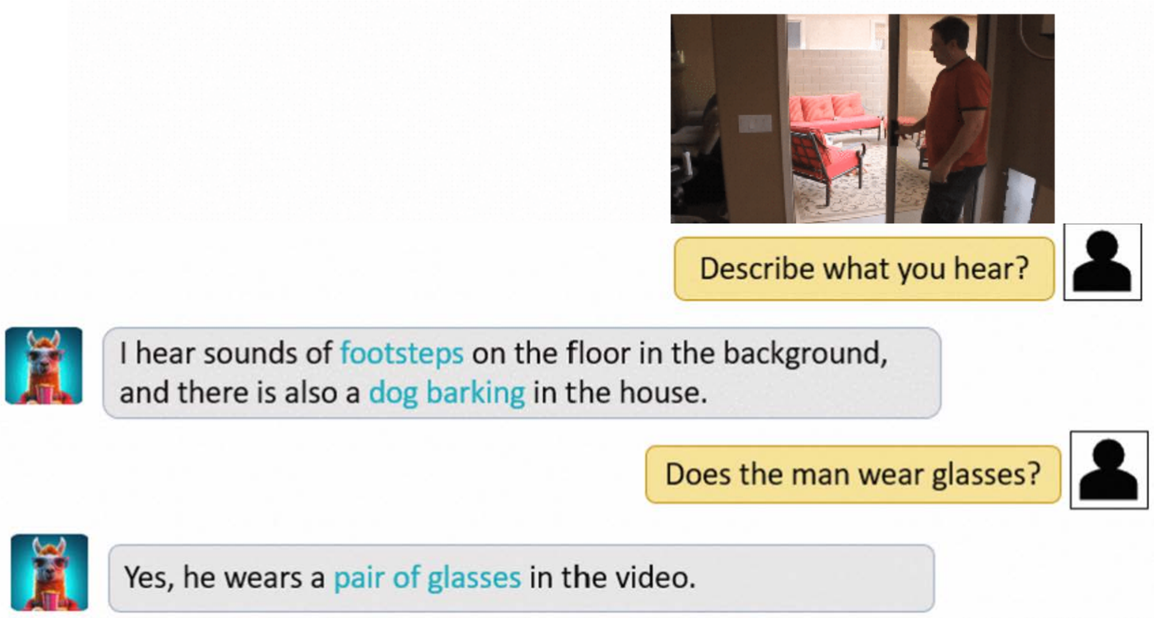
Visual Question Answering

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.
User	What is unusual about this image?
GPT-4 [32]	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.
User	Can you explain this meme in detail?
BLIP-2	a man is sitting on the back of a yellow cab
User	Can you explain this meme in detail?
OpenFlamingo	The man is drying his clothes on the hood of his car.

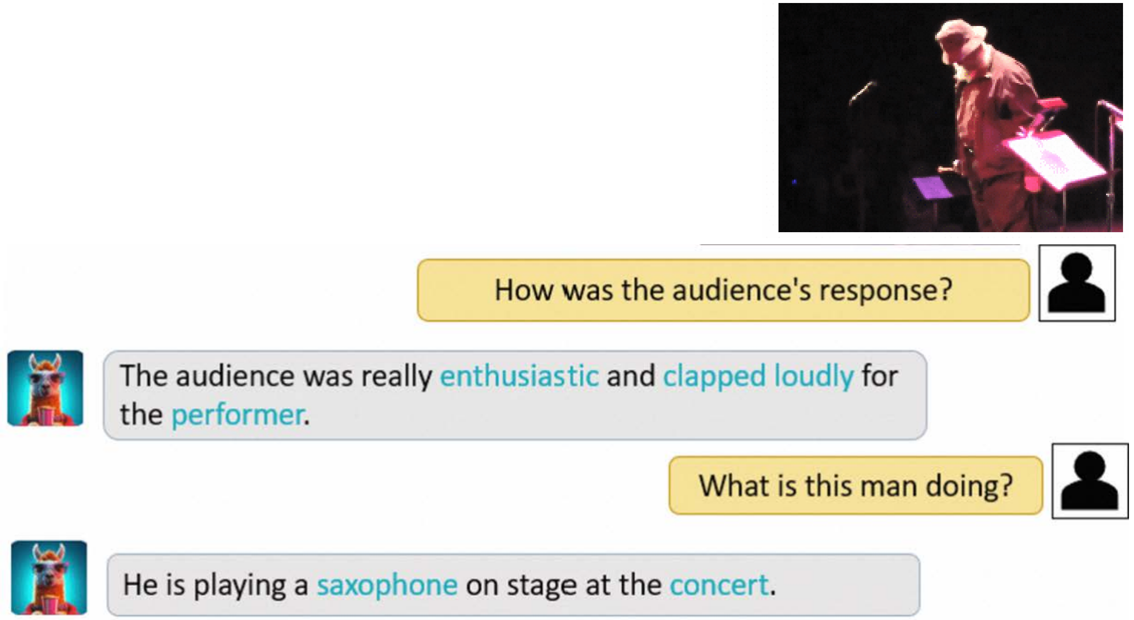
Visual input example, Chicken Nugget Map:	
	
User	Can you explain this meme in detail?
LLaVA	The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from 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.
User	Can you explain this meme?
GPT-4 [32]	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.
User	Can you explain this meme in detail?
BLIP-2	sometimes i just look at pictures of the earth from space and marvel how beautiful it is
User	Can you explain this meme in detail?
OpenFlamingo	It's a picture of a chicken nugget on the International Space Station.

Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual Instruction Tuning." arXiv:2304.08485 (2023).

Audio-visual integration perception ability



A case that Video-LLaMA answers the questions based on the background sound and visual content of the video



A case showing Video-LLaMA's ability to identify the sound of applause in a video and infer the positive response from the audience. Additionally, it infers that a man is playing the saxophone on stage based on the visual content.

Zhang, Hang, Xin Li, and Lidong Bing. "Video-llama: An instruction-tuned audio-visual language model for video understanding." arXiv preprint arXiv:2306.02858 (2023).

Common-knowledge concept recognition

Do you know who the man on the right is?

The man on the right is Jon Snow, played by Kit Harington in the television series Game of Thrones.

Who is the lady on the left?

The lady on the left is likely Daenerys Targaryen, played by Emilia Clarke in the television series Game of Thrones.

What is the relationship between them?

Jon Snow and Daenerys Targaryen are both characters in the television series Game of Thrones. In the show, they have a romantic relationship. They first meet in the seventh season, and their attraction to each other grows over the course of the eighth season. Despite their initial attraction, they have a complicated relationship due to their respective backgrounds and the challenges they face in the series

Do you know the building in this picture?

The building in the image appears to be the United States Capitol. It is a neoclassical style building located in Washington, D.C., the capital of the United States. It serves as the home of the U.S. Congress and is one of the most iconic landmarks in the country

Zhang, Hang, Xin Li, and Lidong Bing. "Video-llama: An instruction-tuned audio-visual language model for video understanding." arXiv preprint arXiv:2306.02858 (2023).

Capture temporal dynamics in videos

Describe this video.

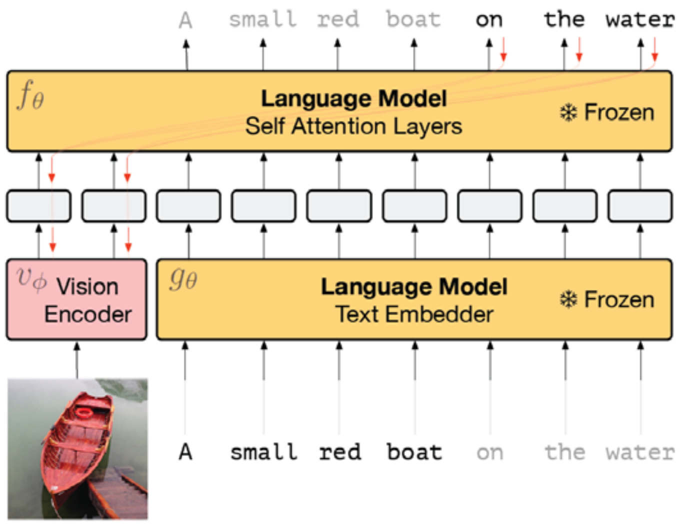
Summarize this video in one sentence.

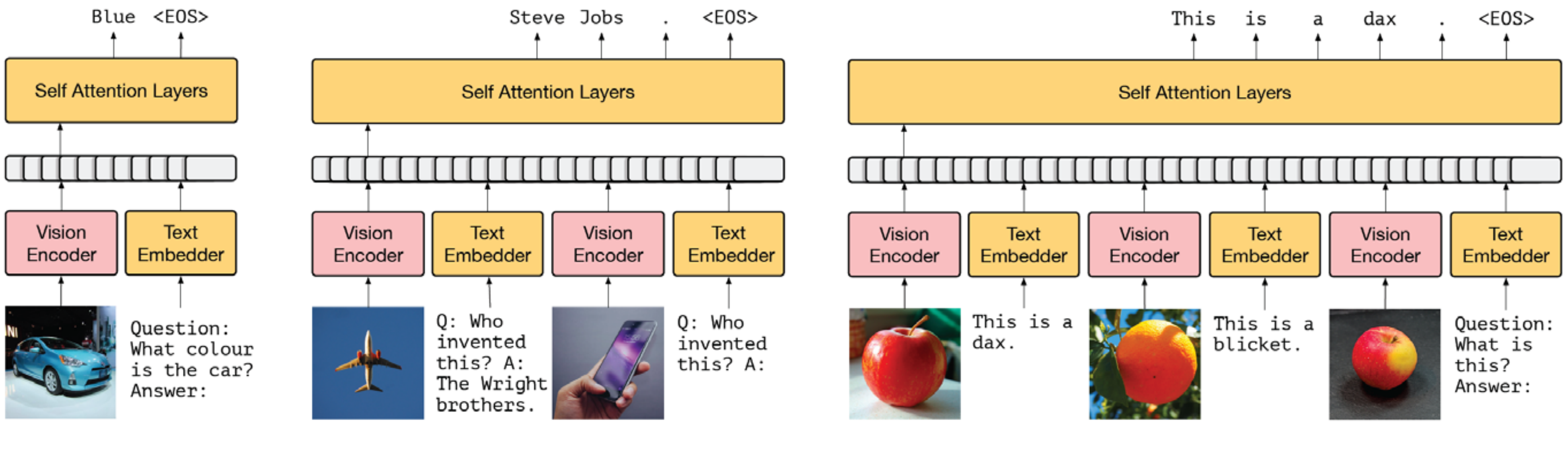
Zhang, Hang, Xin Li, and Lidong Bing. "Video-llama: An instruction-tuned audio-visual language model for video understanding." arXiv preprint arXiv:2306.02858 (2023).

Story and Song generation; comic understanding

Zhu, Deyao, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. "Minigpt-4: Enhancing vision-language understanding with advanced large language models." arXiv:2304.10592 (2023).

Frozen: Multimodal few-shot learning



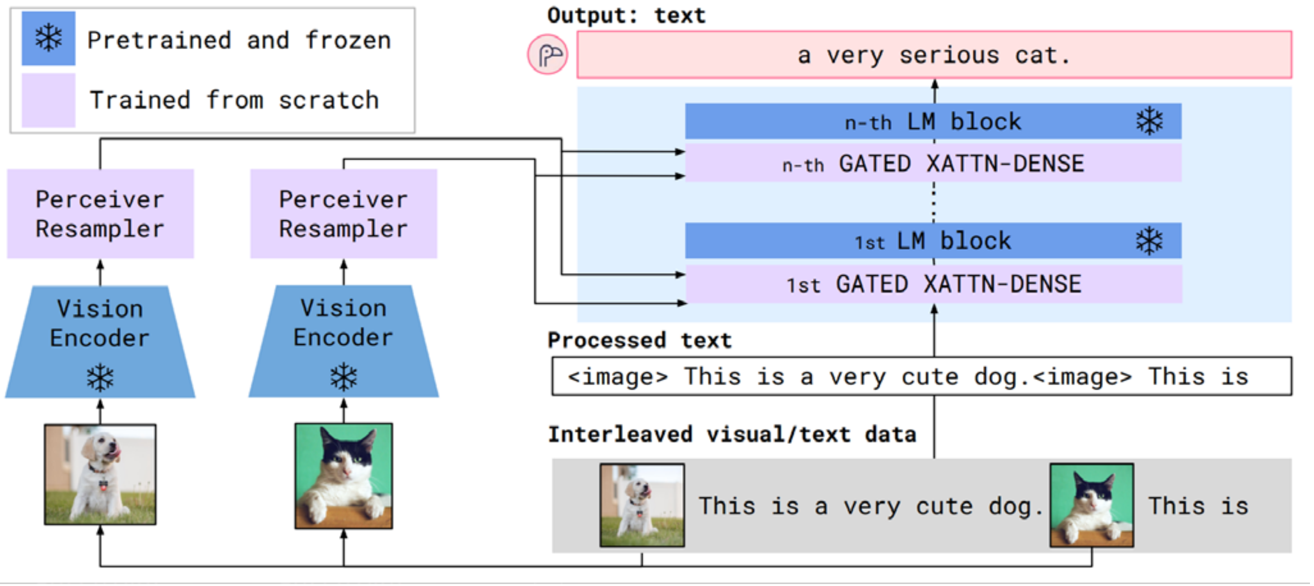


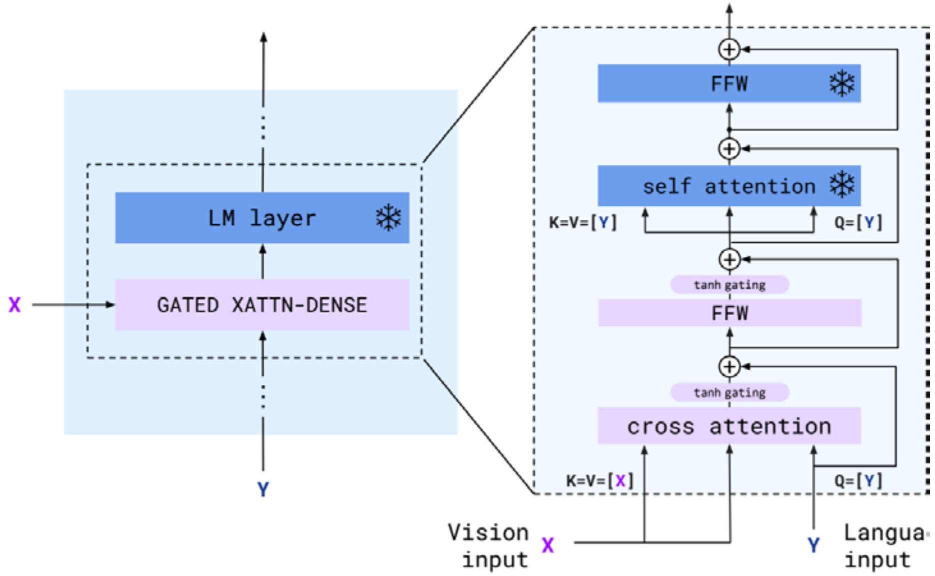
(a) **0-shot VQA** (b) **1-shot outside-knowledge VQA** (c) **Few-shot image classification**

Gradients through a frozen language model's self attention layers are used to train the vision encoder.

Inference-Time interface via in-context learning.

Flamingo



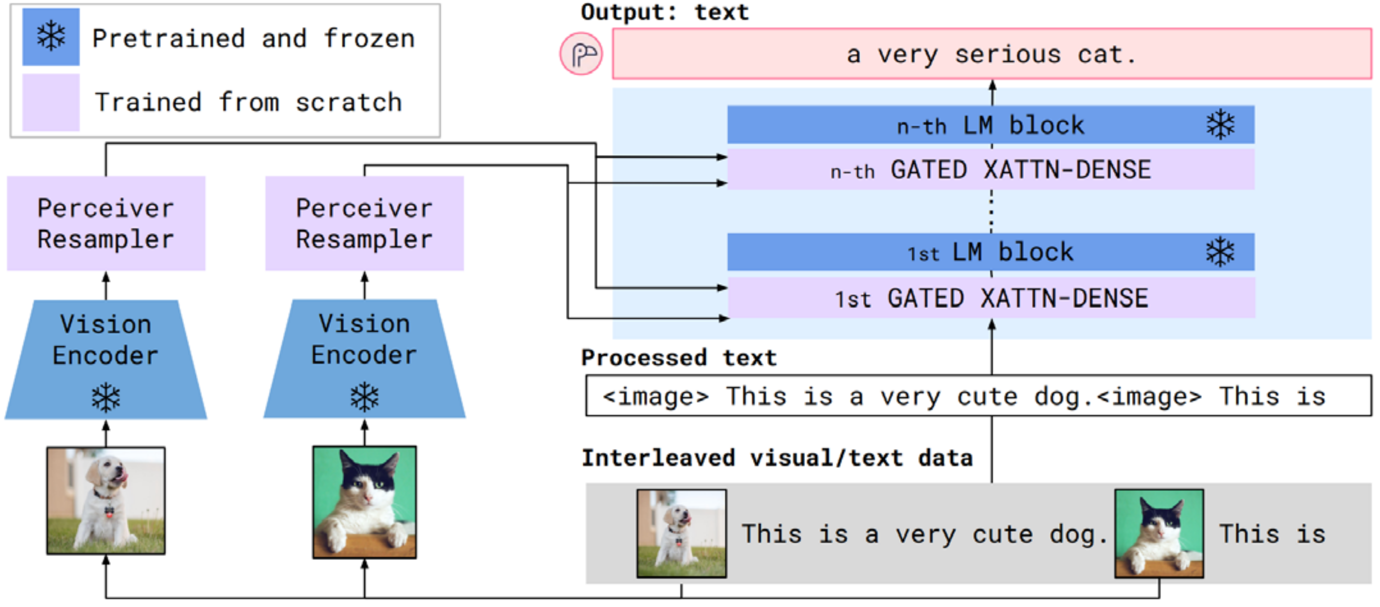


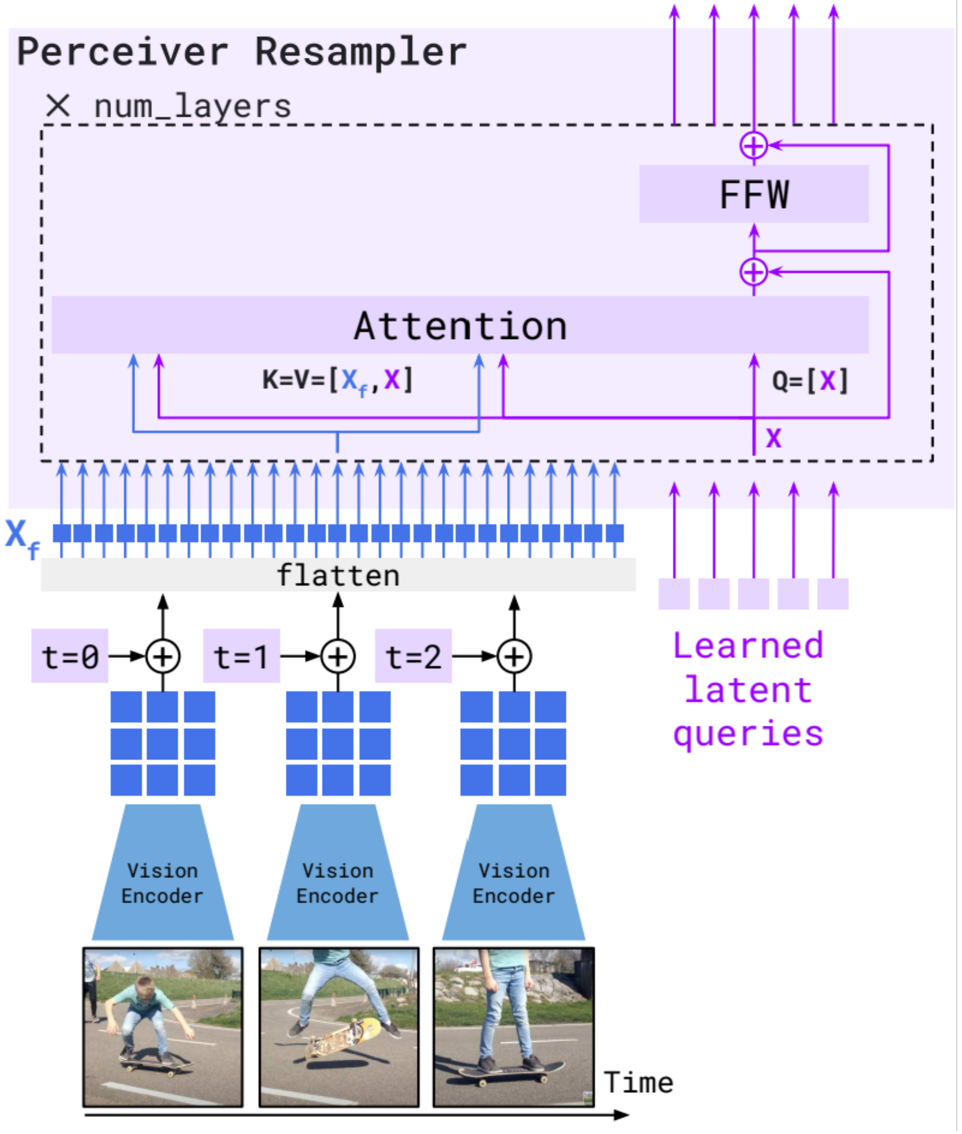
- Vision Encoder: Pretrained ResNet.
- Flamingo-3B, 9B and 80B.
- At a given text token, the model attends to the visual tokens of the image that appeared just before.

- Datasets:
  - MultiModal MassiveWeb (M3W): Interleaved image and text dataset.
  - Pairs of image/video and text: ALIGN dataset, LTIP (Long Text & Image Pairs), VTP (Video & Text Pairs)

Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc et al. "Flamingo: a visual language model for few-shot learning." NIPS 35 (2022): 23716-23736.

Flamingo

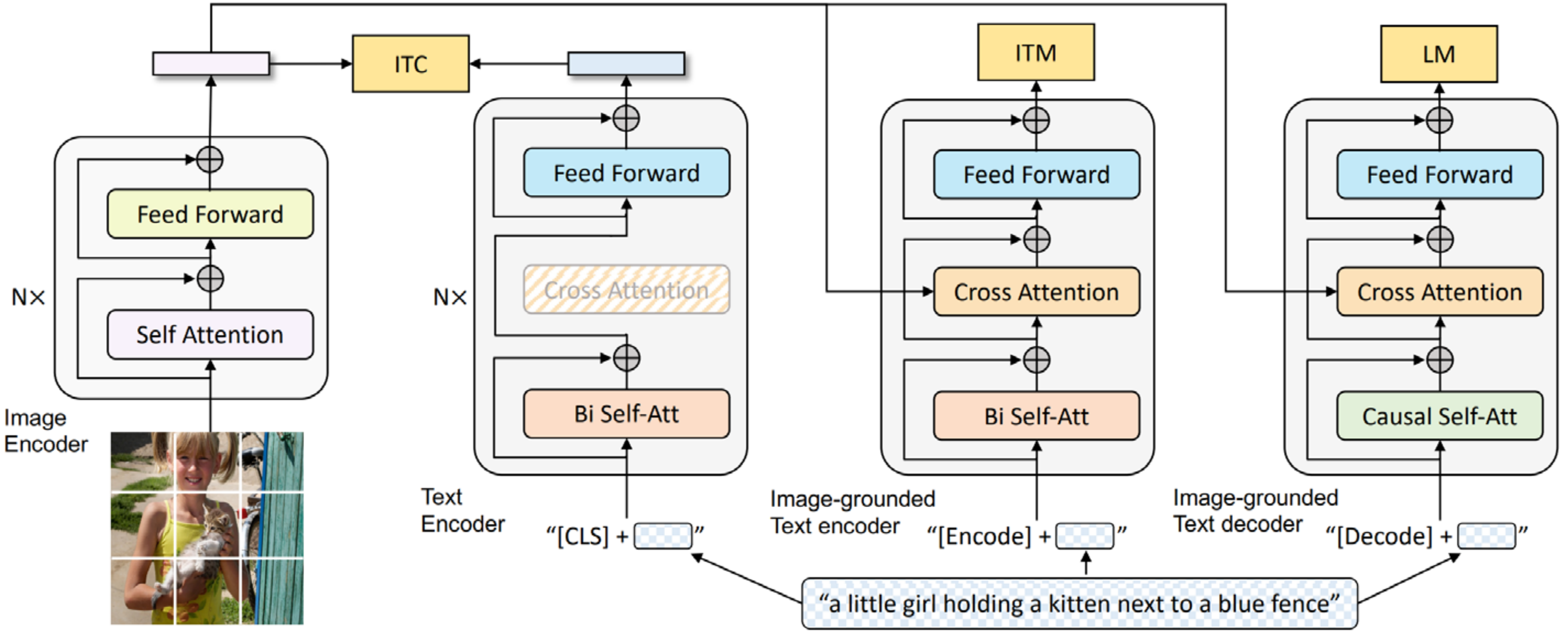




Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc et al. "Flamingo: a visual language model for few-shot learning." NIPS 35 (2022): 23716-23736.

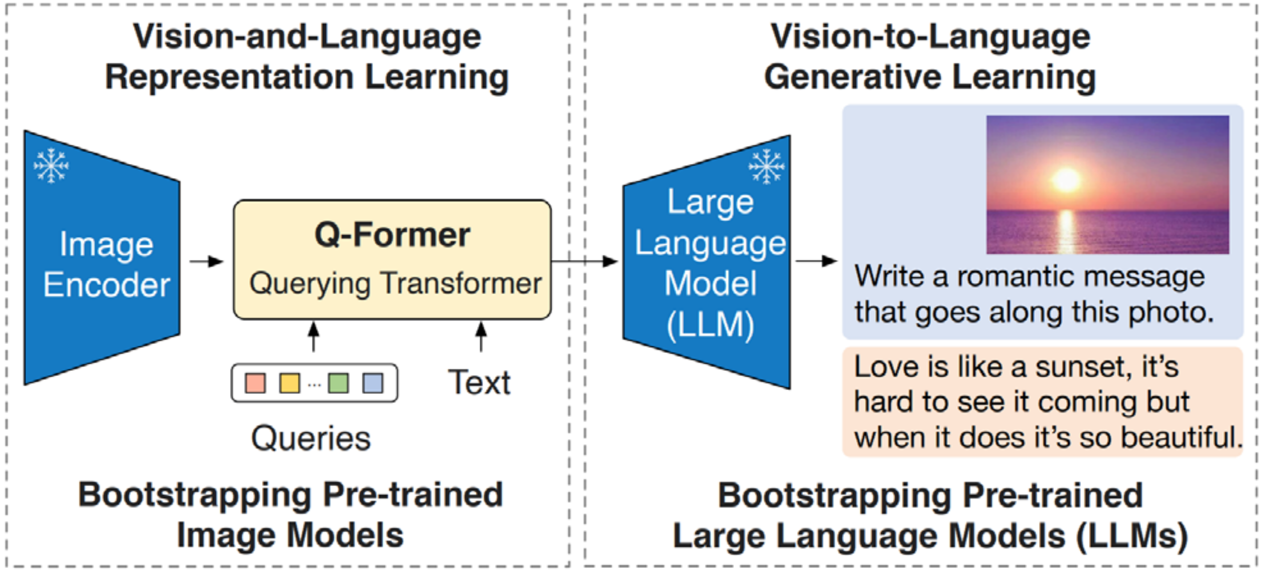
BLIP

- Vision-Language Pre-training
  - Models work for understanding or generation – not both.
  - Trained on small hand-labelled data like COCO, or on noisy web data like CC.
- Multimodal mixture of Encoder-Decoder (MED)
  - Unimodal encoder: ViT-B/16 and ViT-L/16; BERT-base
  - image-grounded text encoder
  - image-grounded text decoder



[Li, Junnan, Dongxu Li, Caiming Xiong, and Steven Hoi. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." In ICML, pp. 12888-12900. PMLR, 2022.]()

BLIP2

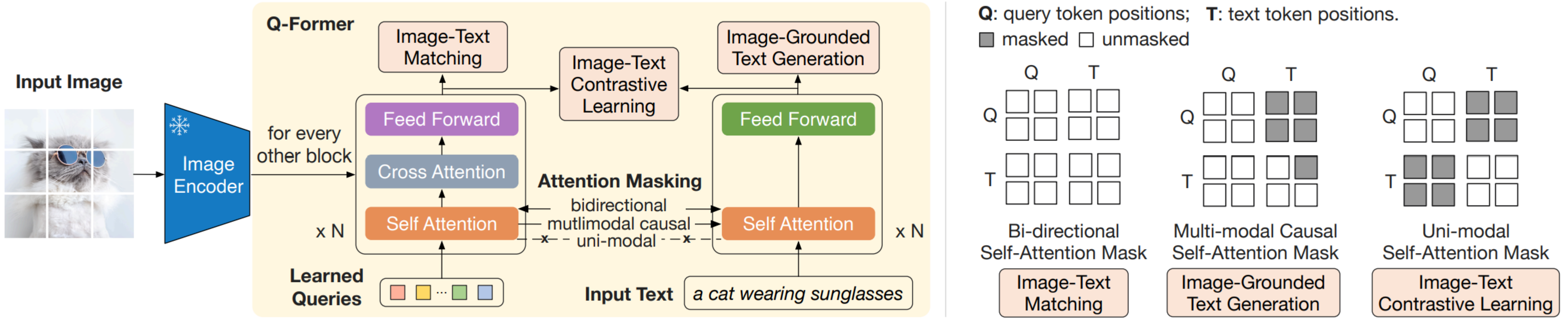


- Pre-trained in two stages:
  - Vision-language representation learning stage with a frozen image encoder
  - Vision-to-language generative learning stage with a frozen LLM.
- Q-Former=image transformer+ text transformer

[Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." arXiv:2301.12597 (2023).]

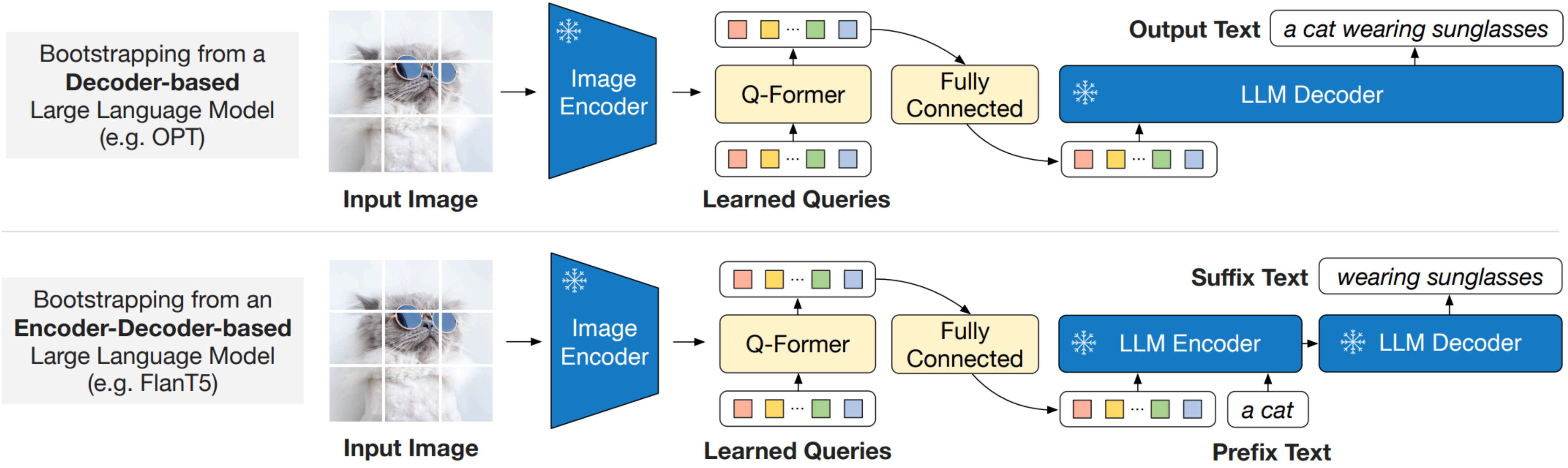
BLIP2

- Q-Former=image transformer+ text transformer
- Queries interact with
  - each other and optionally text through self-attention layers
  - frozen image features through cross-attention layers



Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." arXiv:2301.12597 (2023).

BLIP2

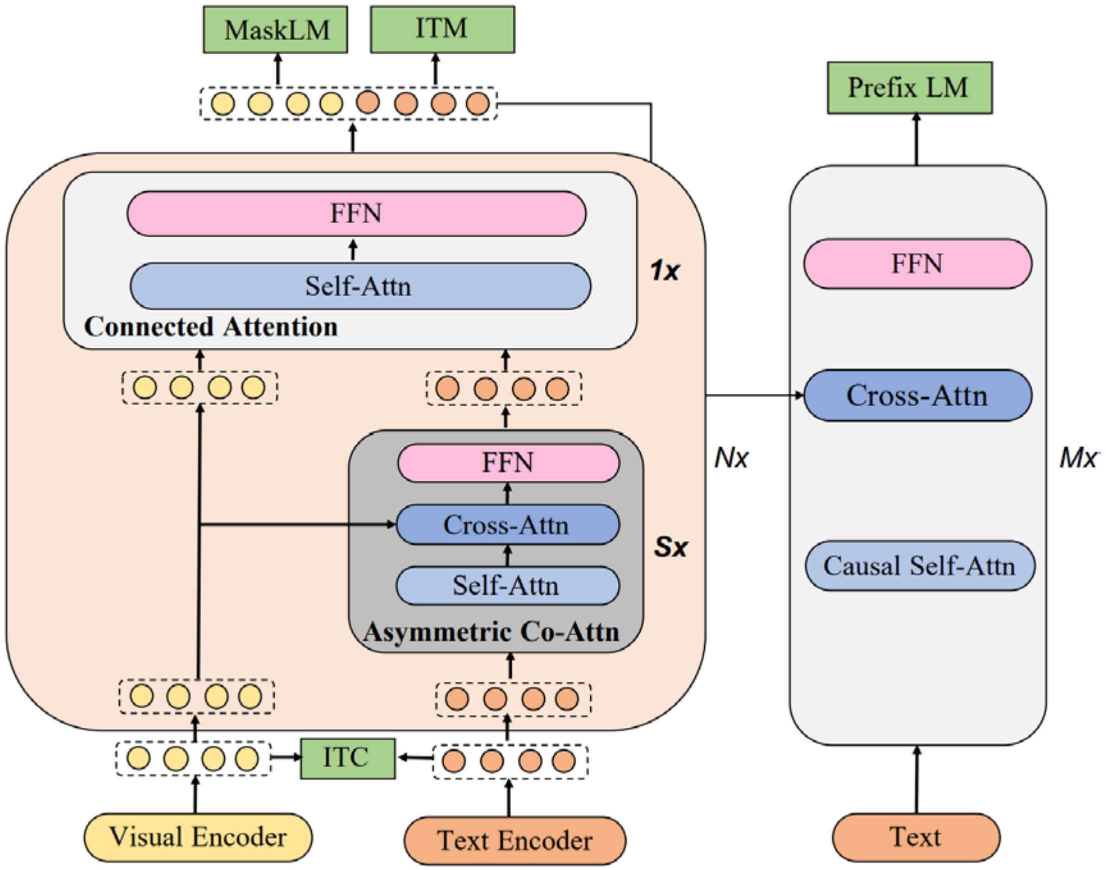


- Pre-training dataset
  - Same as BLIP
  - 129M images from COCO, Visual Genome, CC3M, CC12M, SBU
  - 115M images from LAION400M dataset.
- Pre-trained image encoder
  - ViT-L/14 from CLIP; ViT-g/14 from EVA-CLIP
- Frozen language model
  - OPT for decoder-based LLMs; FlanT5 for encoder-decoder-based LLMs.

Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." arXiv:2301.12597 (2023).

mPLUG

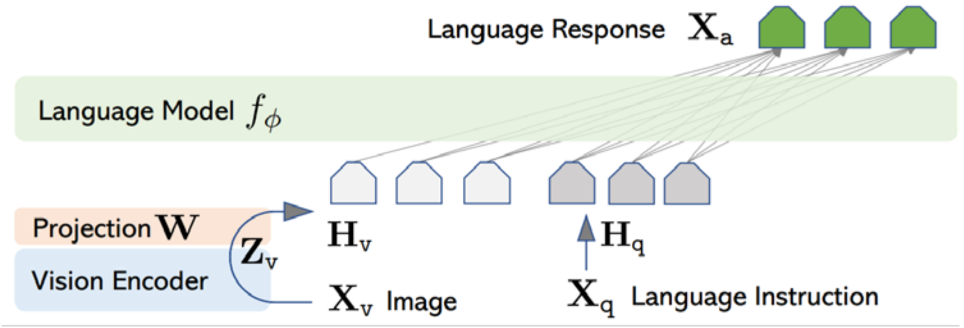
- Instruction-following image-LLMs: LLaVA, mPLUG and MiniGPT4
- Pre-trained end-to-end on largescale image-text pairs
  - 3 understanding tasks (Image-Text Contrastive Learning, Image-Text Matching, MLM)
  - 1 generation task (Prefix LM).
- Pre-training data
  - 14M images with texts from MS COCO, Visual Genome, Conceptual Captions, Conceptual 12M, SBU Captions.
- First 6L of BERTbase for text encoder
- Last 6L of BERTbase for cross-modal skip-connected network
- 12-layer Transformer for the decoder.
- CLIP-ViT for Visual encoder (ViT-B/16 or ViT-L/14)



[Li, Chenliang, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye et al. "mplug: Effective and efficient vision-language learning by cross-modal skip-connections." arXiv:2205.12005 (2022).]

LLaVa

- Use COCO image captioning data and language-only GPT-4.
- 158K language-image instruction-following samples: 58K conversations, 23K detailed description, 77k complex reasoning.
- All generated using few-shot prompts to GPT-4.
- Stage 1: Pre-training for Feature Alignment
  - Filter CC3M to 595K image-text pairs.
  - Pretrain using conv data.
  - Visual encoder and LLM weights frozen.
  - Define loss over answer tokens.
- Stage 2: Fine-tuning End-to-End.
  - Keep the visual encoder weights frozen.
  - Update both the pre-trained weights of the projection layer and LLM.
  - Multimodal Chatbot: Finetune on all 158K



$$
\begin{aligned}
&\mathbf{X}_{\texttt{system-message}}\ \texttt{<STOP>}\ \backslash\texttt{n}\\
&\texttt{Human : }\mathbf{X}^1_{\texttt{instruct}}\ \texttt{<STOP>}\ \backslash\texttt{n Assistant: }\mathbf{X}^1_a\ \texttt{<STOP>}\ \backslash\texttt{n}\\
&\texttt{Human : }\mathbf{X}^2_{\texttt{instruct}}\ \texttt{<STOP>}\ \backslash\texttt{n Assistant: }\mathbf{X}^2_a\ \texttt{<STOP>}\ \backslash\texttt{n}\cdots
\end{aligned}
$$

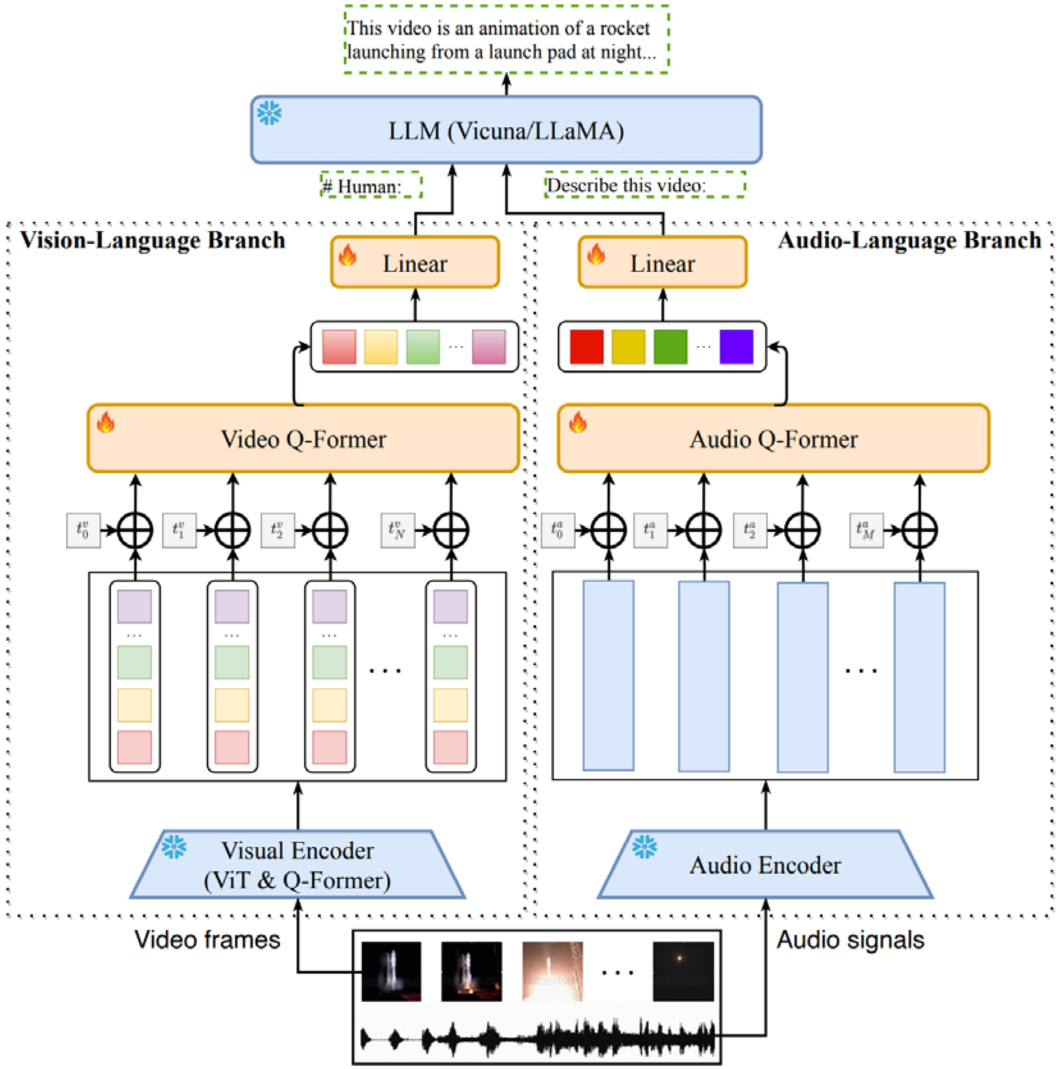
$$
\mathbf{X}^t_{\texttt{instruct}} = \begin{cases} \text{Random choose } [\mathbf{X}^1_q, \mathbf{X}_v] \text{ or } [\mathbf{X}_v, \mathbf{X}^1_q], & \text{the first turn } t = 1 \\ \mathbf{X}^t_q, & \text{the remaining turns } t > 1 \end{cases}
$$

- LLM=LLaMa
- Pre-trained CLIP visual encoder ViT-L/14
- 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.

Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual Instruction Tuning." arXiv:2304.08485 (2023).

Video-LLaMA

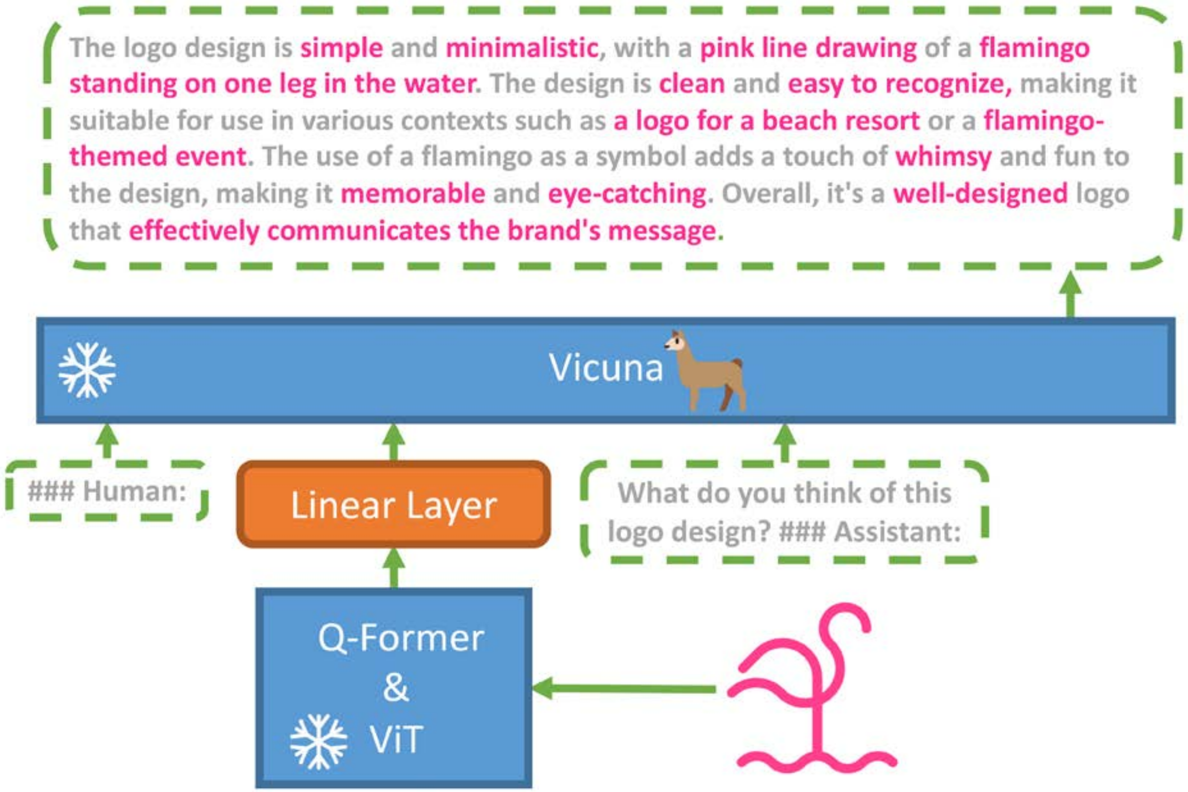
- Video Q-former: aggregates frame-level representations
  - shares same architecture with Query Transformer (Q-Former) in BLIP-2
- Video and audio soft prompts
- Audio encoder: Audio spectrogram Transformer from ImageBind.
- Multi-branch Cross-Modal Training
  - Vision-language:
    - Pre-train on (a) video caption dataset, Webvid-2M, with a video-clips-to-text generation task (b) image-caption data, CC595k.
    - Fine-tune on a video-based conversation dataset to execute visual instruction tuning.
      - image-detail-description dataset from MiniGPT-4
      - image-instruction dataset from LLaVA
      - video-instruction dataset from Video-Chat.
  - Audio-language:
    - Pre-train audio-related components on an audio caption dataset.



Zhang, Hang, Xin Li, and Lidong Bing. "Video-llama: An instruction-tuned audio-visual language model for video understanding." arXiv:2306.02858 (2023).

MiniGPT-4

- Aligns a frozen visual encoder with a frozen advanced LLM, Vicuna, using one projection layer.
- Stage 1 (pretraining)
  - Vision encoder and LLM are frozen
  - Uses a combined image captioning dataset that includes images from LAION, Conceptual Captions, and SBU.
- Stage 2 (finetuning)
  - 3,500 detailed image description pairs
  - ###Human: <Img><ImageFeature></Img><Instruction>###Assistant:
  - <Instruction> could be “Describe this image in detail” or “Could you describe the contents of this image for me”.



Zhu, Deyao, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. "Minigpt-4: Enhancing vision-language understanding with advanced large language models." arXiv:2304.10592 (2023).

Meeting notes with multiple speakers (speaker diarization)

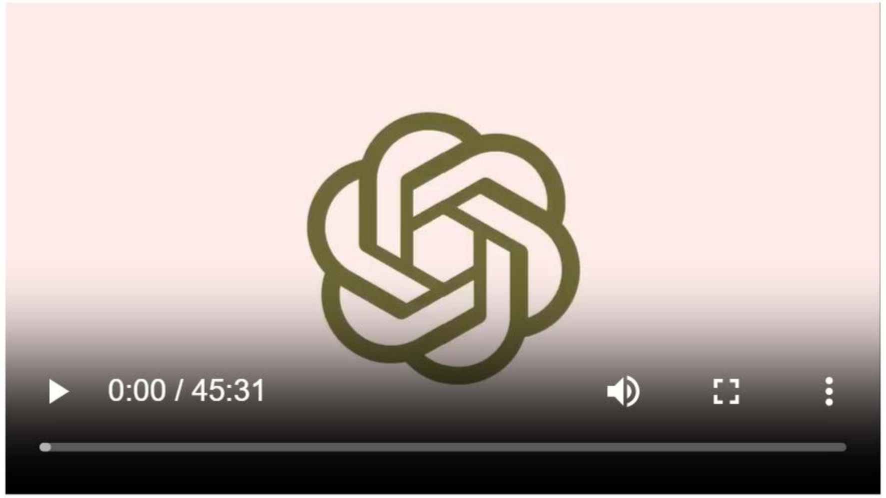
- Input: How many speakers in this audio and what happened?
- Output
  - There are four speakers in the audio. It sounds like a project meeting where the project manager, Mark, is introducing himself and asking the team members to introduce themselves and state their roles. After Mark introduces himself, three additional speakers introduce their roles: Dirk as the marketing expert, Nick as the user interface designer, and Xavier as the industrial designer. They briefly mention their responsibilities for the project.

- Input
  - Can you transcribe it with speaker names?
- Output
  - **Mark:** Okay, good morning. This is our first team meeting.
  - **Team:** Morning.
  - **Mark:** I will be your project manager for today, for this project. My name is Mark and I'll be giving this presentation for you to kick the project off. This is the agenda for today. Well, of course, we're new to each other, so I'd like to get acquainted first. So, let's do that first. I mean, let's start with you. Can you introduce yourself? You are a marketing expert?
  - **Dirk:** Yes. Uh, my name is Dirk, Dirk Meinfelt. Uh, I will be, uh, during this project, the marketing expert. And I will see what, uh, the user wants and, uh, what we can do, uh, with the new product, project, product.
  - **Mark:** Excellent. And you are?
  - **Nick:** Nick de Boer, user interface designer. I'm going to, uh, look at the technical design from the user point of view.
  - **Mark:** Okay.
  - **Xavier**: My name is Xavier Joerkens. I'm the industrial designer, and there are three main questions that I have to find an answer to.

https://openai.com/index/hello-gpt-4o/

Lecture summarization

- Input
  - Here is a video of a presentation on techniques for maximizing LLM performance. Could you give me a detailed summary of the presentation?



https://openai.com/index/hello-gpt-4o/

Output

Certainly! The presentation focused on techniques for maximizing the performance of large language models (LLMs) and was delivered by Colin Jarvis and John Allard from OpenAI. The session was part of OpenAI's first developer conference and aimed to provide insights into optimizing LLMs for specific tasks.

Introduction

- **Colin Jarvis** is the head of OpenAI's solutions practice in Europe, and **John Allard** is an engineering lead on the fine-tuning product team.
- The presentation highlighted the challenges of optimizing LLMs, emphasizing that there is no one-size-fits-all solution.

Challenges of Optimizing LLMs

1. **Separating Signal from Noise**: It's challenging to identify the exact problem.
2. **Abstract Performance Metrics**: Measuring performance can be difficult.
3. **Choosing the Right Optimization**: It's hard to know which approach to use.

Optimization Flow

- The presenters introduced a framework for optimizing LLMs based on two axes:
  - **Context Optimization**: What the model needs to know.
  - **LLM Optimization**: How the model needs to act.
- The framework includes four quadrants:
  - **Prompt Engineering**: The starting point for optimization.
  - **Retrieval-Augmented Generation (RAG)**: For context optimization.
  - **Fine-Tuning**: For LLM optimization.
  - **All of the Above**: Combining all techniques.

Lecture summarization

Prompt Engineering

- **Strategies:**
  - Write clear instructions.
  - Split complex tasks into simpler subtasks.
  - Give the model time to think.
  - Test changes systematically.
- **Good for:**
  - Testing and learning early.
  - Setting a baseline.
- **Not good for:**
  - Introducing new information.
  - Replicating complex styles.
  - Minimizing token usage.

Retrieval-Augmented Generation (RAG)

- **Overview:**
  - RAG involves retrieving relevant documents and using them to generate responses.
- **Good for:**
  - Introducing new information.
  - Reducing hallucinations.
- **Not good for:**
  - Embedding broad domain knowledge.
  - Teaching new formats or styles.
  - Minimizing token usage.
- **Success Story:**
  - The presenters shared a success story where they improved accuracy from 45% to 98% using RAG.

Fine-Tuning

- **Overview:**
  - Fine-tuning involves continuing the training process on a smaller, domain-specific dataset.
- **Benefits:**
  - Improves performance on specific tasks.
  - Improves efficiency.
- **Good for:**
  - Emphasizing existing knowledge.
  - Customizing structure or tone.
  - Teaching complex instructions.
- **Not good for:**
  - Adding new knowledge.
  - Quick iteration.
- **Success Story:**
  - The presenters shared a success story from Canva, where fine-tuning improved performance significantly.

Best Practices

1. **Start with Prompt Engineering and Few-Shot Learning**.
2. **Establish a Baseline**.
3. **Start Small and Focus on Quality**.

Combining Fine-Tuning and RAG

- The presenters highlighted the benefits of combining fine-tuning and RAG for optimal performance.

Application of Theory

- The presenters applied the theory to a practical challenge, the Spider 1.0 benchmark, achieving high accuracy using both RAG and fine-tuning.

Conclusion

- The presentation concluded with a summary of the optimization flow and emphasized the importance of iteratively improving LLM performance using the discussed techniques.

https://openai.com/index/hello-gpt-4o/

Summary

- Frozen
- Flamingo
- BLIP
- BLIP2
- mPLUG
- LLaVa
- Video-LLaMA
- MiniGPT-4

Thanks!

- HomePage: https://sites.google.com/view/manishg/
- Google Scholar: https://scholar.google.co.in/citations?user=eX9PSu0AAAAJ
- LinkedIn: http://aka.ms/manishgupta
- YouTube (Data Science Gems): https://www.youtube.com/@dlByManish