

# Text Generation with Multimodal Inputs

Advanced Large Language Models

ELL8299 · AIL861 · ELL881




Manish Gupta

Principal Applied Scientist, Microsoft

<https://sites.google.com/view/manishg/>


g/

# Visual conversation




Explain the advantages of this product.

The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.




Tell me something about the history of this place.

The great wall of china was built in 221 bc by emperor qin shihuang to protect the emperor's capital from invaders from the north.



Write down the facts that you know about this flower.

It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.




Is this photo unusual?

Yes, it's a house that looks like it's upside down.

How could someone get out of the house?

It has a slide on the side of the house.




What are shown in the photo?

A man and a chicken.

What does the man feel and why?

He is scared of the chicken because it is flying at him.



What are the ingredients I need to make this?

Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.

What is the first step?

Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.

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 \(2023\).](#)





# Knowledge grounded image desc

## 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.



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 Jungi 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.
User	[Start a new conversation, and clear the history]
LLaVA	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:

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.




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




Describe what you hear?

I hear sounds of **footsteps** on the floor in the background, and there is also a **dog barking** in the house.

Does the man wear glasses?

Yes, he wears a **pair of glasses** in the video.



How was the audience's response?

The audience was really **enthusiastic** and **clapped loudly** for the **performer**.

What is this man doing?

He is playing a **saxophone** on stage at the **concert**.

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).



Advanced LLMs



Manish Gupta

# Story and Song generation; comic understanding



Can you write a story around this image for my child?



Sure, here's a story based on the image:

Once upon a time, there was a brown bear, a white rabbit, and a black cat who lived in a forest. They were good friends and loved to spend time together. One day, they decided to have a picnic under a tree. The bear brought some honey, the rabbit brought some carrots, and the cat brought some milk.

As they sat around the table, the bear said, "I'm so hungry, I could eat a whole honey pot!" The rabbit giggled and said, "I'm so hungry, I could eat a whole bag of carrots!" The cat meowed and said, "I'm so hungry, I could drink a whole bowl of milk!"

They all laughed and dug in to their food. The bear loved the honey, the rabbit loved the carrots, and the cat loved the milk. They were all very happy and content.

After they finished their meal, they



Can you write me a master rap song that rhymes very well based on this image?



Sure, here's a rap song based on the image you provided:

Verse 1:  
I'm an old lady, but I'm still flyin'  
I've got my headphones on and I'm feelin' high  
I'm in the studio, makin' some beats  
I'm the queen of hip hop, can't nobody touch my feet

Chorus:  
I'm the grandma of rap, I'm still spittin' fire  
I've got my green jacket on, I'm lookin' fly  
I've got my sunglasses on, I'm shining bright  
I'm the queen of the game, I'm makin' it right

Verse 2:  
I've been in the game for years, I've paid my dues  
I've got a microphone in my hand and I'm singin' the blues  
I've got the crowd in the palm of my hand  
I'm the queen of rap, I'm takin' over the land



Who are the people in the image?



The people in the image are Goku and Vegeta, two characters from the Dragon Ball series. Goku is a saian warrior and the main protagonist of the series, while Vegeta is a saian prince and the main antagonist. They are both known for their powerful abilities and their rivalry.

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).





# Text extraction and detailed reasoning based on image

## Prompt

Please identify the text in the picture.

### Unesco announces its newest geoparks around the world

9 April 2024

By Lynn Brown, Features correspondent

Share



Unesco Geoparks represent a balance of unique geological features, cultural touchpoints and a focus on sustainability (Credit: Chris Karagelis)

From dinosaur fossils in Brazil to the soaring monasteries of Meteora, these 18 new geological sites highlight a particular region's natural, cultural and intangible heritage.

Travellers interested in deep dives into geology, culture and sustainability have several new destinations to place on their bucket lists. Unesco just announced the designation of 18 new **Global Geoparks**, bringing the total to 213 sites representing 48 countries around the world.

From crystal-clear karst lakes to an archipelago formed during the last ice age, these parks are more than just sites with unique geological features; they also serve to highlight a particular region's natural, cultural and intangible heritage. "Unesco Global Geoparks represent landscapes where internationally significant geological heritage intertwines with human history and supports rich biodiversity," explains António Abreu, director of the Division of Ecological and Earth Sciences at Unesco. "Each new geopark tells a unique story, using geology as its lens."

Here are five of the newest geoparks on Unesco's list:

## MiniCPM-Llama3-V 2.5

**Title:** Unesco announces its newest geoparks around the world

**Date and Author:** 9 April 2024

By Lynn Brown, Features correspondent

**Image Caption:** Unesco Geoparks represent a balance of unique geological features, cultural touchpoints and a focus on sustainability (Credit: Chris Karagelis)

**Body Text:** From dinosaur fossils in Brazil to the soaring meteors of Meteora, these 18 new geological sites highlight a particular region's natural, cultural and intangible heritage.

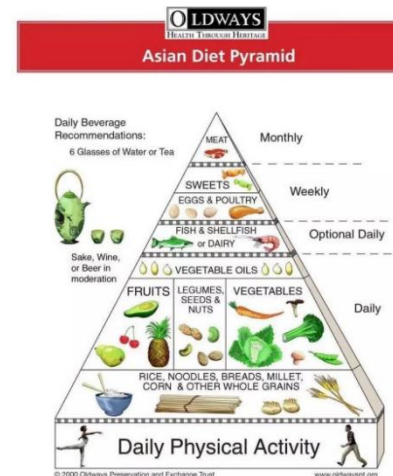
Travellers interested in deep dives into geology, culture and sustainability have several new destinations to place on their bucket list. Unesco just announced the designation of 18 new Global Geoparks, bringing the total to 213 sites representing 48 countries around the world.

From crystal-clear karst lakes to an archipelago formed during the last ice age, these parks are more than just unique geological features; they also serve to highlight a particular region's natural, cultural and intangible heritage. Unesco Global Geoparks represent landscapes where internationally significant geological heritage intertwines with human history and supports rich biodiversity, explains António Abreu, director of the Division of Ecological Sciences at Unesco. "Each new geopark tells a unique story, using geology as its lens."

Here are five of the newest geoparks on Unesco's list:

## Prompt

Based on this picture, make a detailed meal plan for me.



## MiniCPM-Llama3-V 2.5

To create a detailed meal plan based on the Asian Diet Pyramid, one would start with the base of the pyramid, which emphasizes daily physical activity as the foundation of a healthy lifestyle. Then, for each food group listed in descending order from the top to the bottom, an individual could plan meals that include the recommended items. For example, for the 'MEAT' group, one might have fish or chicken dishes for protein. The 'SWEETS' and 'EGGS & POULTRY' groups suggest including eggs or poultry like chicken or turkey. The 'FISH & SHELLFISH' group could feature seafood such as salmon or shrimp. The 'SWEETS' group could include fruits like apples or bananas. The 'VEGETABLE OILS' section suggests incorporating vegetable oils like olive or canola oil into cooking. The 'LEGUMES, SEEDS & NUTS' group might include beans, lentils, or nuts. The 'VEGETABLES' group is broad and could include a variety of vegetables such as broccoli, carrots, or bell peppers. Lastly, the 'RICE, NOODLES, BREADS, MILLET, CORN & OTHER WHOLE GRAINS' group would be the staple carbohydrates in meals. This meal plan should be tailored to an individual's dietary needs and preferences while following the principles outlined in the Asian Diet Pyramid.

[Yao, Yuan, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai et al. "Minicpm-v: A gpt-4v level mllm on your phone." arXiv:2408.01800 \(2024\).](#)



Advanced LLMs



Manish Gupta



# Understanding images with extreme aspect ratio


BB

C

What will we eat on the Moon? The food is literally out of this world

2 hours ago  
By Laura Hall

Share



Good food is crucial for a successful deep space mission (Credit: Getty Images)

The Moon may be the final frontier for mankind, but what will we eat when we get there? Pasta and protein bars made out of thin air are just the beginning.


Space fever is approaching at warp speed. In the next two years, Nasa plans to send astronauts back to the Moon via its **Artemis programme**; the International Space Station (ISS), designed to orbit for 15 years but now hitting its 26th year in space, will soon be **replaced**; and scientists are looking seriously into the possibility of manned deep space missions. Add to that a proliferation of tourism projects rocketing deep-pocketed individuals up to the edge of space and it begs one question for a food writer like me: what will we eat when we get there?

"Food is something that keeps astronauts sane," says Dr Sonja Brungs, astronaut operations deputy lead at the **European Space Agency**. "Good food, proper food with a lot of variety, tailored to the needs of the individual astronauts is crucial for a successful deep space mission. I think people underestimate how important it is."

Currently, astronauts are given small food pouches containing prepared meals. These

meals are made by specialised food-production companies and then freeze-dried, dehydrated or thermostabilised. Astronauts add water to heat or cool the meals to eat; they can also bring along a special meal that reminds them of home (this too has to be carefully formulated and thermostabilised).

There are some no-gos: anything that crumbs, like bread, can't be taken into space as the crumbs can easily become airborne in the low-gravity environment, meaning they could be inhaled or get into vital equipment. Salt is limited, due to the fact that the body stores sodium differently in space, **leading to accelerated osteoporosis**, and alcohol is also not permitted as it affects the waste water recycling system in the ISS.



ESA astronaut Andreas Mogensen holds chocolate mousse on his last trip (Credit: ESA/Nasa)

"Novelty is definitely an issue," says Brungs. "Astronauts who are in space for just six months miss crunchiness and texture. It is really important for mental wellbeing to have a variety of textures, and especially for deep space missions, having a variety of foods to eat."


In 2021, Nasa launched a **Deep Space Food Challenge** to discover new ways to create food in space with limited resources producing minimal waste, while also providing safe, nutritious and tasty food that can perform on a long-term deep space mission.

**"We make food out of thin air, quite literally - Artuu Luukanen"**

Solar Foods, based in Helsinki, is one of the eight companies that has reached the challenge's final phase. Their remarkable concept: using space waste to create protein.

"We make food out of thin air, quite literally," says Artuu Luukanen, Solar Foods' senior vice president in Space and Defence. His company discovered an edible microbe in the Finnish countryside that grows by feeding on a mixture of carbon dioxide, hydrogen and oxygen. The result is a source of protein from bacteria. The protein can be blended with a range of flavours or textures to create various kinds of nutritious food, such as pasta, protein bars, alternative meats and even an egg replacement

"We started thinking about space food because in any space habitat, you have two key waste gases available: hydrogen and carbon dioxide," Luukanen said. "So what we are talking about here is really not just a food manufacturing technology for space, but something that will be an integral part of the environmental control and life support system."



The ISS has its own tiny vegetable garden on board where astronauts study plant growth in microgravity (Credit: Nasa/Amanda Griffin)

Solar Foods' protein can be turned into a paste or powder and blended with flour and more typical food ingredients to create protein enriched foods such as pasta, protein bars and even chocolate. Experiments are continuing to discover whether it can be mixed with oils and **turned into something with a texture of a steak, using a 3D printer**.

Fresh food is also a consideration: while vitamin tablets can help, astronauts need fresh produce, **and experiments continue into how to grow vegetables in this unique zero-gravity, zero-sunlight environment**. The ISS has its own tiny vegetable garden on board, known as Veggie, where **astronauts study plant growth in microgravity**.

Please describe the content in the picture.

MiniCPM-Llama3-V 2.5

The image is a screenshot of an article from **the BBC website**, titled **"What will we eat on the Moon? The food is literally out of this world."** The article is authored by Laura Hall and was published 2 hours ago at the time of the screenshot. The main visual is a photograph of an astronaut, identified as **Andreas Mogensen, who is smiling and holding a chocolate mousse on a spoon**. He is wearing a blue shirt with the ESA (European Space Agency) logo and a name tag. The background of the photo shows the interior of a spacecraft or space station. The text of the article discusses the challenges of providing food in space, mentioning the ISS (International Space Station) and the Artemis program. It also touches on the nutritional aspects of food in space, **the use of 3D printers for food production, and experiments with growing plants in microgravity**. The article includes a **quote from Luukanen** about the challenges of food production in space and mentions the ISS's vegetable garden. The overall tone of the article is informative, with a focus on the technological and scientific aspects of space food.

*\* The colored underlines are for illustration purposes only and are not used as model input.*

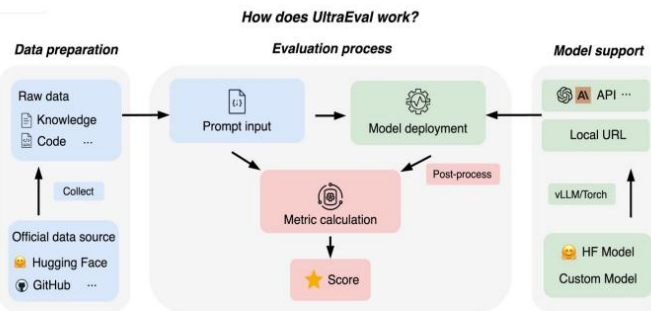
Advanced LLMs

Manish Gupta

# Flowchart understanding and reviewing interior designs

## Prompt

How does it work?



## MiniCPM-Llama3-V 2.5

UltraEval appears to be a system or process designed to evaluate data, likely in the context of a scientific or research project. The flowchart outlines the steps involved in this process:

**1.Data Preparation:** This stage involves collecting and organizing raw data, which can come from official data sources such as GitHub or Hugging Face. These are represented by icons indicating their origin.

**2.Evaluation Process:** After the data is prepared, it undergoes an evaluation process that includes several steps:

- **Prompt Input:** This could involve providing prompts or instructions for the model.

- **Model Deployment:** Here, the model is likely trained or configured to perform evaluations based on the input.

- **Local URL:** This step may involve accessing or using a local URL, possibly for deploying the model or retrieving results.

- **Post-process:** This could refer to further processing of the model's output, such as cleaning, filtering, or formatting data.

**3. Metric Calculation:** Following the post-process, metrics are calculated to evaluate the performance of the model. This is indicated by a graph icon, suggesting quantitative analysis.

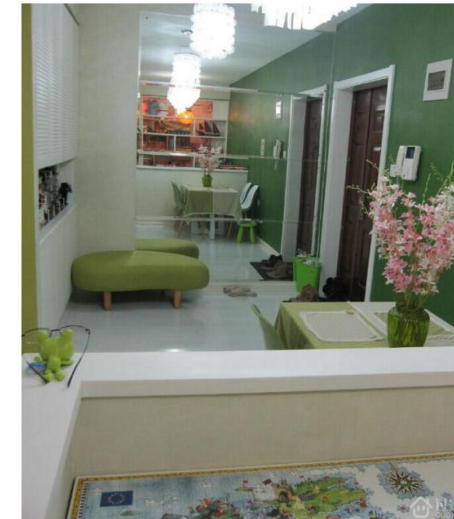
**4. Score:** The final stage is to determine the score, which is likely the outcome of the metric calculation. This score would reflect the model's performance or the quality of the data after evaluation.

**5. Model Support:** Throughout the process, there is support for various models, including vLLM/Torch, HF Model, and Custom Model. These models are likely used at different stages of the evaluation process.

In summary, UltraEval seems to be a structured approach to evaluating data using machine learning models, with a focus on performance metrics and customization options for different types of models.

## Prompt

What do you think are the issues with this interior design? Can you provide a detailed analysis of the flaws in this interior design and what suggestions would you propose to address these issues?



## MiniCPM-Llama3-V 2.5

The issues with this interior design include a lack of cohesion in color scheme, cluttered space, and mismatched furniture styles. To address these issues, one could start by selecting a unifying color palette to create a more harmonious environment. Additionally, decluttering and organizing the space can help create a more open and inviting atmosphere. Finally, replacing or reupholstering mismatched furniture with pieces that share a similar style can improve the overall aesthetic of the room.

[Yao, Yuan, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai et al. "Minicpm-v: A gpt-4v level mlm on your phone." arXiv:2408.01800 \(2024\).](#)





# Understanding menu cards and infographics

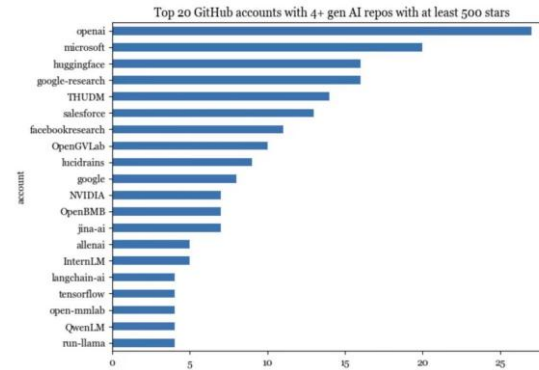
## Prompt

Can you list all of the items under pancakes?



## Prompt

Analyze this form, I want to know the trends and data



## MiniCPM-Llama3-V 2.5

The bar chart provides a visual representation of the popularity of GitHub accounts that have 4+ generation artificial intelligence repositories with at least 500 stars. The trends indicated by the chart suggest that there is a significant variation in the number of stars received by these accounts. The account with the longest bar, 'openai', has the highest number of stars, while 'run-llama' has the shortest bar, indicating it has the fewest stars among the listed accounts.

The data shows that 'openai' has significantly more stars than the other accounts, with its bar extending beyond the 25-star mark. This suggests that 'openai' is the most popular among the listed accounts. Other accounts, such as 'microsoft', 'huggingface', and 'google-research', also have relatively long bars, indicating they are quite popular as well.

On the other hand, 'run-llama' has the shortest bar, indicating it has the fewest stars among the listed accounts. This could suggest that it is either a newer account or less well-known compared to the others.

Overall, the chart provides a snapshot of the popularity of these GitHub accounts based on the number of stars they have received, with 'openai' being the most popular and 'run-llama' being the least popular among the listed accounts.

## MiniCPM-Llama3-V 2.5

Certainly. The items listed under the "PANCAKES" section on the menu are:

Original Pancake; Oat Pancake with a Twist; Cherry-Topped Pancakes; Yogurt Pancakes; Banana-Spice Cakes; Red Velvet Pancake; Pancakes with Apricots; Tropical Pancakes; Crepes Suzette; Shortcut Apple Pancakes

Each item is priced at \$10.

[Yao, Yuan, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai et al. "Minicpm-v: A gpt-4v level mlm on your phone." arXiv:2408.01800 \(2024\).](#)





# QA on bill images



I only had the sliced meat, how much do I need to pay? include a 18% tip

Here's the breakdown of the cost and the tip calculation, based on the receipt:

## 1. Identify the cost of the sliced meat:

- The receipt lists "1x Zürcher Geschnnetzelttes + Rösti" at CHF 36.50. This is the sliced meat dish with a side of Rösti (fried potatoes).

## 2. Calculate the 18% tip:

- Tip amount =  $\text{CHF } 36.50 \times 0.18 = \text{CHF } 6.57$

## 3. Calculate the total amount to pay:

- Total =  $\text{CHF } 36.50 + \text{CHF } 6.57 = \text{CHF } 43.07$

Therefore, you need to pay approximately CHF 43.07, including an 18% tip.

Team, Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin et al. "Gemma 3 technical report." *arXiv preprint arXiv:2503.19786* (2025).

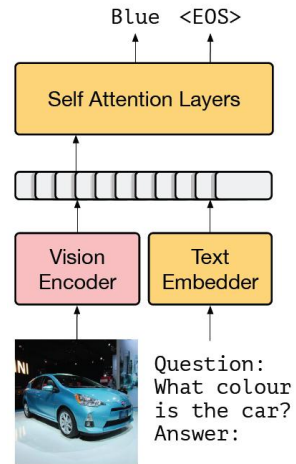
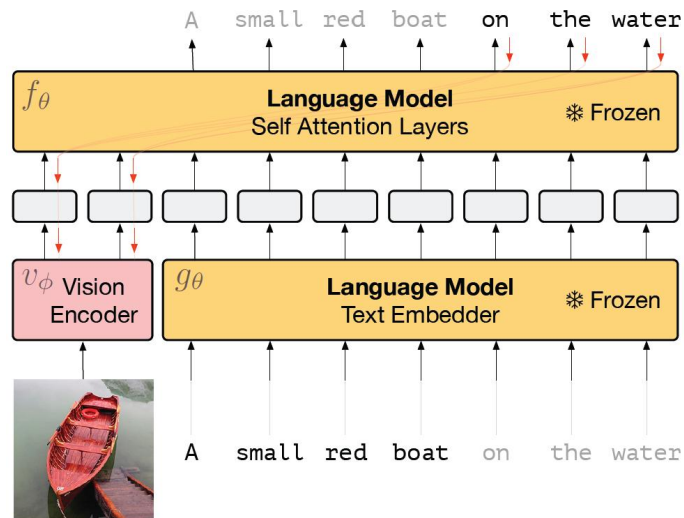


Advanced LLMs

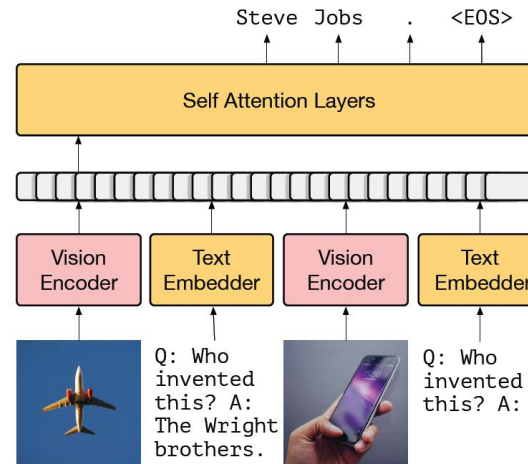


Manish Gupta

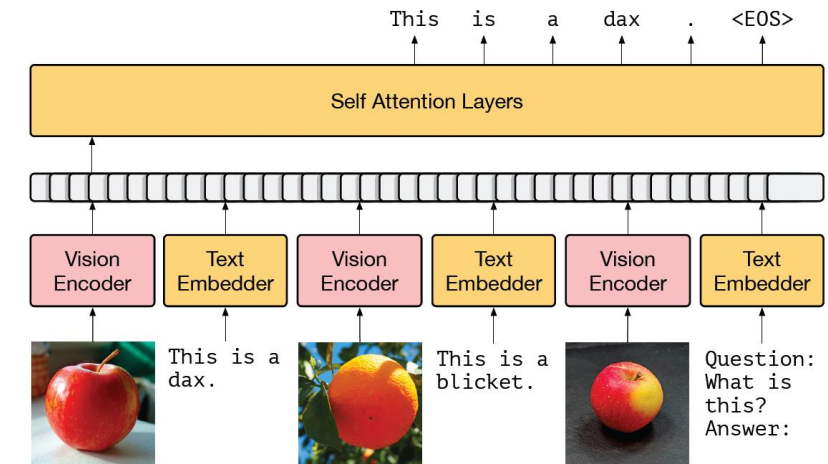
# Frozen: Multimodal few-shot learning



(a) 0-shot VQA



(b) 1-shot outside-knowledge VQA

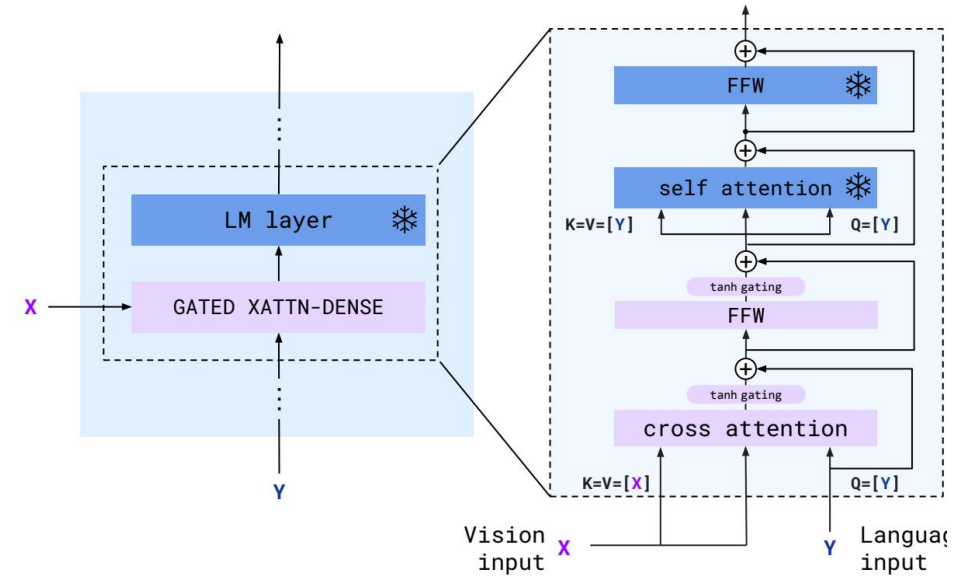
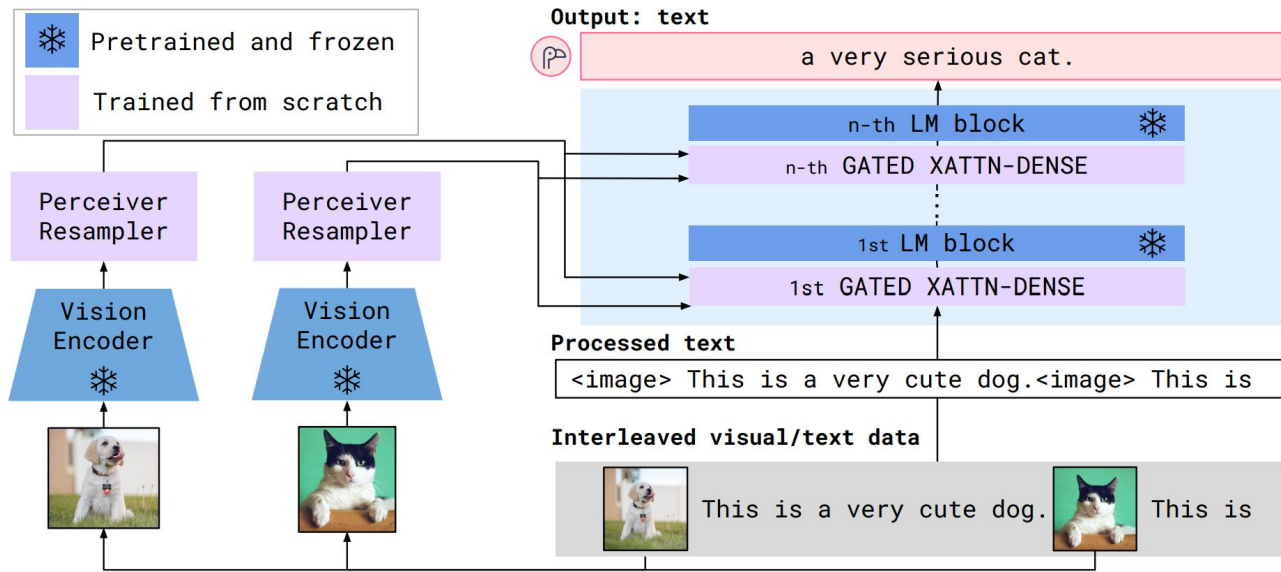


(c) Few-shot image classification

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.

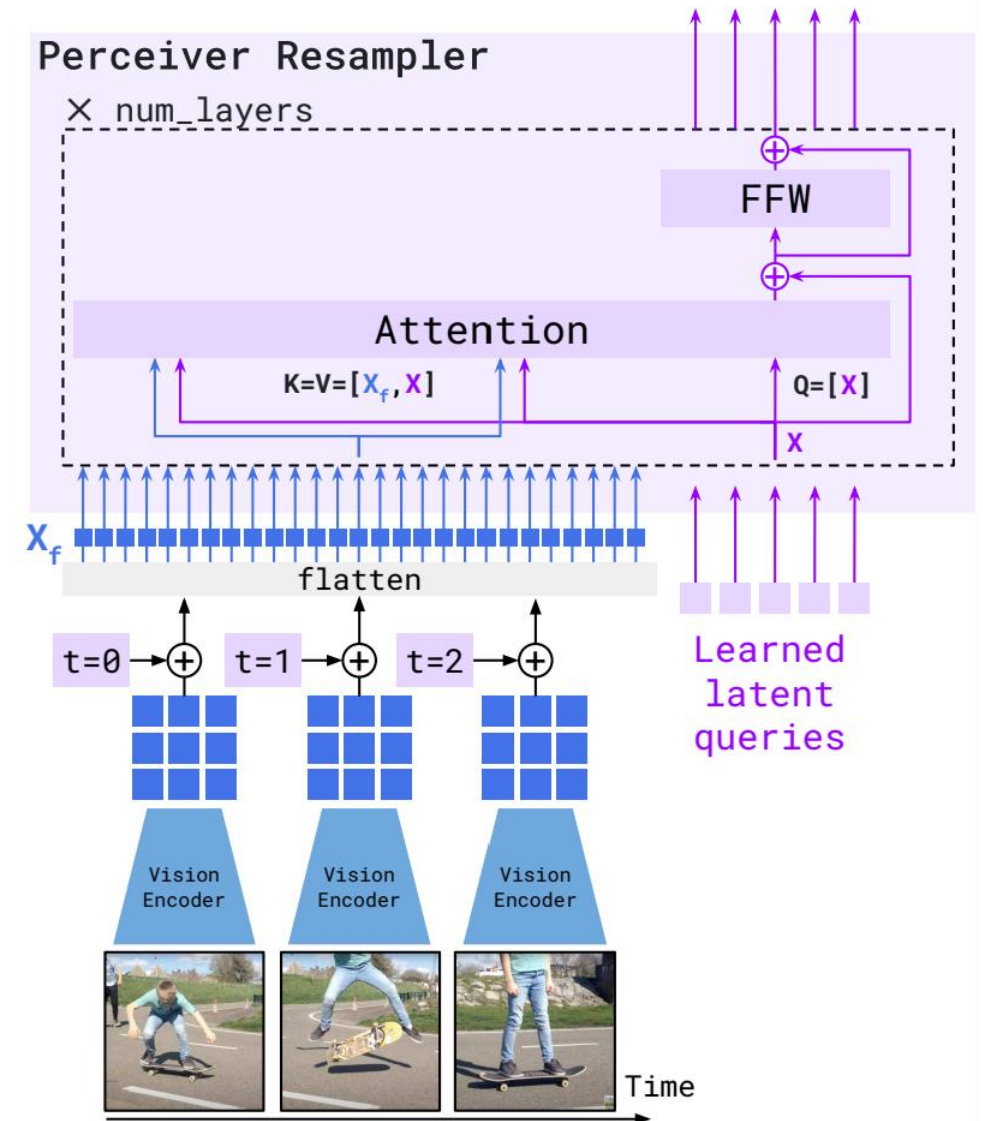
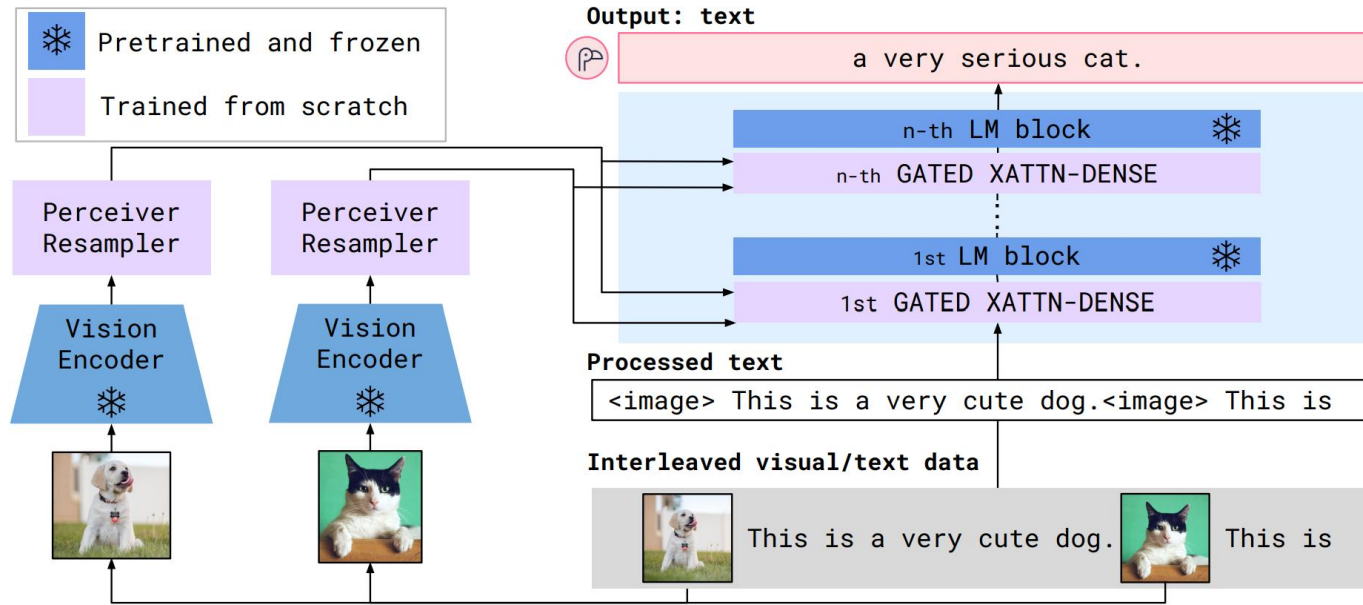
- Datasets:
  - MultiModal MassiveWeb (M3W): Interleaved image and text dataset.
  - Pairs of image/video and text: ALIGN dataset, LTIP (Long Text & Image Pairs), VTP (Video & Text Pairs)

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

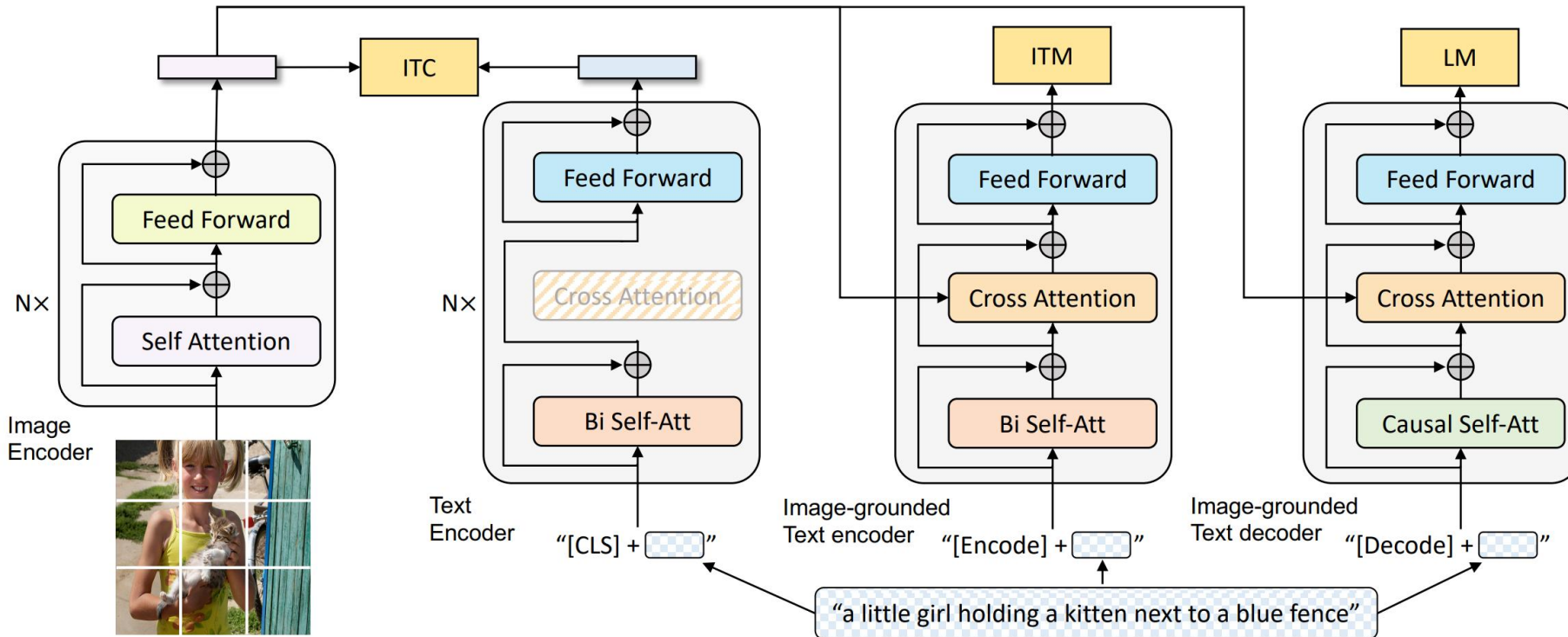


Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc et al. "Flamingo: a visual language model for few-shot learning." NIPS 35 (2022): 23716-23736.



# BLIP: Bootstrapping Lang-Image Pre-training

- Works for both understanding and generation
- 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

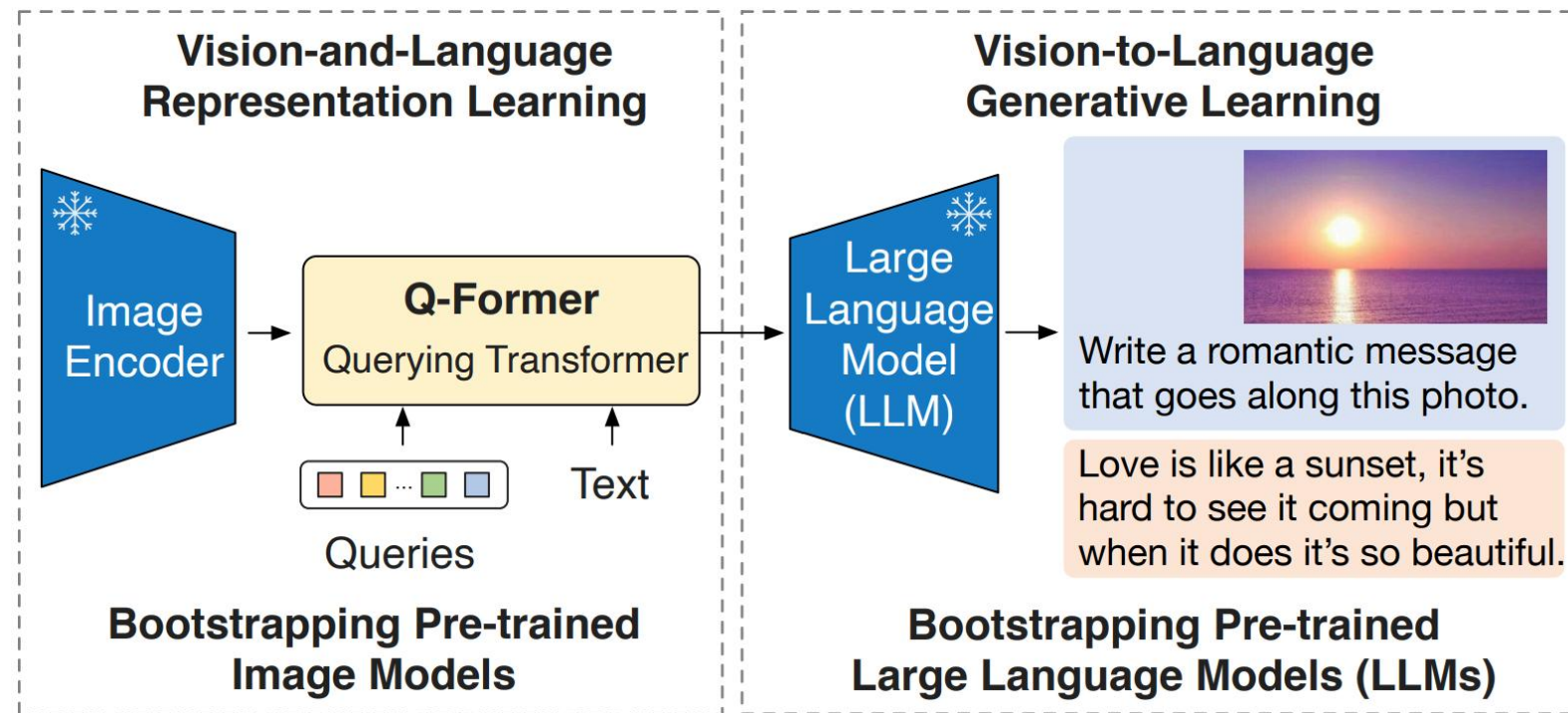


[Li, Junnan, Dongxu Li, Caiming Xiong, and Steven Hoi. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." In ICML, pp. 12888-12900. PMLR, 2022.](#)





# BLIP2



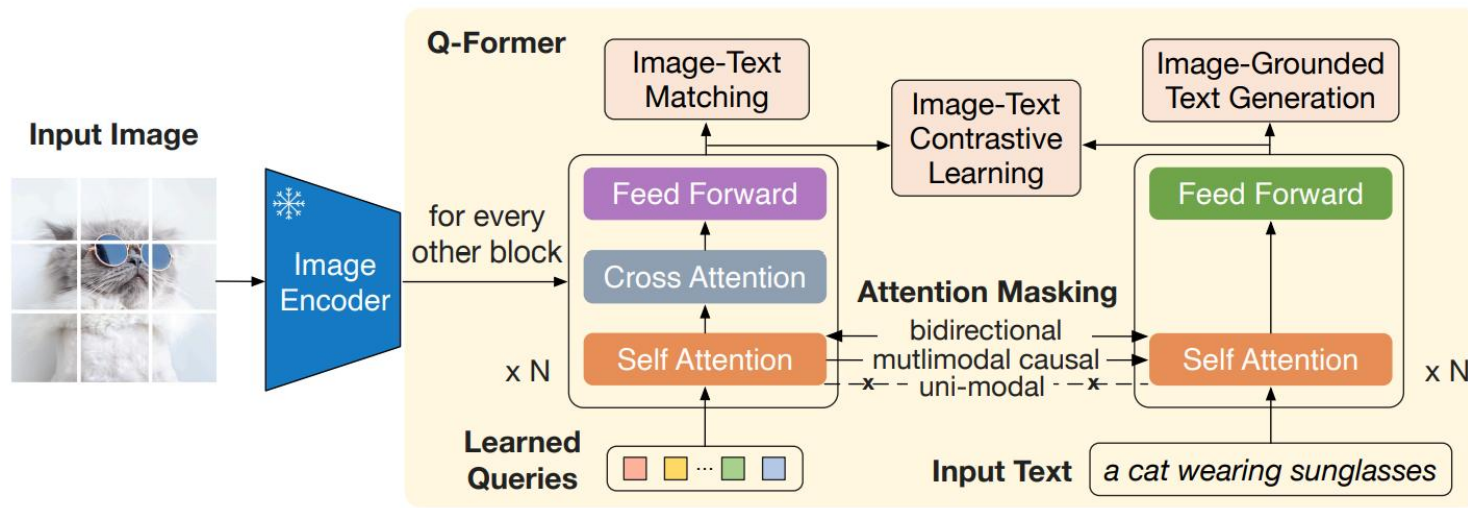
- Pre-trained in two stages
  - Vision-language representation learning stage with a frozen image encoder
  - Vision-to-language generative learning stage with a frozen LLM.
- Q-Former=image transformer+ text transformer

[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 \(2023\).](#)

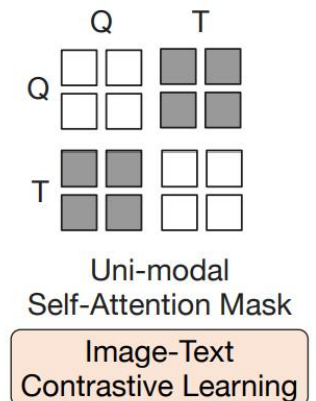
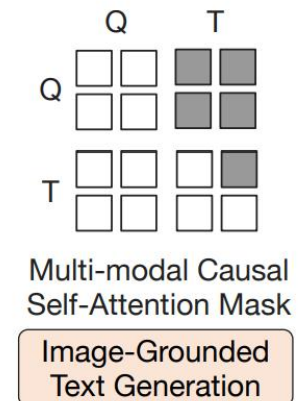
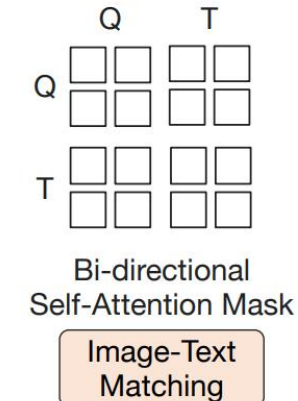


# BLIP2

- 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



**Q:** query token positions; **T:** text token positions.  
■ masked □ unmasked

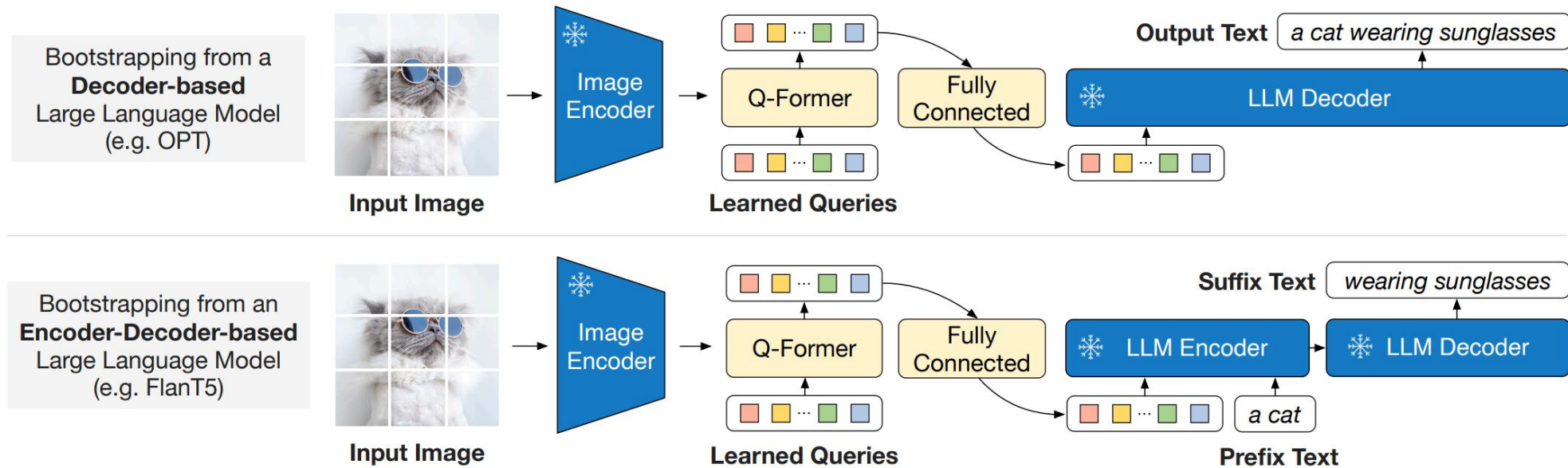


[Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." arXiv:2301.12597 \(2023\).](#)





# BLIP2



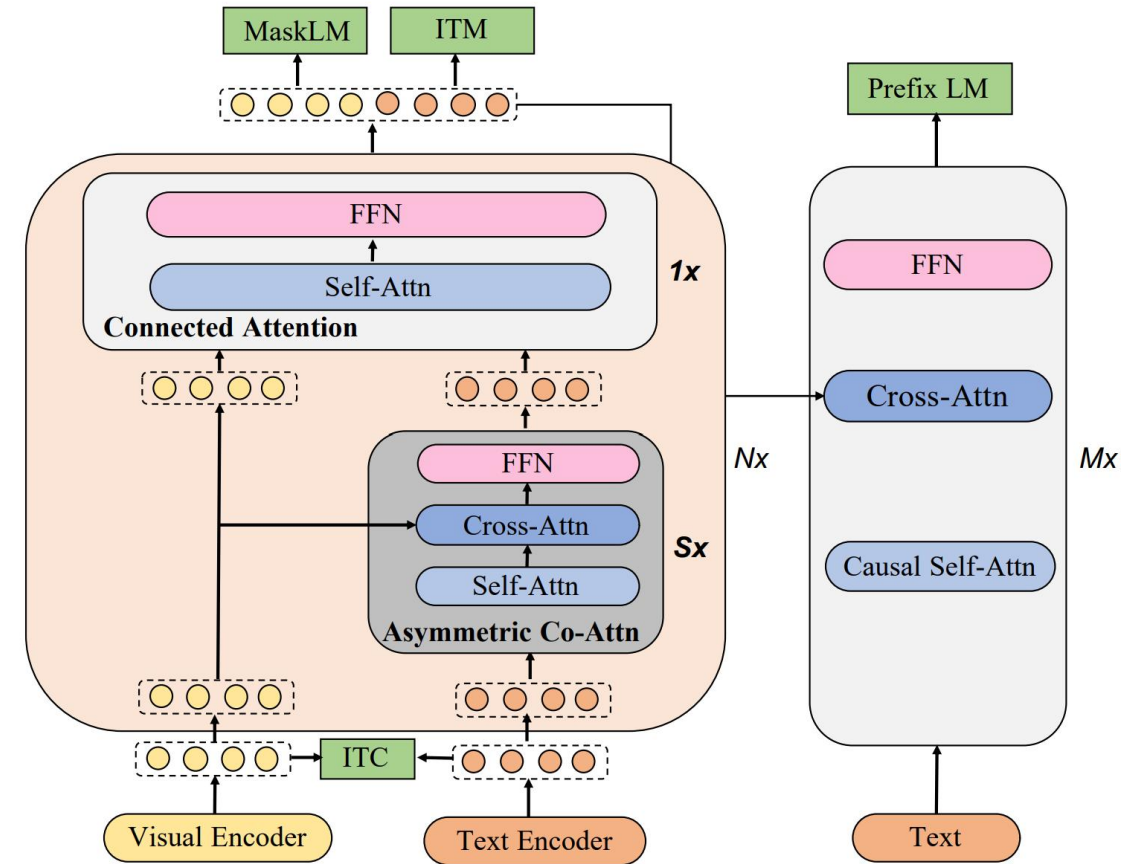
- Pre-training dataset
  - Same as BLIP
  - 129M images from COCO, Visual Genome, CC3M, CC12M, SBU
  - 115M images from LAION400M dataset.
- Pre-trained image encoder
  - ViT-L/14 from CLIP; ViT-g/14 from EVA-CLIP
- Frozen language model
  - OPT for decoder-based LLMs; FlanT5 for encoder-decoder-based 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 \(2023\).](https://arxiv.org/abs/2301.12597)



# mPLUG

- Instruction-following image-LLMs: LLaVA, mPLUG and MiniGPT4
- Pre-trained on 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)



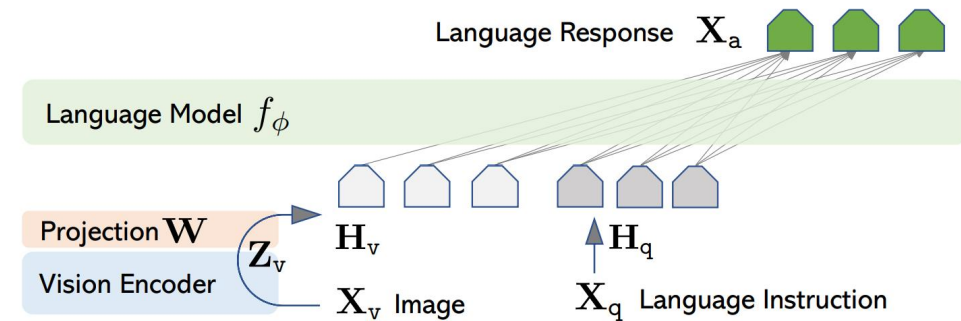
[Li, Chenliang, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye et al. "mplug: Effective and efficient vision-language learning by cross-modal skip-connections." arXiv:2205.12005 \(2022\).](#)





# LLaVa

- Stage 1: Pre-training for Feature Alignment
  - 595K image-text pairs.
  - Pretrain using 58K conv generated by GPT4.
  - 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 projection layer and LLM.
  - Finetune on 158K language-image instruction-following samples: 58K conversations, 23K detailed description, 77k complex reasoning.
    - Generated using few-shot prompts to GPT-4.



```
Xsystem-message <STOP> \n
Human : Xinstruct1 <STOP> \n Assistant: Xa1 <STOP> \n
Human : Xinstruct2 <STOP> \n Assistant: Xa2 <STOP> \n ...
```

$$X_{\text{instruct}}^t = \begin{cases} \text{Random choose } [X_q^1, X_v] \text{ or } [X_v, X_q^1], & \text{the first turn } t = 1 \\ X_q^t, & \text{the remaining turns } t > 1 \end{cases}$$

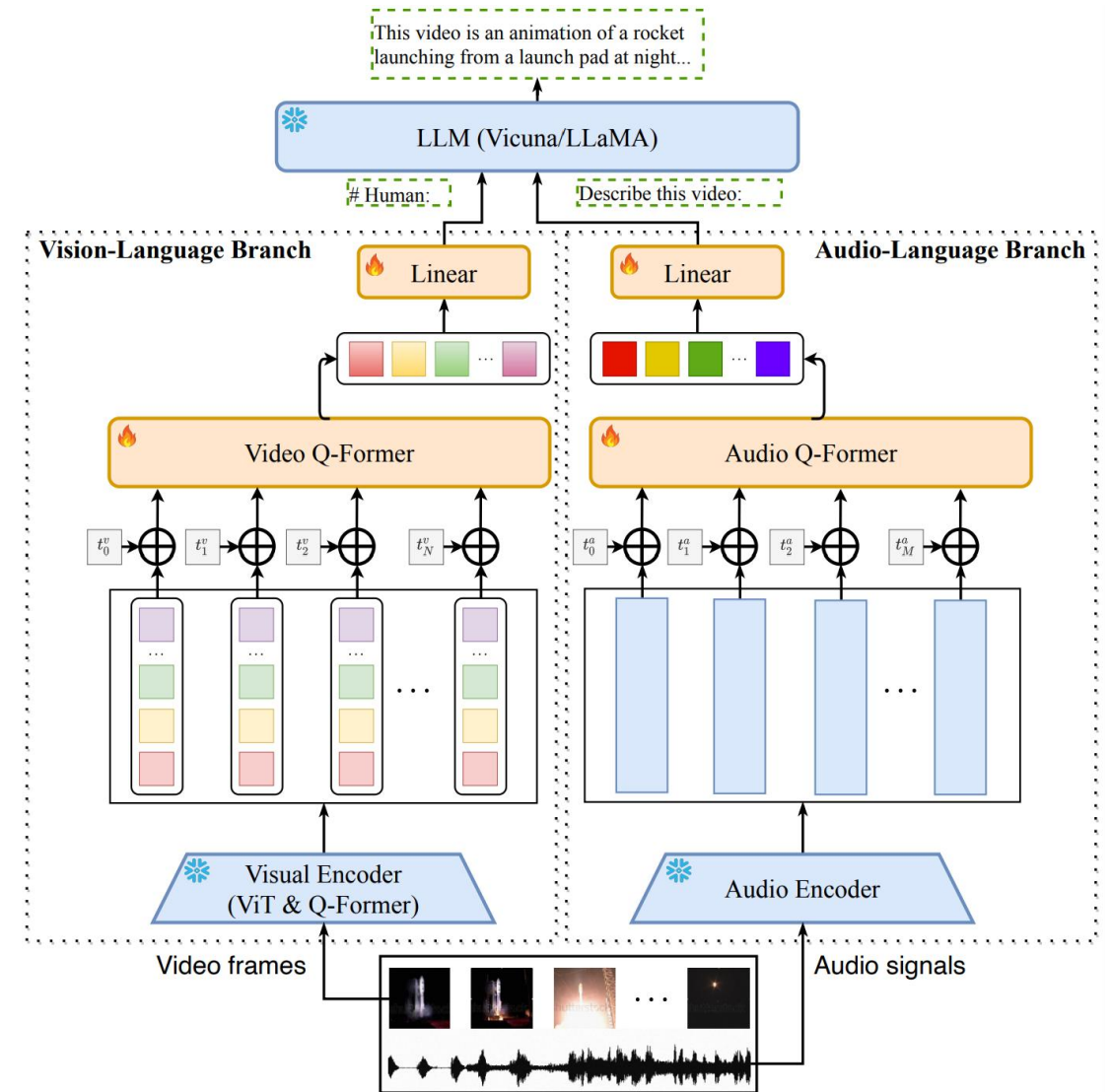
- LLM=LLaMa
- Pre-trained CLIP visual encoder ViT-L/14
- system-message = A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions.

[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
    - (b) image-caption data
  - Audio-language
    - Pre-train on an audio caption dataset.



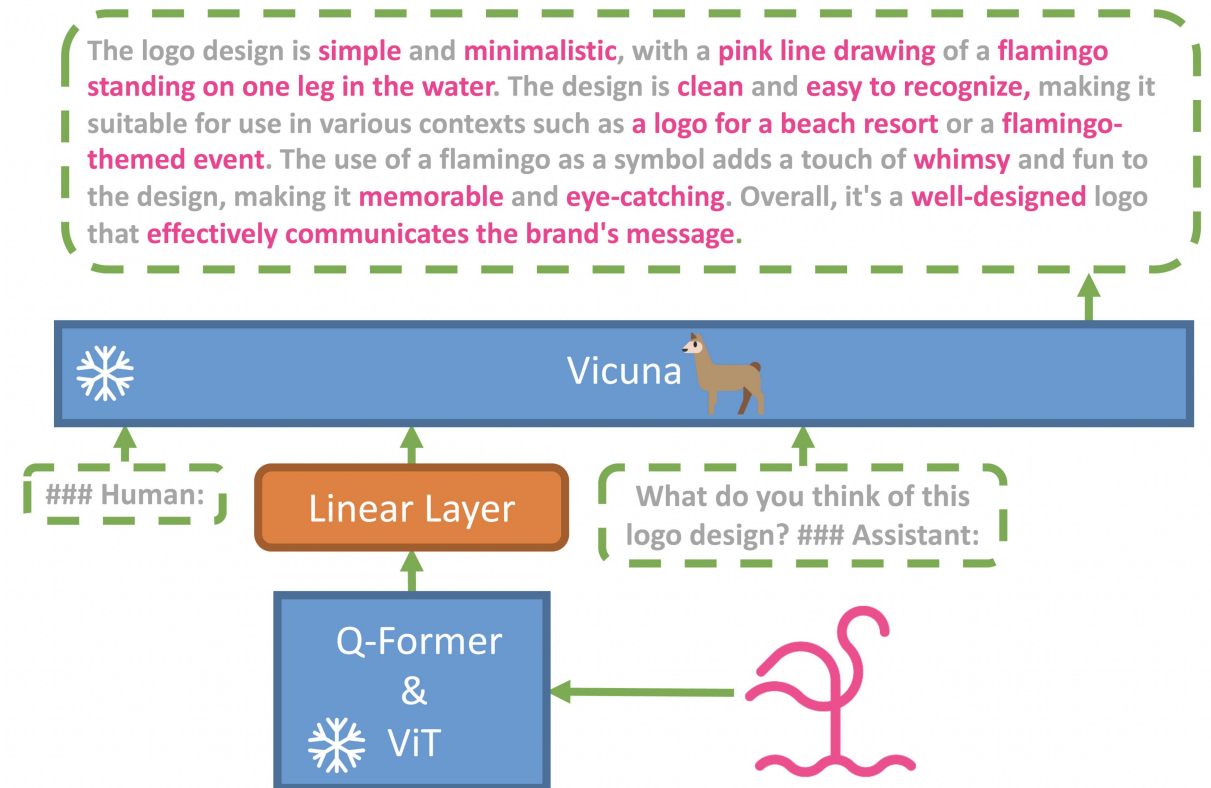
Zhang, Hang, Xin Li, and Lidong Bing. "Video-llama: An instruction-tuned audio-visual language model for video understanding." arXiv:2306.02858 (2023).





# MiniGPT-4

- Aligns a frozen visual encoder with a frozen Vicuna with 1 projection layer.
- Stage 1 (pretraining)
  - Vision encoder and LLM are frozen
  - Image captioning dataset (LAION, Conceptual Captions, and SBU)
- Stage 2 (finetuning)
  - 3,500 detailed image description pairs
  - ###Human:  
<Img> <ImageFeature> </Img> <Instruction>###Assistant:
  - <Instruction> is "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 Elhoseini  
"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 are 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.

<https://openai.com/index/hello-gpt-4b/>



Advanced LLMs

- 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.

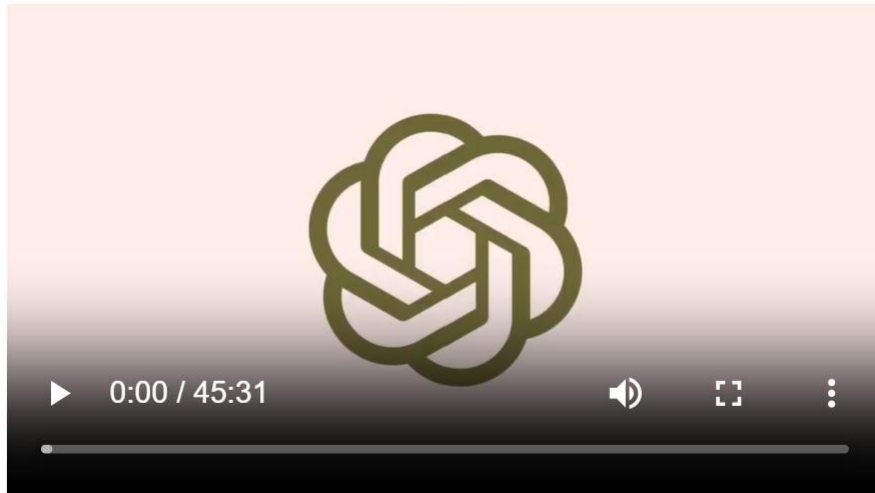
**Xavier:** My name is Xavier Joerkens. I'm the industrial designer, and there are three main questions that I have to find an answer to.



# 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?*



## 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.

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





# 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.

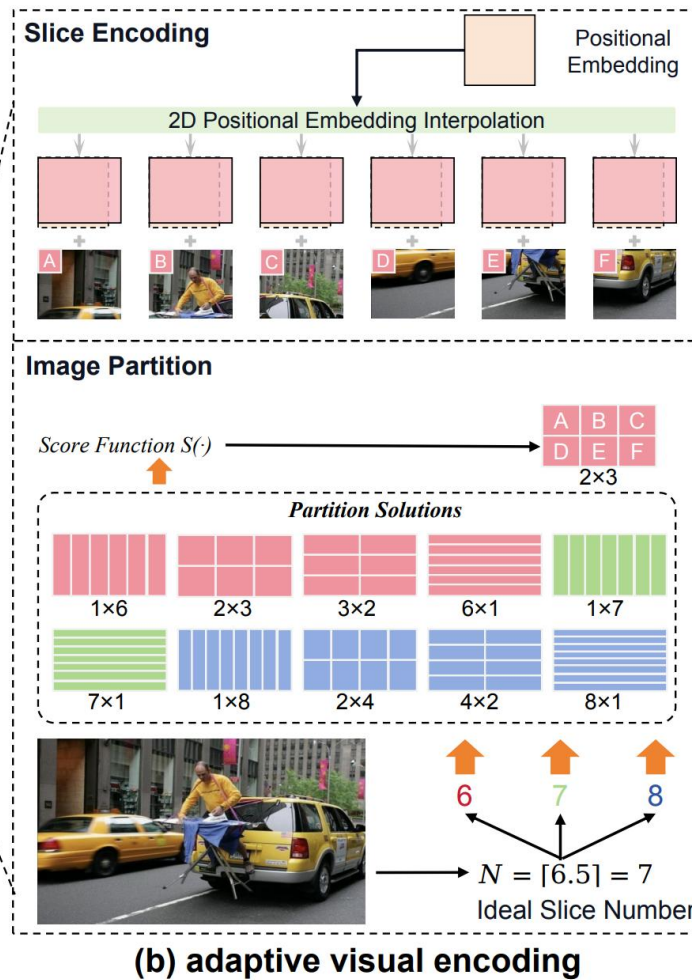
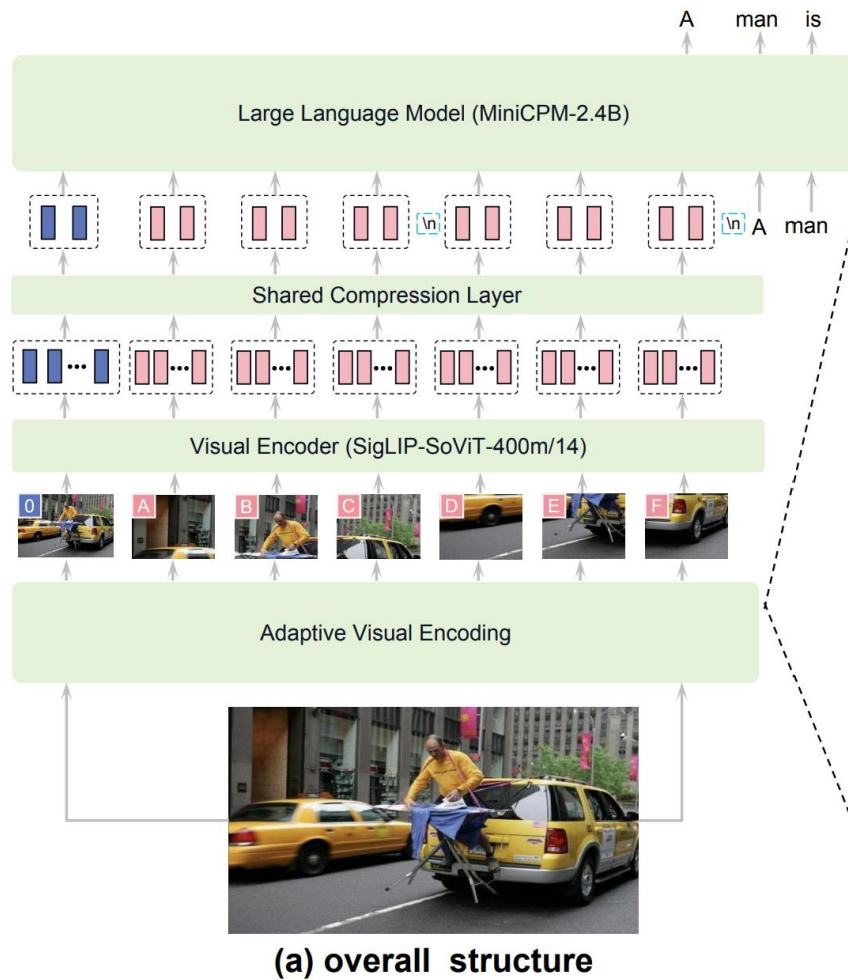
## Conclusion

- 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/>



# MiniCPM-V



- Adaptive Visual Encoding
  - Respect raw input aspect ratio
  - Preserve sufficient visual details (high resolution).
- Include orig image as a slice
- Compression layer: a perceiver resampler structure with 1 layer cross-attention.
  - Visual tokens of each slice are compressed into 64 queries for MiniCPM V1&2 and 96 tokens for MiniCPM-Llama3-V 2.5
- Spatial Schema: Wrap tokens of each slice by two special tokens <slice> and </slice>, and "<n>" to separate slices from different rows.

Yao, Yuan, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai et al. "Minicpm-v: A gpt-4v level mllm on your phone." arXiv:2408.01800 (2024).





# Help user to find flights (1/4)



Find round trip flights from SEA to NYC on 5th next month and filtered by price in ascending order.



**X Thought:** To proceed with finding flights from SEA to NYC on the 5th of next month, I need to first select the departure city. Click on the "From" field to open the dropdown menu for selecting the departure city.

**Action:** click (start\_box: [0.247, 0.236, 0.247, 0.236])



**X Thought:** To proceed, I need to type "SEA" into the "Origin" field. This will set the departure city to Seattle, which is the first step in specifying the flight search criteria. Type "SEA" into the "Origin" field to set the departure city to Seattle.

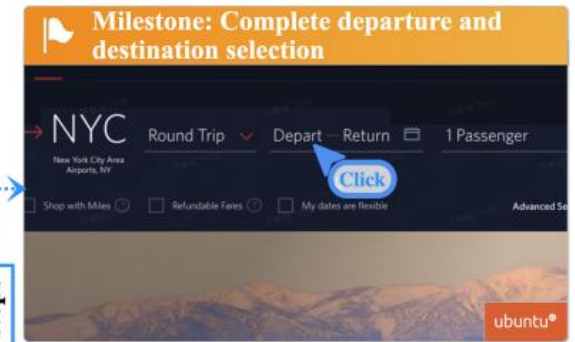
**Action:** type (content='SEA')



**X Thought:** To confirm the departure city as "SEA" (Seattle, WA), I need to select it from the dropdown menu. This ensures that the correct city is set for the flight search, allowing me to move forward with entering the destination and other search criteria. Click on the "SEA - Seattle, WA" option in the dropdown menu to confirm the departure city.

**Action:** click (start\_box: [0.297, 0.456, 0.297, 0.456])

Skip several steps



**X Thought:** I have completed departure and destination city selection. To proceed with the task, I need to specify the travel dates. Since the "Depart" field is highlighted, I will click on it to open the date selection interface, where I can choose the 5th of next month. Click on the "Depart" field to open the date selection interface for specifying the departure date.

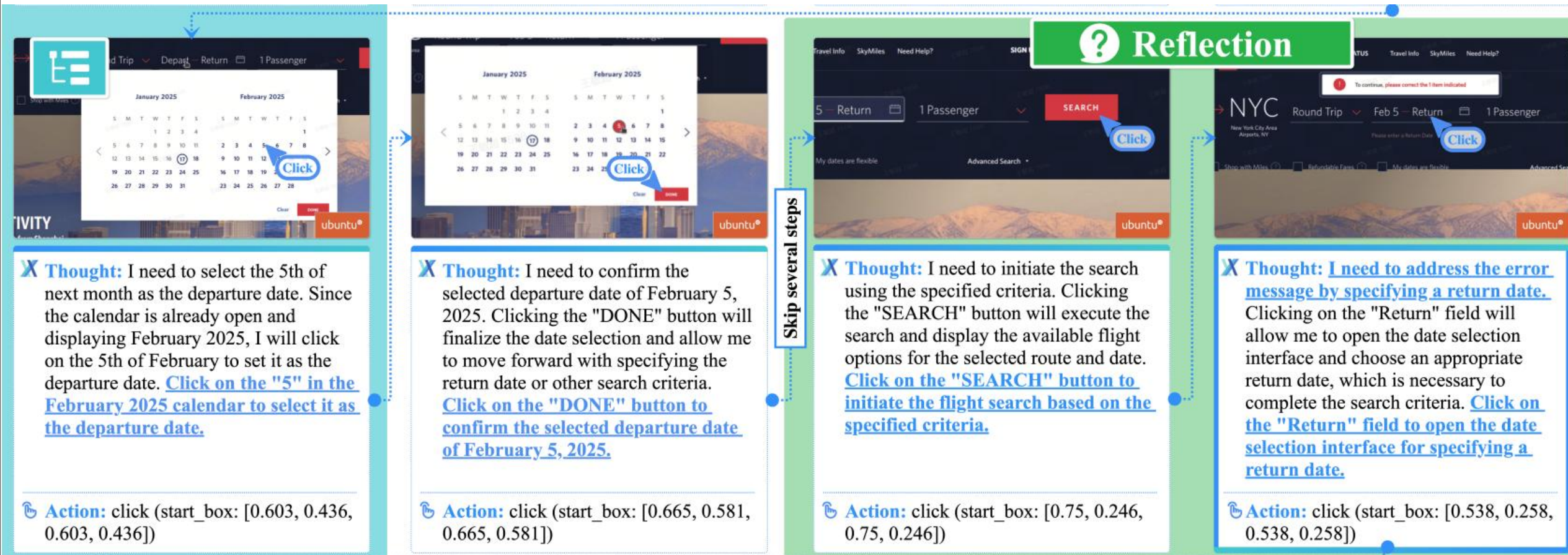
**Action:** click (start\_box: [0.5, 0.246, 0.5, 0.246])

Qin, Yujia, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang et al. "Ui-tars: Pioneering automated gui interaction with native agents." arXiv:2501.12326 (2025).





# Help user to find flights (2/4)



Qin, Yujia, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang et al. "Ui-tars: Pioneering automated gui interaction with native agents." arXiv:2501.12326 (2025).



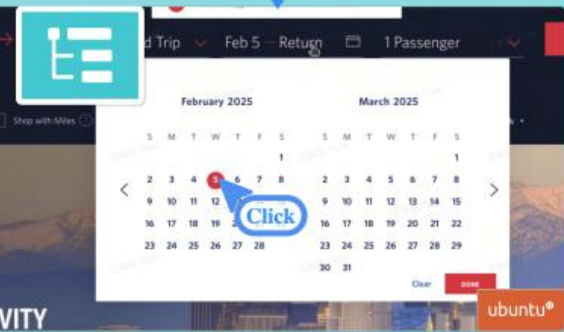
Advanced LLMs



Manish Gupta




# Help user to find flights (3/4)



**X Thought:** I need to specify a return date for a round trip. Since the calendar is already open and displaying February 2025, I will click on the 5th of February to set it as the return date, ensuring the search criteria are complete. Click on the "5" in the February 2025 calendar to select it as the return date.

**Action:** click (start\_box: [0.465, 0.451, 0.465, 0.451])

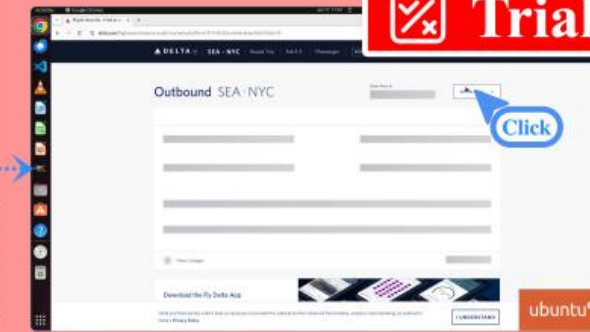
Skip several steps



**Milestone: Complete date selection**

**X Thought:** I need to initiate the search using the specified criteria. Clicking the "SEARCH" button will execute the search and display the available flight options for the selected route and date. Click on the "SEARCH" button to initiate the flight search based on the specified criteria.

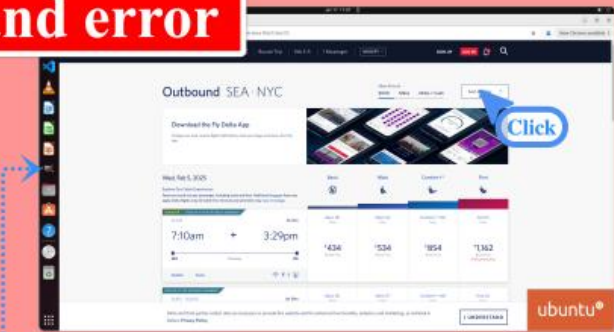
**Action:** click (start\_box: [0.75, 0.246, 0.75, 0.246])



**Trial and error**

**X Thought:** I need to apply filters to sort the results by price in ascending order. Clicking on the "Sort & Filter" button will open the filtering options, allowing me to set the desired sorting criteria and refine the search results accordingly. Click on the "Sort & Filter" button to access the filtering options for sorting the flight results by price in ascending order.

**Action:** click (start\_box: [0.761, 0.263, 0.761, 0.263])



**X Thought:** Due to network problems, my last action did not show the filter dropdown menu. Now the page is working now, and to proceed with the task, I need to apply filters again. Click on the "Sort & Filter" button.

**Action:** click (start\_box: [0.761, 0.263, 0.761, 0.263])

Qin, Yujia, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang et al. "Ui-tars: Pioneering automated gui interaction with native agents." arXiv:2501.12326 (2025).

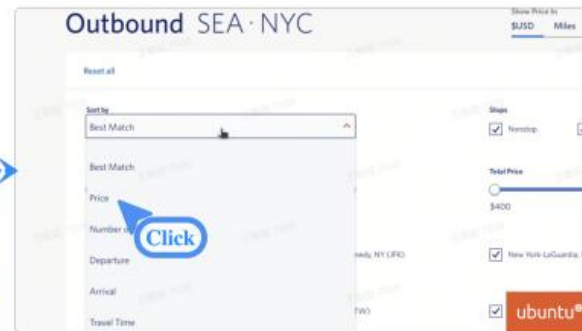


# Help user to find flights (4/4)



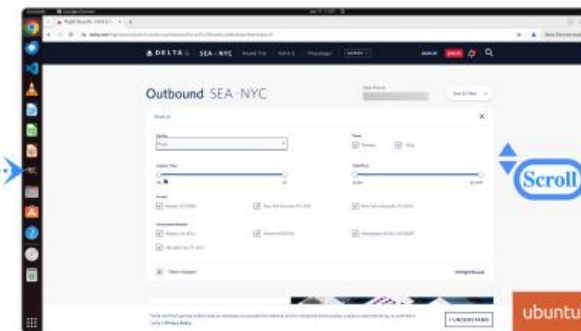
**X Thought:** I need to sort the results by price in ascending order. The "Sort by" dropdown menu is the appropriate control for this, and interacting with it will allow me to select the desired sorting criterion. Clicking on the dropdown will reveal the available sorting options, including the one for price. Click on the "Sort by" dropdown menu to reveal the sorting options and prepare to select the "Price" option.

**Action:** click (start\_box: [0.342, 0.418, 0.342, 0.418])



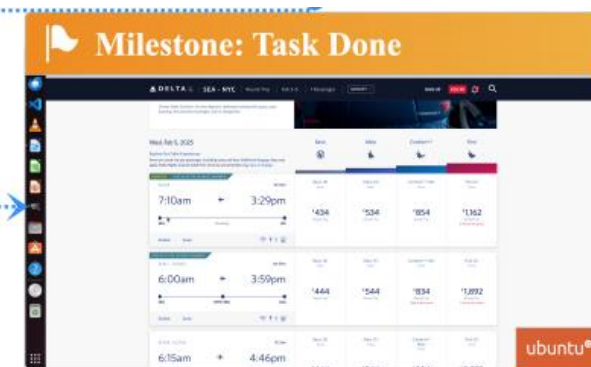
**X Thought:** I need to sort the results by price in ascending order. The "Price" option in the "Sort by" dropdown menu is the appropriate choice for this, and selecting it will ensure the results are displayed in the desired order, making it easier to identify the most affordable options. Click on the "Price" option in the "Sort by" dropdown menu to sort the flight results by price in ascending order.

**Action:** click (start\_box: [0.243, 0.521, 0.243, 0.521])



**X Thought:** I need to scroll down to see more details about the filtered flights. The page is at the top position, and scrolling down will show the details about the 10 flight results. Scroll down to see detailed results.

**Action:** scroll (direction='down')



**X Thought:** The task of finding round trip flights from SEA to NYC on the 5th of next month, filtered by price in ascending order, has been successfully completed. The results are now sorted by price, and all relevant filters have been applied. No further actions are required to fulfill the task, and I can now submit the task as completed. Submit the task, confirming that it has been successfully completed.

**Action:** finished

Qin, Yujia, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang et al. "Ui-tars: Pioneering automated gui interaction with native agents." arXiv:2501.12326 (2025).



Advanced LLMs



Manish Gupta



# UI-TARS Training data

## ③ Question Answering

Query: *What is that app with the waveform-like icon?*

Answer: The one you're referring to appears to be the Windows Task Manager icon, which displays a simple performance metric graph. Task Manager is used to monitor system performance and manage running programs and processes. From the screen, we can see that the user has opened this software, but it is currently minimized.

## ④ Set of Mark Perception

Query: What are the elements inside the yellow bounding box?

Answer: It's a button containing the text "Post a reply" with a blue background and white text.

Query: What is the text inside the green circle section?

Answer: Microsoft Community <Micr...

## ⑤ Dense Captioning

This is a Windows interface screenshot showing a maximized Outlook app window. In the navigation bar, you can see File, Home, Send/Receive... Currently, the second email under the "older" category is right-clicked, with the title "Question has a reply: When Windows 11 22H2 version will be released?"... In the expanded right-click menu, there are several options including Copy, Quick Print, Reply... From the bottom of the screenshot, you can see that the user has many apps open, including File Explorer, an application with a green circular icon. Chrome...

## ⑥ State Transition Captioning

The first screenshot shows a Windows desktop with an Outlook app maximized ... On it, you can see that the user has selected the second email under the "Older" category, which brought up a right-click context menu...

The second screenshot is largely identical to the first one, with the main difference being: that the Rules option in the right-click menu is expanded, showing additional options like "Always Move Messages From: Microsoft Community", "Always Move Messages To: learntechwithedi@outlook.com"... Additionally, the mouse is hovering over the "Manage Rules & Alerts" option.

From the screen content, we can infer that the user clicked on the Rules option in the right-click menu of the second email, which expanded into a submenu.

## ② Element Description

### Visual Description

A button with the text "NewEmail" and an envelope-shaped icon above it.

### Position information

Located in the upper-left corner of the page, it's the first button under the Home tab. To its right is the "New Items" button.

### Element Function

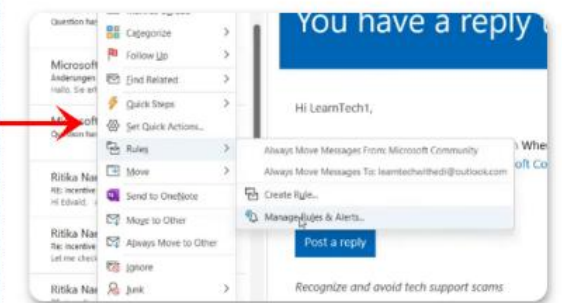
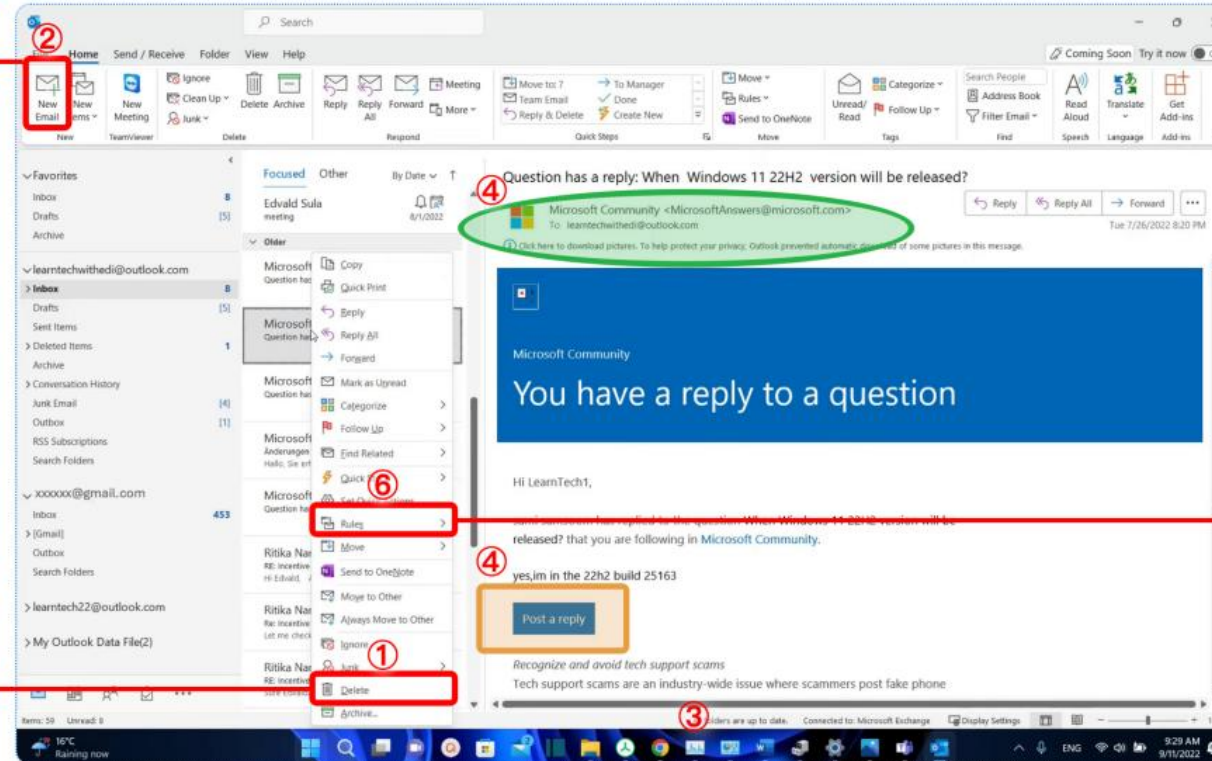
This button is interactive, clicking it will take you to the new email composition page.

### Element Type: Button

## ① Task Grounding

Query: Help me delete This Email.

Answer: [0.27,0.86,0.34,0.89]



(Second screenshot)



# How is UI-TARS trained?

- Tasks

- Enumerate all visible elements within a screenshot and generate their element descriptions, conditioned on the screenshot.
- Caption the entire interface while maintaining accuracy and minimizing hallucinations.
- Identify and describe the differences between two consecutive screenshots and determine whether an action, such as a mouse click or keyboard input, has occurred.
- QA, Set of Mark prompting.
- Predict the coordinates of the elements it

$P(t_n, a_n \mid \text{instruction}, t_1, a_1, \dots, (o_{n-i}, t_{n-i}, a_{n-i})_{i=1}^N, o_n)$ .  
top-right corner labeled Submit

Environment	Action	Definition
Shared	Click(x, y)	Clicks at coordinates (x, y).
	Drag(x1, y1, x2, y2)	Drags from (x1, y1) to (x2, y2).
	Scroll(x, y, direction)	Scrolls at (x, y) in the given direction.
	Type(content)	Types the specified content.
	Wait()	Pauses for a brief moment.
	Finished()	Marks the task as complete.
Desktop	CallUser()	Requests user intervention.
	Hotkey(key)	Presses the specified hotkey.
	LeftDouble(x, y)	Double-clicks at (x, y).
Mobile	RightSingle(x, y)	Right-clicks at (x, y).
	LongPress(x, y)	Long presses at (x, y).
	PressBack()	Presses the “back” button.
	PressHome()	Presses the “home” button.
	PressEnter()	Presses the “enter” key.

[Qin, Yujia, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang et al. "Ui-tars: Pioneering automated gui interaction with native agents." arXiv:2501.12326 \(2025\).](#)



# Summary

- Frozen
- Flamingo
- BLIP
- BLIP2
- mPLUG
- LLaVa
- Video-LLaMA
- MiniGPT-4
- MiniCPM-V
- UI-TARS





# 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>

