

Topics:

- Vision-Language Models

CS 4644-DL / 7643-A
ZSOLT KIRA

- Project due April 29 11:59pm (grace period **May 1st**)
- Fill out CIOS! <https://b.gatech.edu/cios>

Visual Math and Text Reasoning



What is "a" in the first equation?
What is y in the last equation?

$5 + a = 7$	$40 + y = 45$
$10 + y = 15$	$39 + a = 41$
$20 + a = 22$	$24 + y = 29$
$48 + a = 50$	$15 + a = 17$
$10 + a = 12$	$34 + y = 39$

The value of "a" in the first equation " $5 + a = 7$ " is 2.

The value of "y" in the last equation " $34 + y = 39$ " is 5.



Visual-Conditioned Joke/Meme



What is funny about it?



The humor in the comic may come from the situation, the characters, or the dialogue. The person is feeling unwell and is asked if they are pregnant or have the coronavirus, which is a current and relevant topic. The humor may come from the absurdity of the situation or the unexpectedness of the question.



Spatial/Coordinate Understanding



Where is the frisbee?
What is the relative position of the frisbee and the man?



The frisbee is located at coordinates (63,119,133,177).

The frisbee is located above the man in the image. The man is falling on the ground while the frisbee is in the air. The frisbee is also closer to the top of the image than the man..



Visual Planning and Prediction



What should I do after kneading the dough, before dividing it into pieces? For how long?



After kneading the dough, you should put it in a warm place before dividing it into pieces, according to the recipe in the image.

You should put the dough in a warm place for 1 hour.



Multi-Image Reasoning



How much in total did I spend on travel?
How much in total did I pay for taxes?

The Shop	
Small Blue Backpack	\$12.99
Medium Blue Backpack	\$19.99
Large Blue Backpack	\$29.99
Small Red Backpack	\$12.99
Medium Red Backpack	\$19.99
Large Red Backpack	\$29.99
Small Green Backpack	\$12.99
Medium Green Backpack	\$19.99
Large Green Backpack	\$29.99
Small Yellow Backpack	\$12.99
Medium Yellow Backpack	\$19.99
Large Yellow Backpack	\$29.99
Total	\$43.83
Tax	\$321.90
Total	\$365.73

You spent \$1299.90 on airfare and \$43.83 on an Uber ride for a total of \$1343.73 on travel.

You paid \$323.23 in total for taxes. This includes \$321.90 for taxes and fees for the flight purchase and \$1.33 for sales tax for the Windmill Family Restaurant.



Multi-Hop Document Understanding



Excluding the regular cab, what is the average total fueling cost?



The average total fueling cost excluding the Ford F150 (Regular Cab) is \$76.55. This is calculated by taking the sum of the total fueling cost of Toyota RAV4 and Honda Civic and dividing it by 2. $(\$82.56 + \$70.55) / 2 = \$76.55$.



Open-World Concept Understanding



What is the brand of these drink? Which one is unsweetened?



The brand of these drinks is "BETTER THAN MILK". The unsweetened one is "BETTER THAN MILK organic OAT DRINK".



Who are the celebrities in this image?



Keira Knightley and Natalie Portman are the celebrities in this image.



Is it safe to eat this mushroom?



Morel mushrooms are generally considered safe to eat when properly cooked.



Can you breakdown this video tutorial to step-by-step instructions, with start and end time?



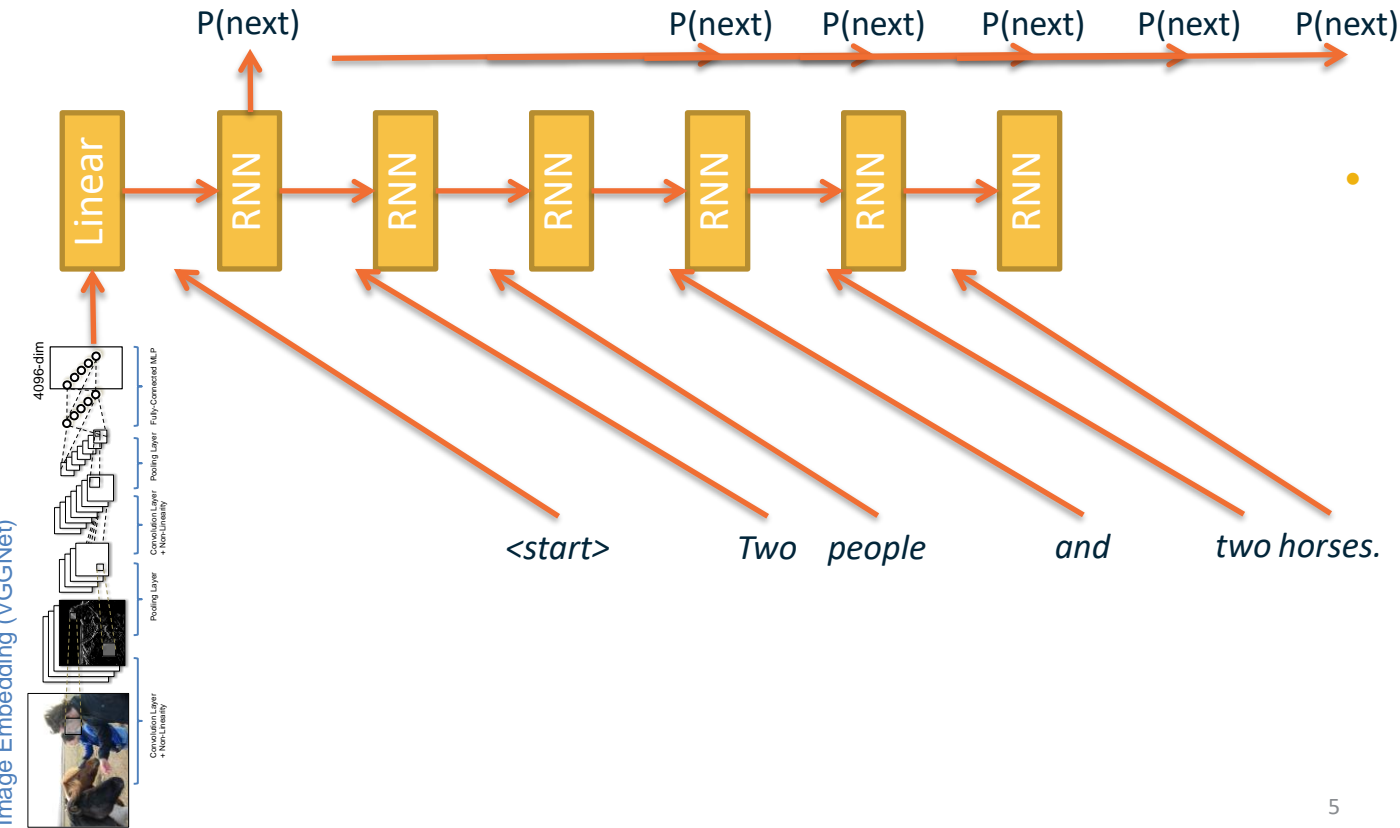
Here's a step-by-step breakdown of the tutorial with start and end time.



... Add Lettuce Add Tomato Add Bacon ...



- Image+LSTM
- CLIP
- ViLbert
- Flamingo
- BLIP/BLIP-2
- LLaVA
- ImageBind / LanguageBind



- **Features:**
 - Pre-trained visual encoder (vector)
 - Linear projection layer to map to “captionable” space
 - Vector serves as initial input to RNN



?

How should we encode
this (representations?)?

How will they be
learned?

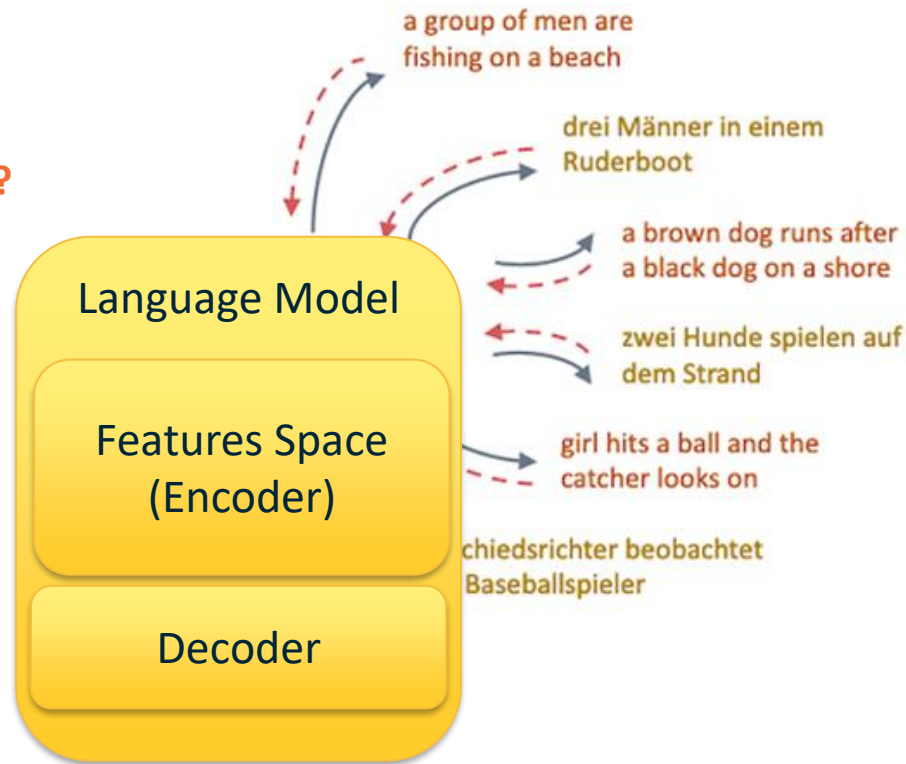
How/what should we train?

Using what data?

What tasks can we do?

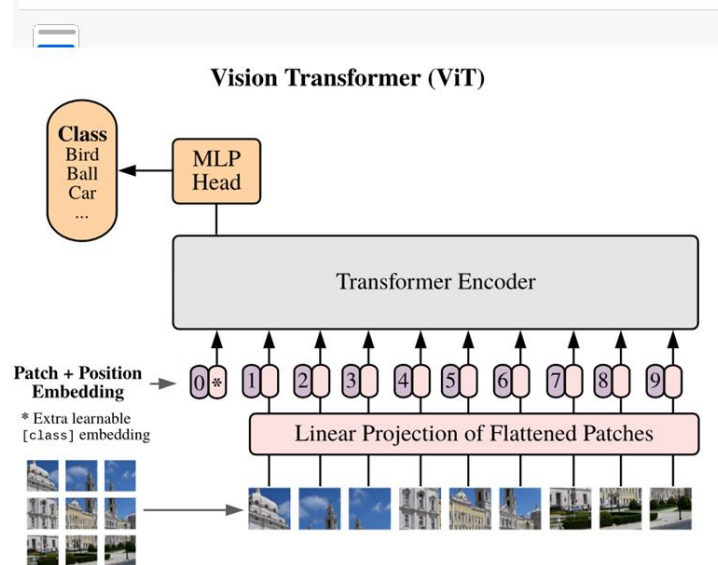


What should the interface be?



Potential ways of representing an image?

- Image encoder
 - Any architecture: ResNet, Vision transform (ViT)
 - Randomly initialized, SL/SSL pre-trained

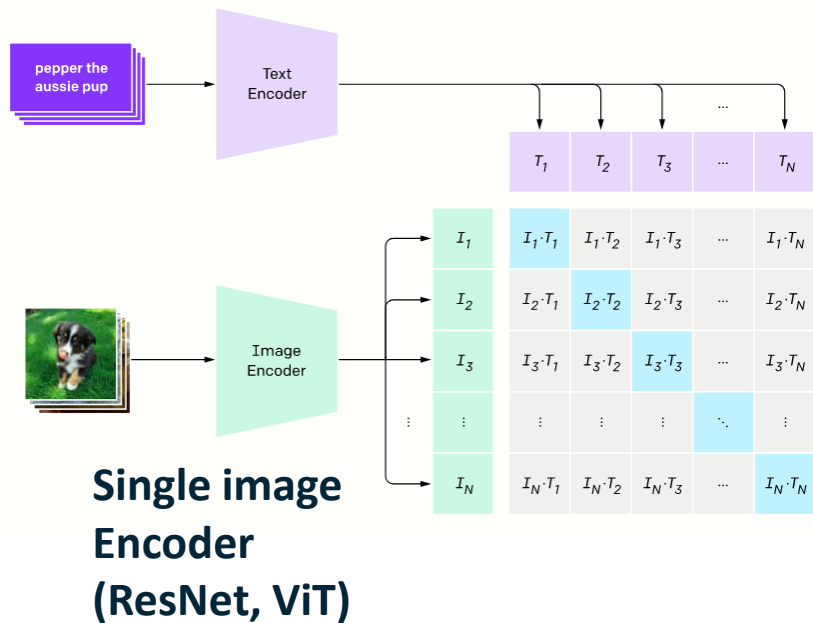


Method of alignment: Contrastive Learning

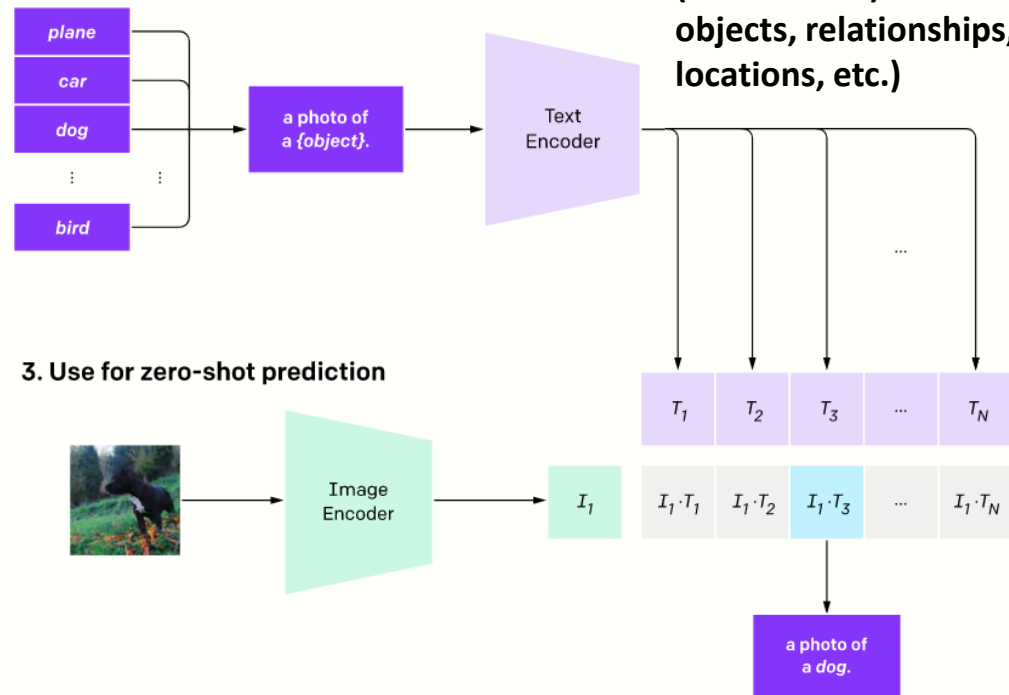
Downside?

Coarse-grained.
Has to represent
(somewhere) notion of
objects, relationships,
locations, etc.)

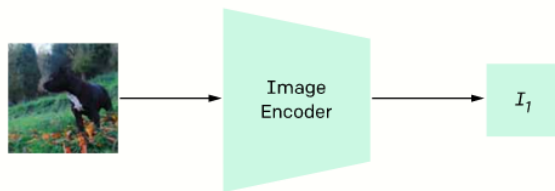
1. Contrastive pre-training



2. Create dataset classifier from label text



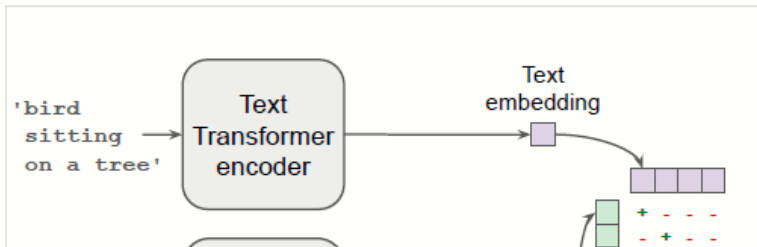
3. Use for zero-shot prediction



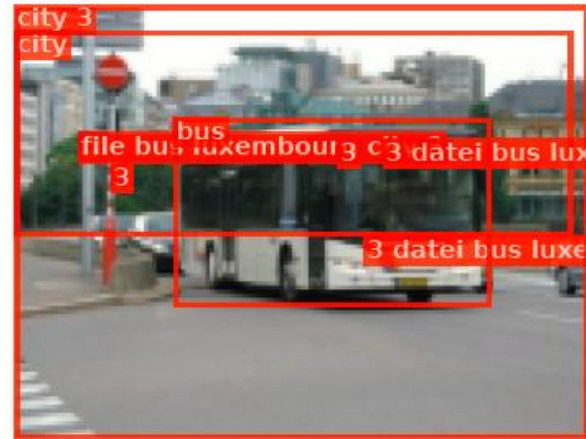
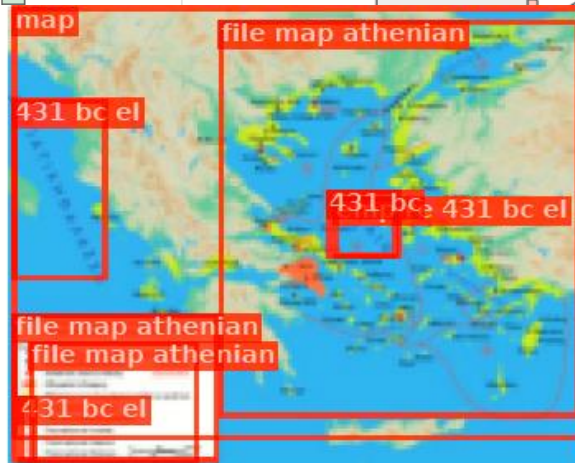
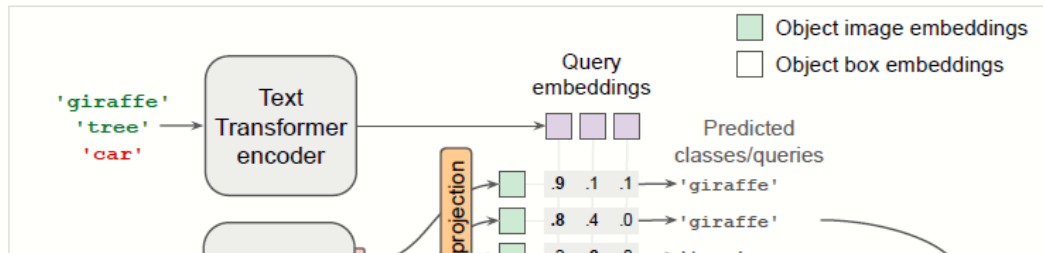
Radford et al., Learning Transferable Visual Models From Natural Language Supervision

CLIP: Learning More Aligned Representations

Image-level contrastive pre-training



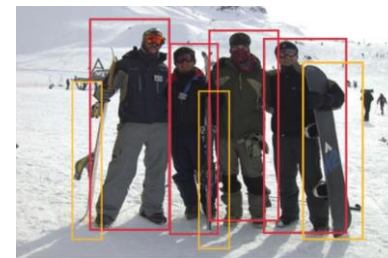
Transfer to open-vocabulary detection



Minderer et al., Simple Open-Vocabulary Object Detection with Vision Transformers
 Minderer et al., Scaling Open-Vocabulary Object Detection

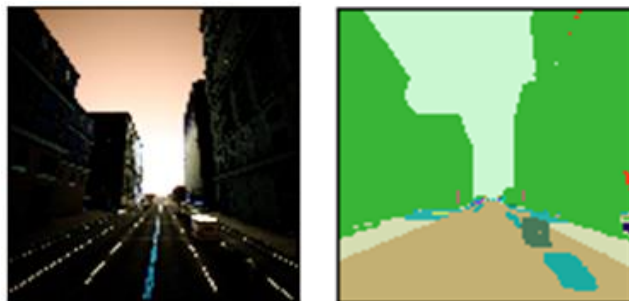
Potential ways of representing an image?

- Image encoder
 - Randomly initialized, SL/SSL pre-trained
- **Alternative:**
 - Bounding boxes/segments/regions + features

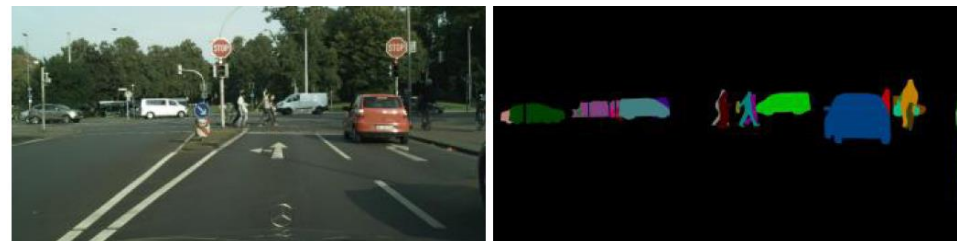


Object Detection

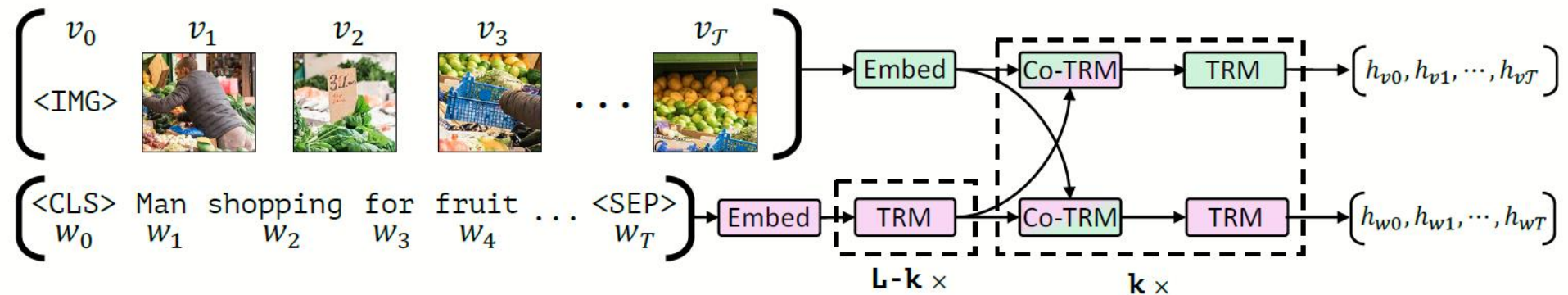
(List of bounding boxes with class distribution per box)



Semantic Segmentation

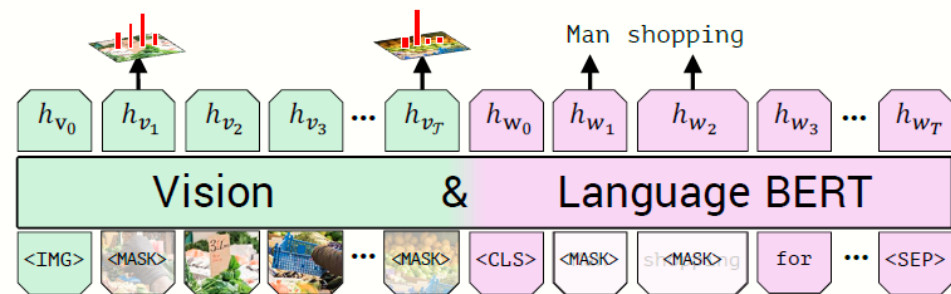


Instance Segmentation

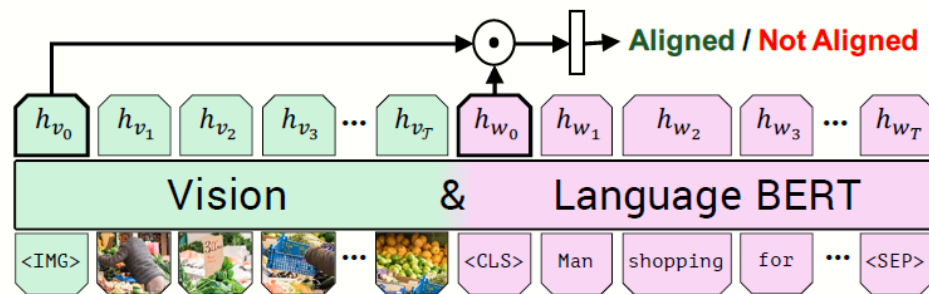


Training: Masked Prediction + Alignment

Interaction/Fusion: Cross-Attention



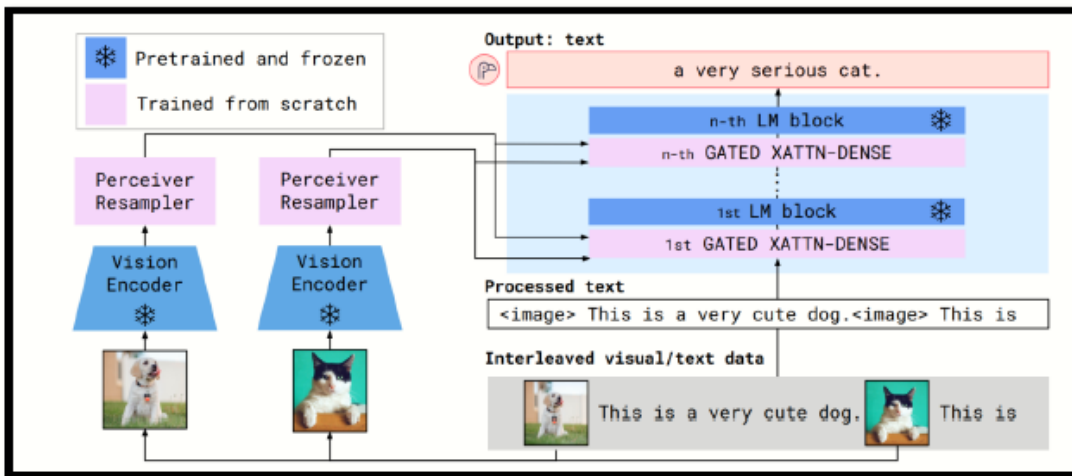
(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

Lu et al., ViBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

- Flamingo:



Language Model

Connection Module

Vision Encoder

Pre-trained: 70B Chinchilla

Perceiver Resampler
Gated Cross-attention + Dense

Pre-trained: Nonnormalizer-Free ResNet (NFNet)

Multimodal Few-Shot Learning with Frozen Language Models

Maria Tsimpoukelli*
DeepMind
mrts@deepmind.com

Jacob Menick*
DeepMind
University College London
jmenick@deepmind.com

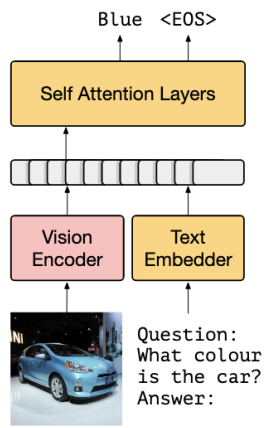
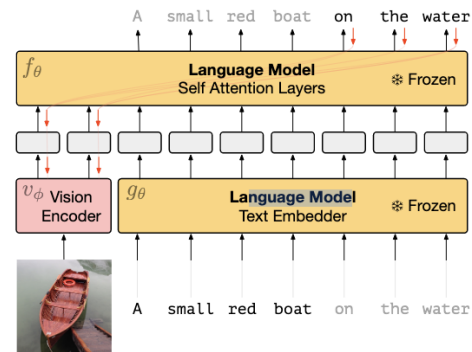
Serkan Cabi*
DeepMind
cabi@deepmind.com

S. M. Ali Eslami
DeepMind
aeslami@deepmind.com

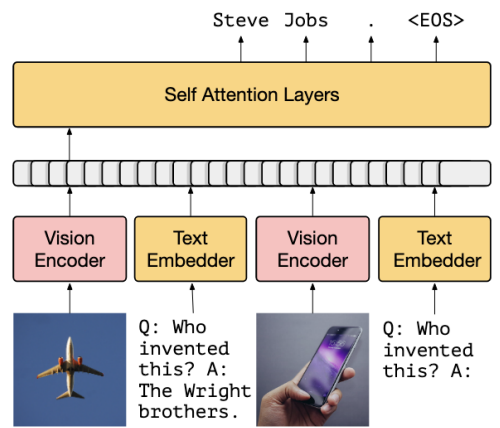
Oriol Vinyals
DeepMind
vinyals@deepmind.com

Felix Hill
DeepMind
felixhill@deepmind.com

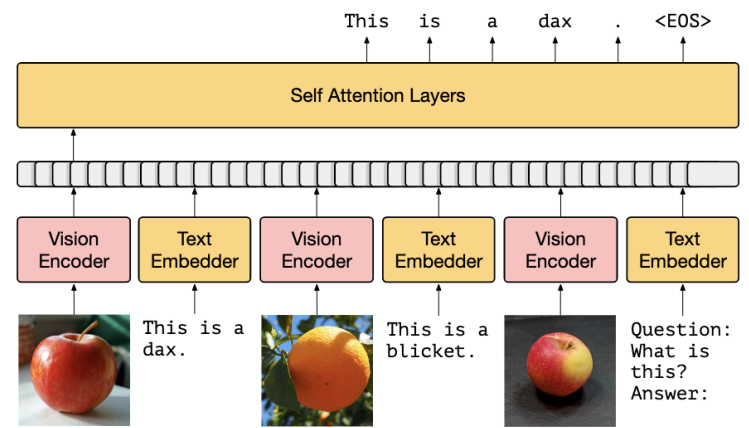
Training:



(a) 0-shot VQA



(b) 1-shot outside-knowledge VQA

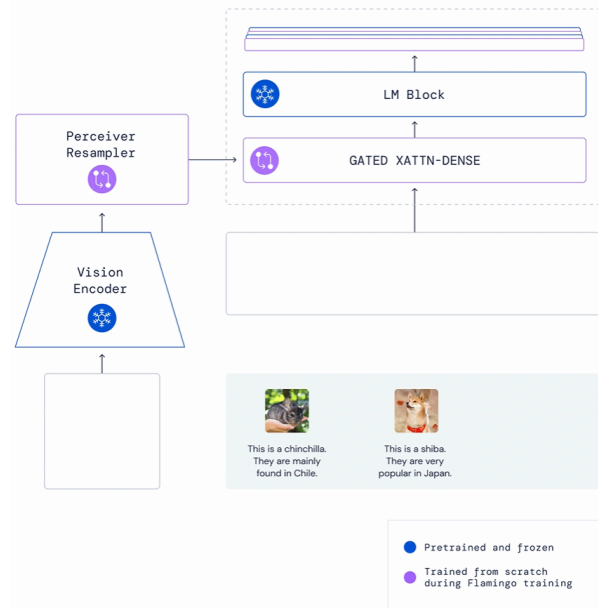


(c) Few-shot image classification

Zero to Few-Shot

Flamingo VLM [4]

- What is Flamingo?
 - It's a Visual Language Model (VLM) for Few-Shot Learning that launched by DeepMind.
- Visual Language Model?
 - Processing images to generate reasonable text.
- What can it do?
 - Applicable to image and video understanding tasks via simply prompting it with a few examples
 - captioning, visual dialogue, classification, visual question answering



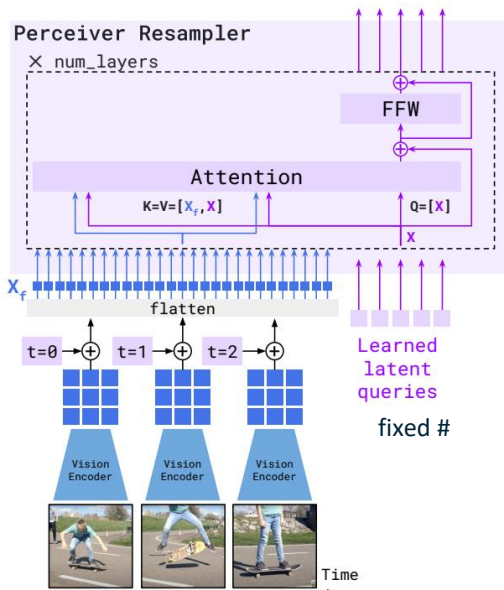
[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, et al. Flamingo: a Visual Language Model for Few-Shot Learning. DeepMind 2022

Flamingo VLM

- Three challenges for training with image/video and text.
 - **Supporting both images and videos**
 - Images /videos :2D structure with high dimensionality.
 - Text: 1D sequence
 - Sol.: Introduce Perceiver Resample module.
 - **The interaction with image/video and text**
 - keep the pretrained model's language understanding and generation capabilities fully intact
 - Sol.: Interleave cross-attention layers with frozen self-attention. gating mechanism.
 - **Obtaining multimodal dataset to induce good generalist capabilities**
 - Dataset with weak matching problem
 - Sol.: combine dataset with standard strong related paired image/text and video/text datasets

Flamingo VLM

- Model structure - Supporting both images and videos



```
def perceiver_resampler(
    x_f, # The [T, S, d] visual features (T=time, S=space)
    time_embeddings, # The [T, 1, d] time pos embeddings.
    x, # R learned latents of shape [R, d]
    num_layers, # Number of layers
):
    """The Perceiver Resampler model."""

    # Add the time position embeddings and flatten.
    x_f = x_f + time_embeddings
    x_f = flatten(x_f) # [T, S, d] -> [T * S, d]
    # Apply the Perceiver Resampler layers.
    for i in range(num_layers):
        # Attention.
        x = x + attention_i(q=x, kv=concat([x_f, x]))
        # Feed forward.
        x = x + ffw_i(x)
    return x
```

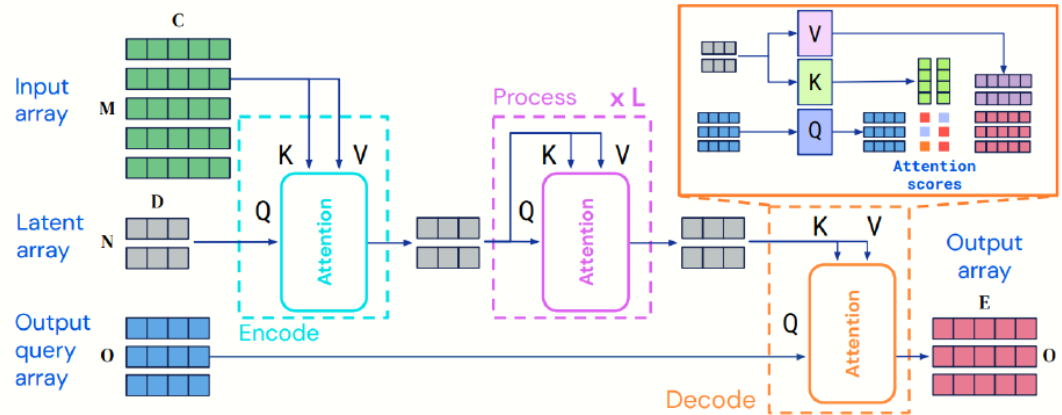
pseudo code

- Using pre-trained ResNet to get visual features X_f
- Compress the encode image into R tokens
- Core of this module : Attention .
 - Query: the learned latent token X
 - Key=Value: the concatenation of X_f and the learned latent token X
 - Better performance by concatenating keys and values obtained from latent
- If the input is video
 - X_f will add time embeddings

Maps a variable size grid of visual features from the Vision Encoder to a fixed number of output token (5 in the figure.)

Perceiver / Perceiver IO: Transformer for general data perception

- General data processing method given data can be mapped into sequence of vectors
- Use cross attention to fetch information from input
- Self attention to process input.
- Use cross attention to fetch relevant information and send to output.



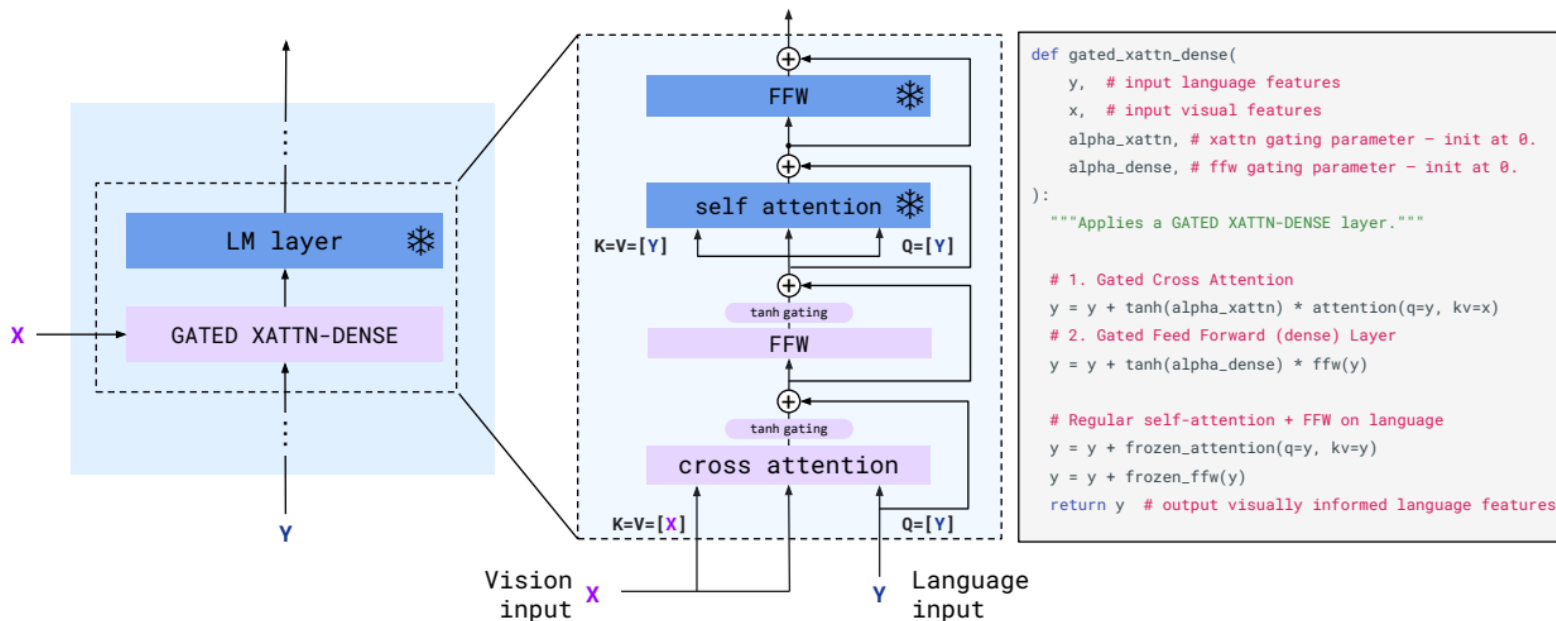
Jaegle, Andrew, et al. "Perceiver: General perception with iterative attention." *ICML*, 2021.

Jaegle, Andrew, et al. "Perceiver io: A general architecture for structured inputs & outputs." *ICLR*, 2021



Flamingo VLM

- Model structure - The interaction with image/video and text

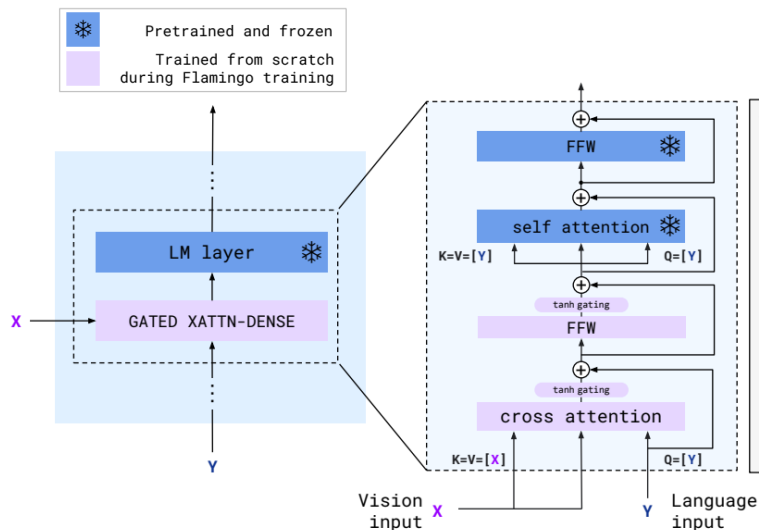


A **Gated Cross attention** mechanism is proposed to fuse images and text.



Flamingo VLM

- Model structure - The interaction with image/video and text

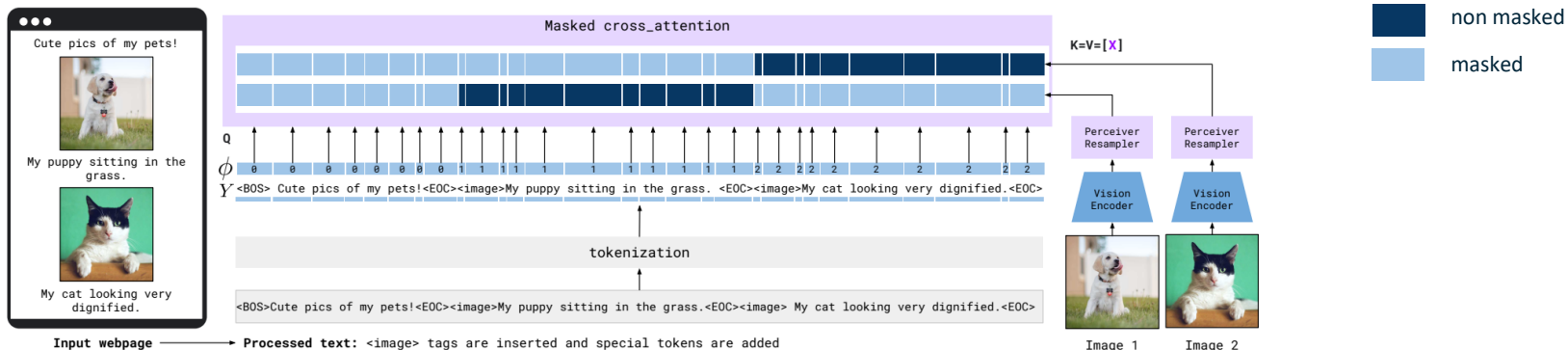


- Frozen LM layers
 - LM: 70B parameter Chinchilla
 - keep pretrained LM's language understanding
- Gated Cross Attention:
 - Query: Y, Key=Value: X
 - Tanh Gating: Initialized with 0 then gradually increases
 - Transitions from a fully trained text-only model to a visual language model.
- The LM can generate text conditioned on the above visual tokens



Flamingo VLM

- Model structure - Interleaved visual data and text support

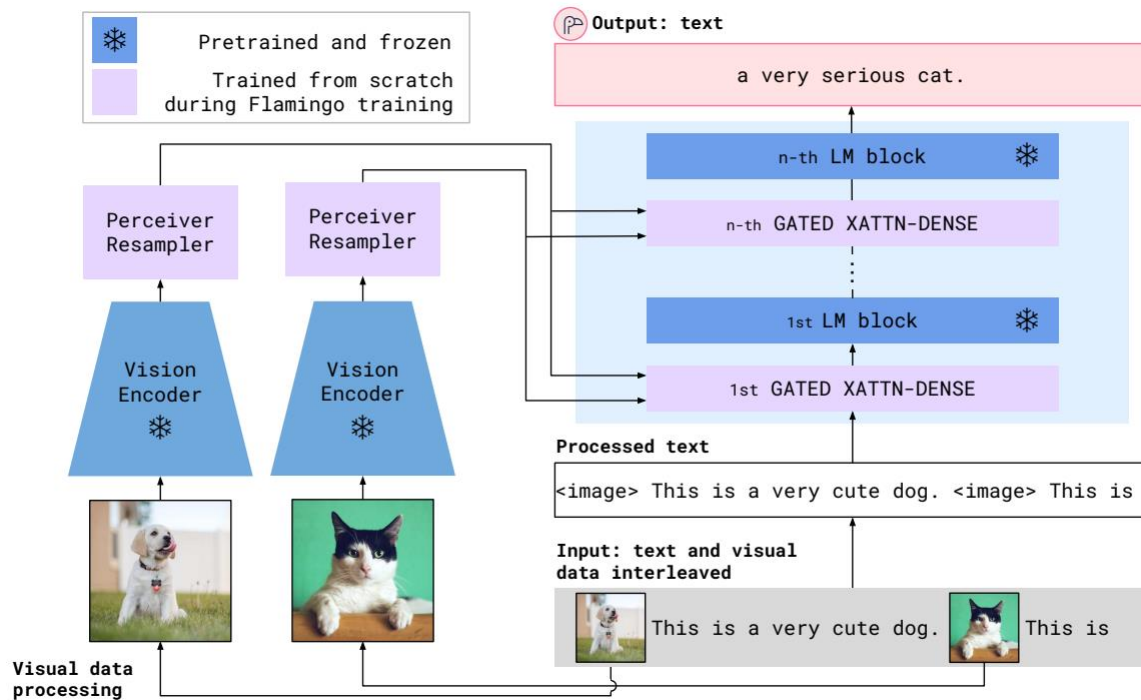


- Multi-visual input support: per-image/video attention masking
- During Cross-attention,
 - each text can only focus on one image before it.
 - Function ϕ : for each token what is the index of the last preceding image
- During final prediction,
 - each token can focus on all the previously text and image



Flamingo VLM

- Overview of the Flamingo Model



- Each image is encoded individually

Slide by Azade Farshad and Mei Sun



Flamingo VLM

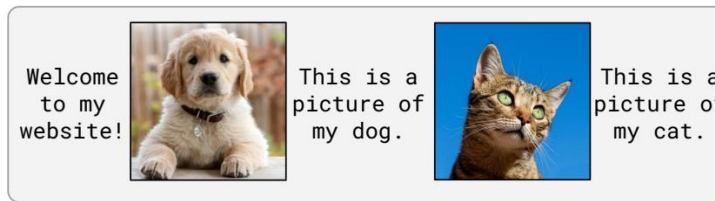
- Model structure - **Obtaining multimodal dataset to induce good generalist capabilities**



Image-Text Pairs dataset
[N=1, T=1, H, W, C]



Video-Text Pairs dataset
[N=1, T>1, H, W, C]



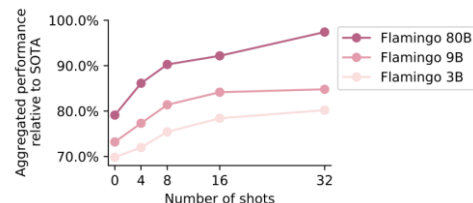
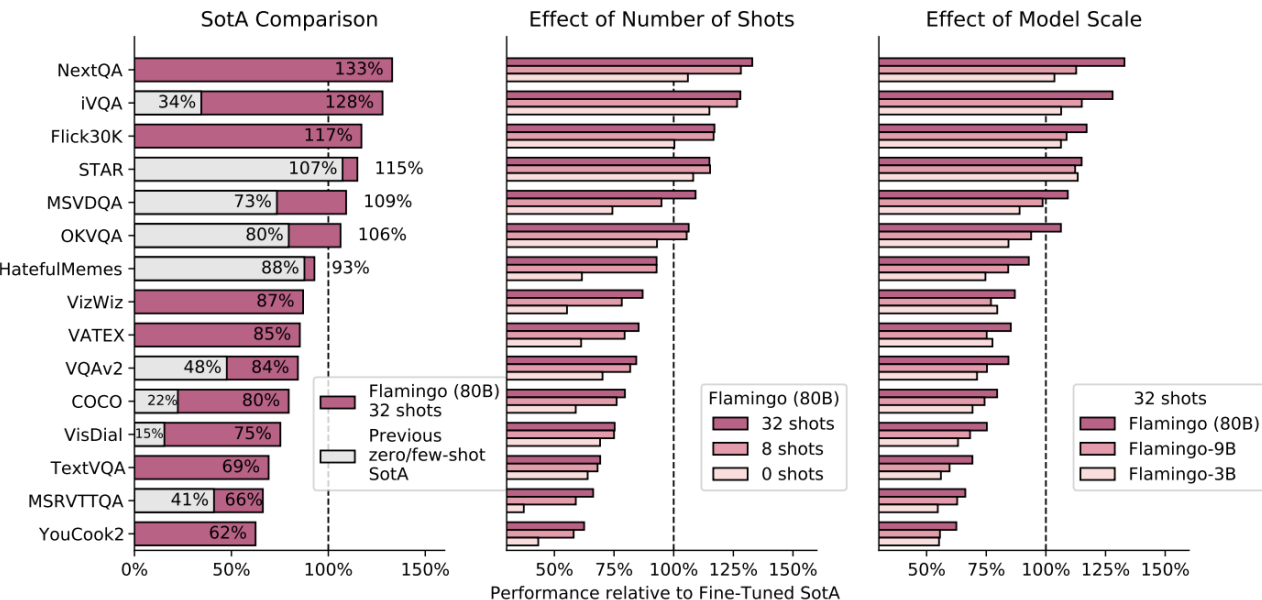
Multi-Modal Massive Web (M3W) dataset
[N>1, T=1, H, W, C]

- M3W: Scrapping 43 million webpages from the Internet
- Training on a mixture of vision and language datasets
 - M3W(185M images+ 182G text)
 - ALIGN(1.8B images with alt-text)
 - LTIP (312M images/text)
 - VTP(27M short video/text)



Flamingo VLM

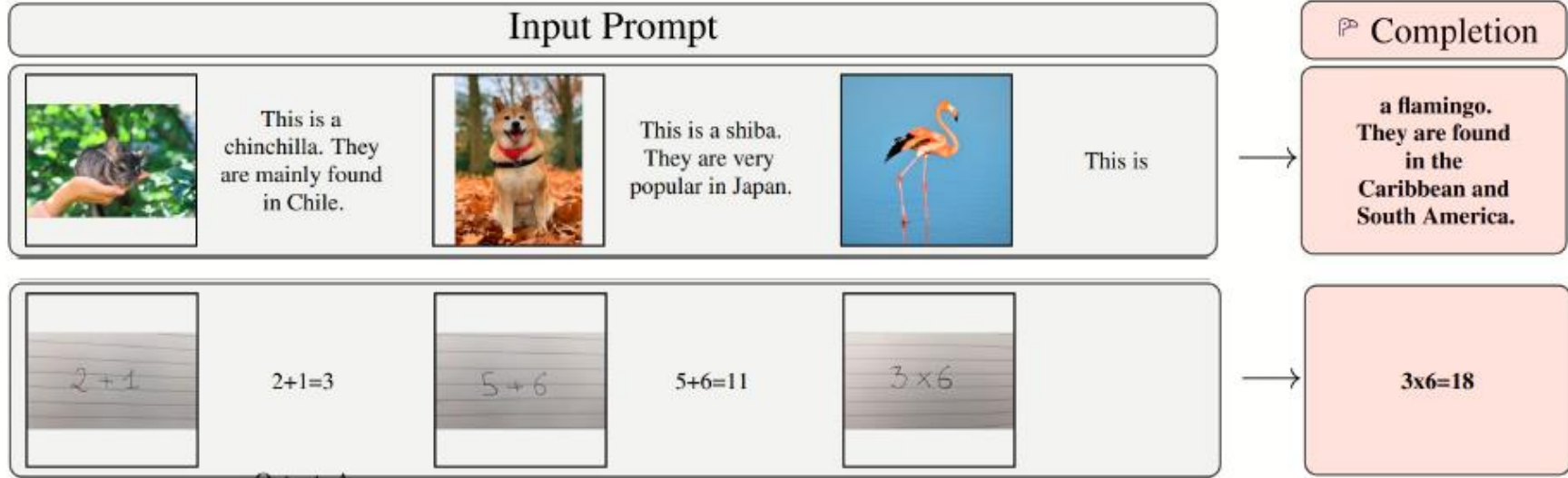
- Result: Overview of the results of the Flamingo models

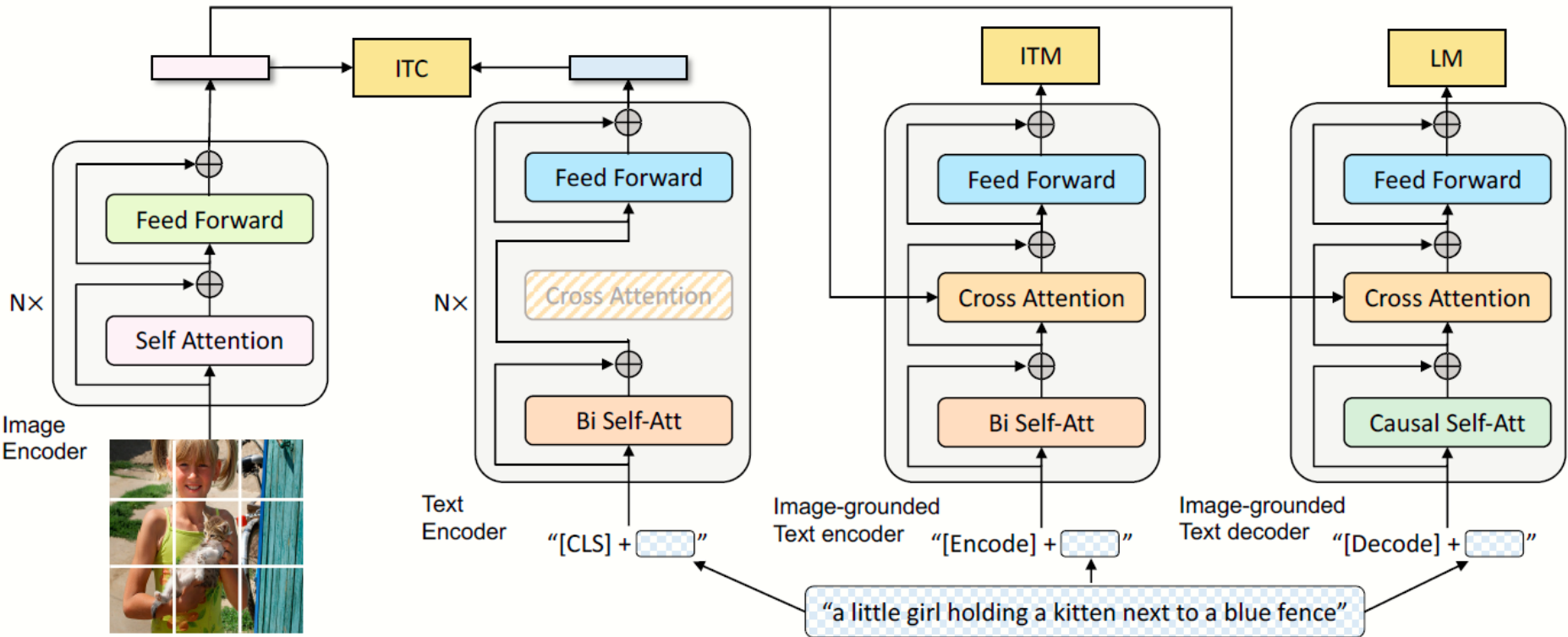


- Larger model sizes and more few-shot examples lead to better performance

- Performance of Flamingo model using different numbers of shots and of different sizes, (without fine-tuned) in comparison with SoTA fine-tuned baseline.

- Flamingo: Multimodal In-Context-Learning





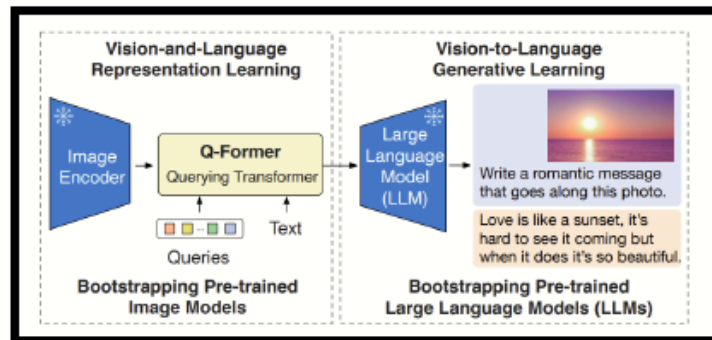
Li et al., BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation



BLIP



• BLIP2



Language Model

Connection Module

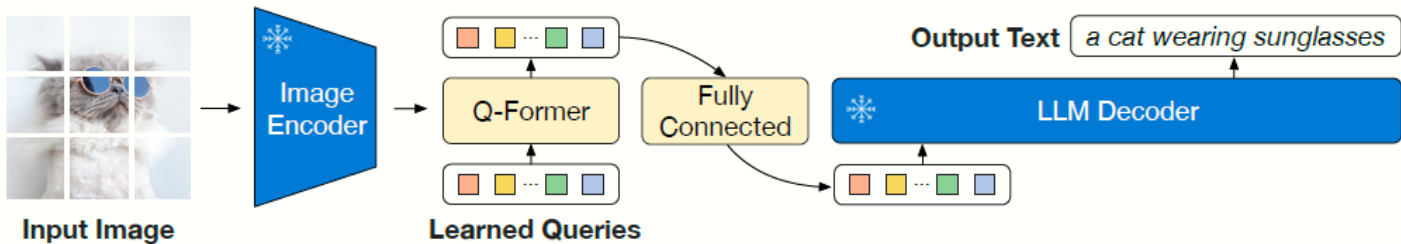
Vision Encoder

Pre-trained: FLAN-T5/OPT

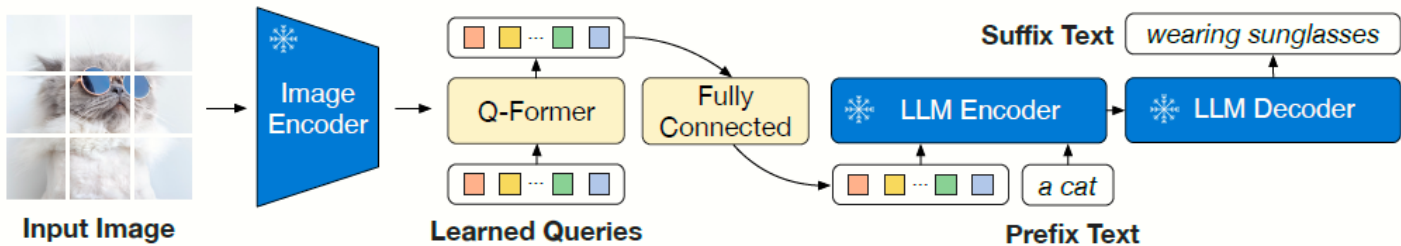
Q-Former: Lightweight
Querying Transformer

Contrastive pre-trained:
EVA/CLIP

Bootstrapping from a **Decoder-based** Large Language Model (e.g. OPT)



Bootstrapping from an **Encoder-Decoder-based** Large Language Model (e.g. FlanT5)



Li et al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

BLIP-2

Models	#Trainable Params	#Total Params	VQAv2		OK-VQA	GQA
			val	test-dev	test	test-dev
VL-T5 _{no-vqa}	224M	269M	13.5	-	5.8	6.3
FewVLM (Jin et al., 2022)	740M	785M	47.7	-	16.5	29.3
Frozen (Tsimpoukelli et al., 2021)	40M	7.1B	29.6	-	5.9	-
VLKD (Dai et al., 2022)	406M	832M	42.6	44.5	13.3	-
Flamingo3B (Alayrac et al., 2022)	1.4B	3.2B	-	49.2	41.2	-
Flamingo9B (Alayrac et al., 2022)	1.8B	9.3B	-	51.8	44.7	-
Flamingo80B (Alayrac et al., 2022)	10.2B	80B	-	56.3	50.6	-
BLIP-2 ViT-L OPT _{2.7B}	104M	3.1B	50.1	49.7	30.2	33.9
BLIP-2 ViT-g OPT _{2.7B}	107M	3.8B	53.5	52.3	31.7	34.6
BLIP-2 ViT-g OPT _{6.7B}	108M	7.8B	54.3	52.6	36.4	36.4
BLIP-2 ViT-L FlanT5 _{XL}	103M	3.4B	62.6	62.3	39.4	<u>44.4</u>
BLIP-2 ViT-g FlanT5 _{XL}	107M	4.1B	<u>63.1</u>	<u>63.0</u>	40.7	44.2
BLIP-2 ViT-g FlanT5 _{XXL}	108M	12.1B	65.2	65.0	<u>45.9</u>	44.7

Table 2. Comparison with state-of-the-art methods on zero-shot visual question answering.

Li et al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Visual Question Answering Examples

Language Modeling: Large Language Models (LLM)



GPT-2



GPT-3



ChatGPT
InstructGPT



GPT-4

What's new?

In-context-learning
Chain-of-thoughts (CoT)

In-context-learning
Chain-of-thoughts (CoT)
Instruction-Following

In-context-learning
Chain-of-thoughts (CoT)
Instruction-Following
Multimodal Input with image

Open Source
Community

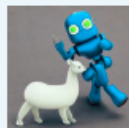
LLaMA



Alpaca



GPT4-Alpaca



Vicuna



Tulu

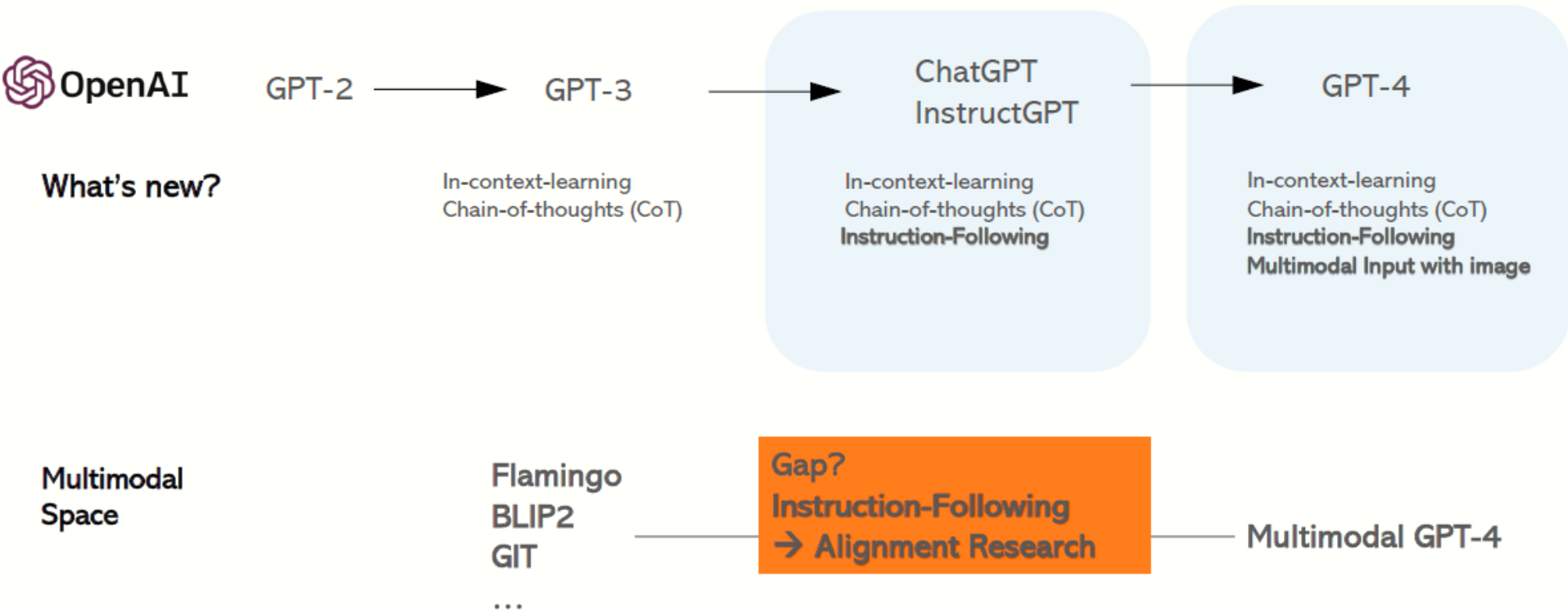


Language Models and Data

Slide by Chunyuan Li



Recap on Language Modeling: Large Language Models (LLM)



GPT4-V Gap

Instruction Tuning

Input → Output

Translation

Hello, Vancouver

你好, 温哥华

Summarization

CVPR is the premier annual computer vision event comprising the main conference and several co-located workshops and short courses. This year, CVPR will be single track such that everyone (with full passport registration) can attend everything.

CVPR: top computer vision event, single-track, accessible to all.

- Task instructions are implicit.
- Individual models are trained, or multi-tasking without specifying the instructions
- Hard to generalize to new tasks in zero-shot

In Language: Various NLP Task Datasets

Instruction Tuning

Instruction

Translate English into Simplified Chinese

Summarize in just 10 words to make the message even more brief and easier to remember.

Input →

Output

Hello, Vancouver

你好, 温哥华

CVPR is the premier annual computer vision event comprising the main conference and several co-located workshops and short courses. This year, CVPR will be single track such that everyone (with full passport registration) can attend everything.

CVPR: top computer vision event, single-track, accessible to all.

- Task instructions are explicit, expressed in natural language
- One single model is trained, multi-tasking with specified instructions
- Natural and easy to generalize to new tasks in zero-shot



In Language: Instruction Tuning

Slide by Chunyuan Li

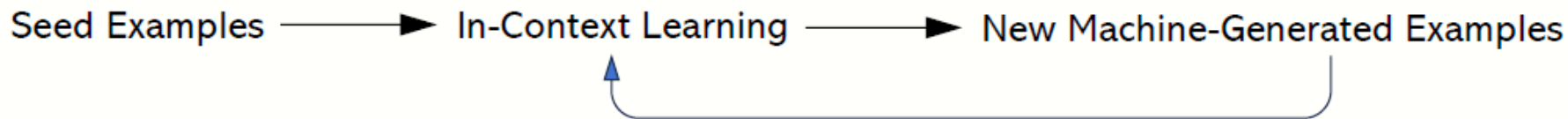


How to collect a diverse set of high-quality instructions and their responses?

- ❑ Human-Human: Collected from humans with high cost
- ❑ Human-Machine: A Strong LLM Teacher such as GPT3 and GPT4






translation example *summarization example*

Please generate new instructions that meet the requirements:



Instruction Tuning with Open-Source LLMs

Self-Instruct with Strong Teacher LLMs & Mixed Human Data

	LLaMA 	Alpaca 	Vicuna 	GPT4-Alpaca 	...	Tulu 
Data Source		GPT-3.5	ShareGPT (Human & GPT)	GPT-4 (text-only)	...	Mixed Data
Instruction- following Data (#Turns)	None	52K	500K (~150K conversions)	52K	...	

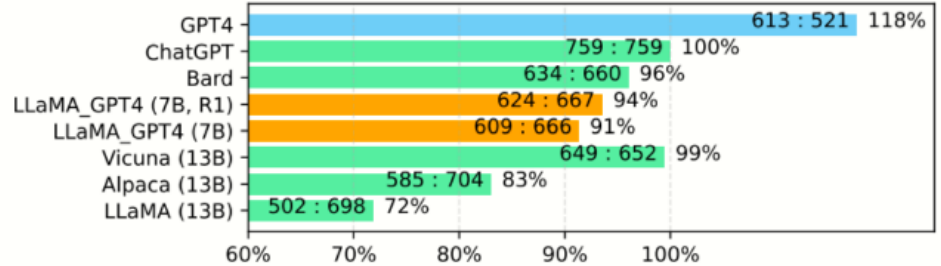
Results on Chatbot

Evaluation Metric: Ask GPT-4 to rate the two model responses (1-10), then compute the ratio, i.e. relative score

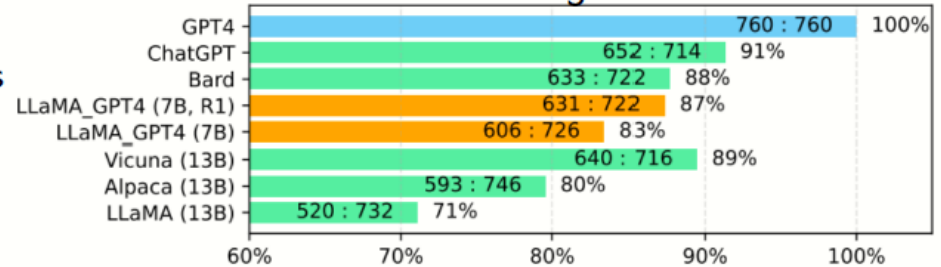
Findings:

- A VERY CONSISTENT Evaluation Metric !
- Opensourced Chatbots mimicked commercial ones

All chatbots against ChatGPT



All chatbots against GPT-4



It works!

Haotian Liu*, Chunyuan Li*, Qingyang Wu, Yong Jae Lee (* Equal contribution)

Self-Instruct with Strong Teacher LLMs

But No Teacher is available on multiGPT4?

	LLaMA	Alpaca	Vicuna
Teacher			
		GPT-3.5	ShareGPT (Human & GPT)
Instruction-following Data	None	52K	700K (70 conversions)

GPT-4-LLM



GPT-4
(text-only)

LLaVA



GPT-4
(text-only)

- 158K multimodal instruction following data
(First & High Quality)

————▶ Multimodal Chatbot

Large Language and Vision Assistant

21

- Rich Symbolic Representations of Images
- In-context-learning with a few manual examples

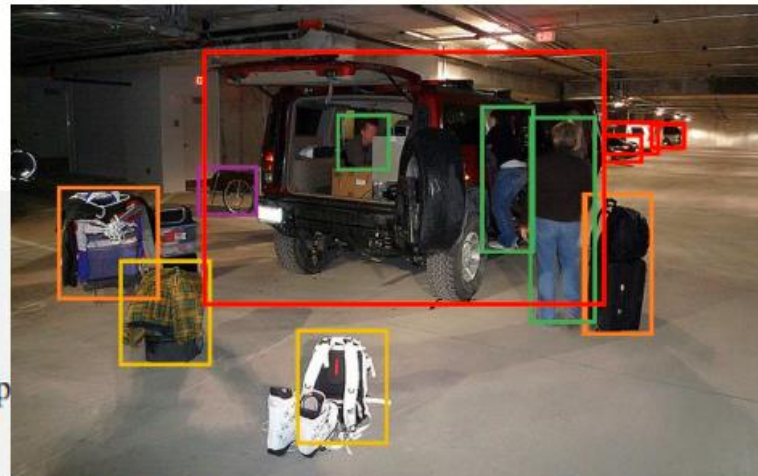
→ Text-only GPT-4

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]



GPT-assisted Visual Instruction Data Generation

Three type of instruction-following responses

Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV).

Question: Where is the vehicle parked?

Answer: The vehicle is parked in an underground parking area, likely in a public garage.

Question: What are the people in the image doing?

Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.

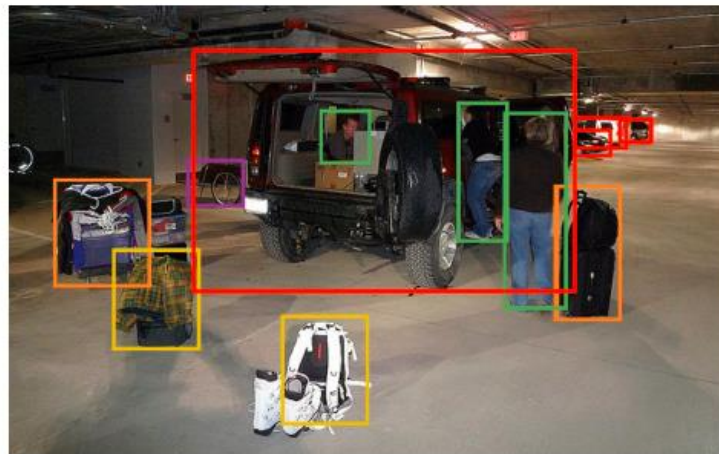
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.

Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Response type 3: complex reasoning

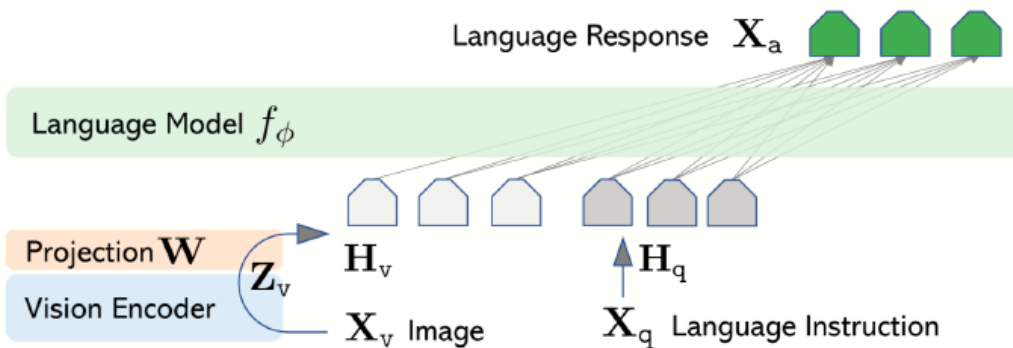
Question: What challenges do these people face?

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



LLaVA: Large Language-and-Vision Assistant

Architecture



Two-stage Training

•Stage 1: Pre-training for Feature Alignment.

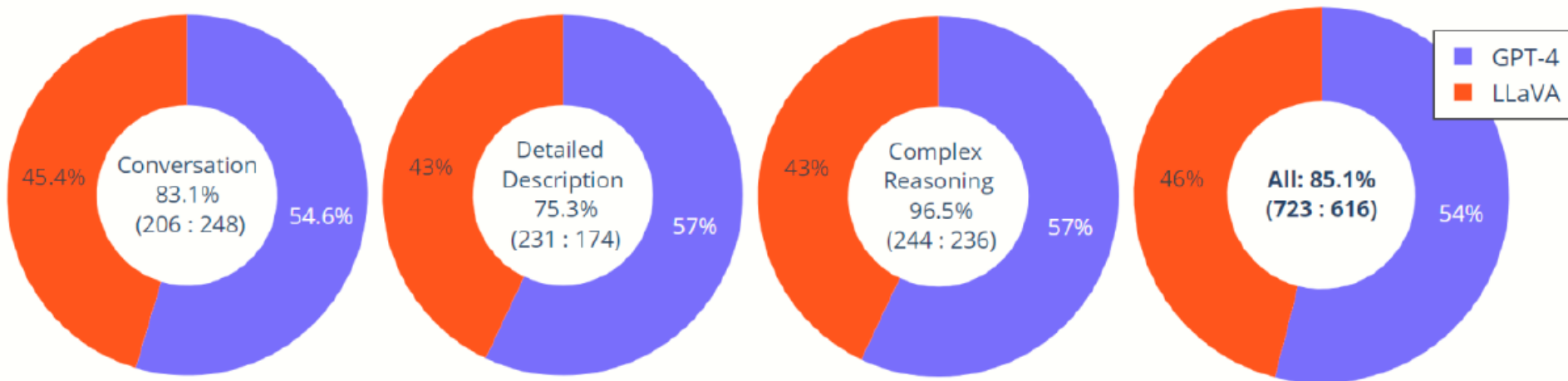
Only the projection matrix is updated, based on a subset of CC3M.

•Stage 2: Fine-tuning End-to-End. Both the projection matrix and LLM are updated

•**Visual Chat:** Our generated multimodal instruction data for daily user-oriented applications.

•**Science QA:** Multimodal reasoning dataset for the science domain.

Visual Chat: Towards building multimodal GPT-4 level chatbot

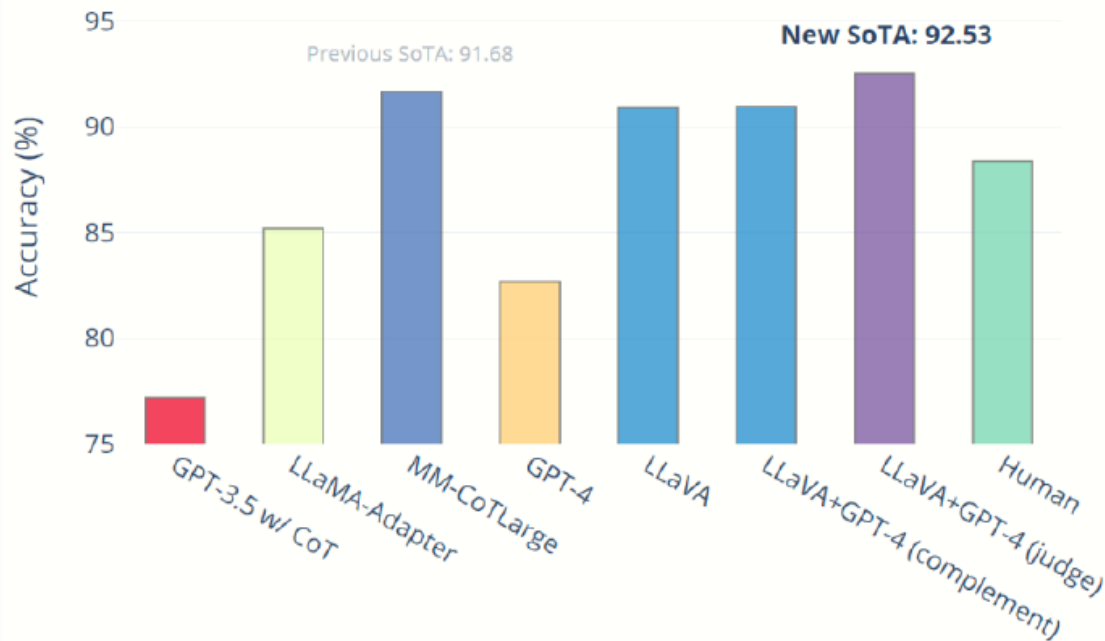


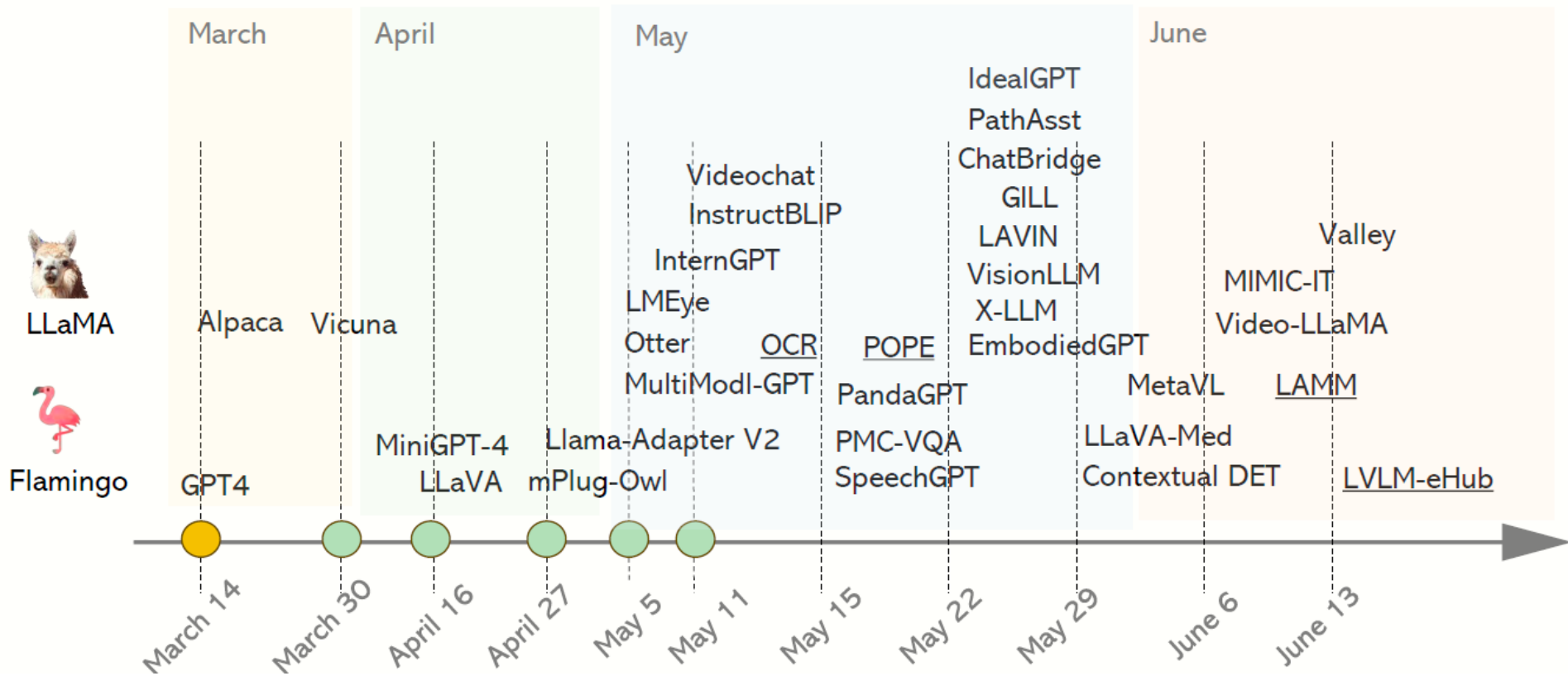
An evaluation dataset with 30 unseen images, 90 new language-image instructions

Overall, LLaVA achieves 85.1% relative score compared with GPT-4

Science QA: New SoTA with the synergy of LLaVA with GPT-4

- LLaVA alones achieve 90.92%
- We use the text-only GPT-4 as the judge, to predict the final answer based on its own previous answers and the LLaVA answers.
- This "GPT-4 as judge" scheme yields a new SOTA 92.53%
- GPT-4 is an effective model ensemble method

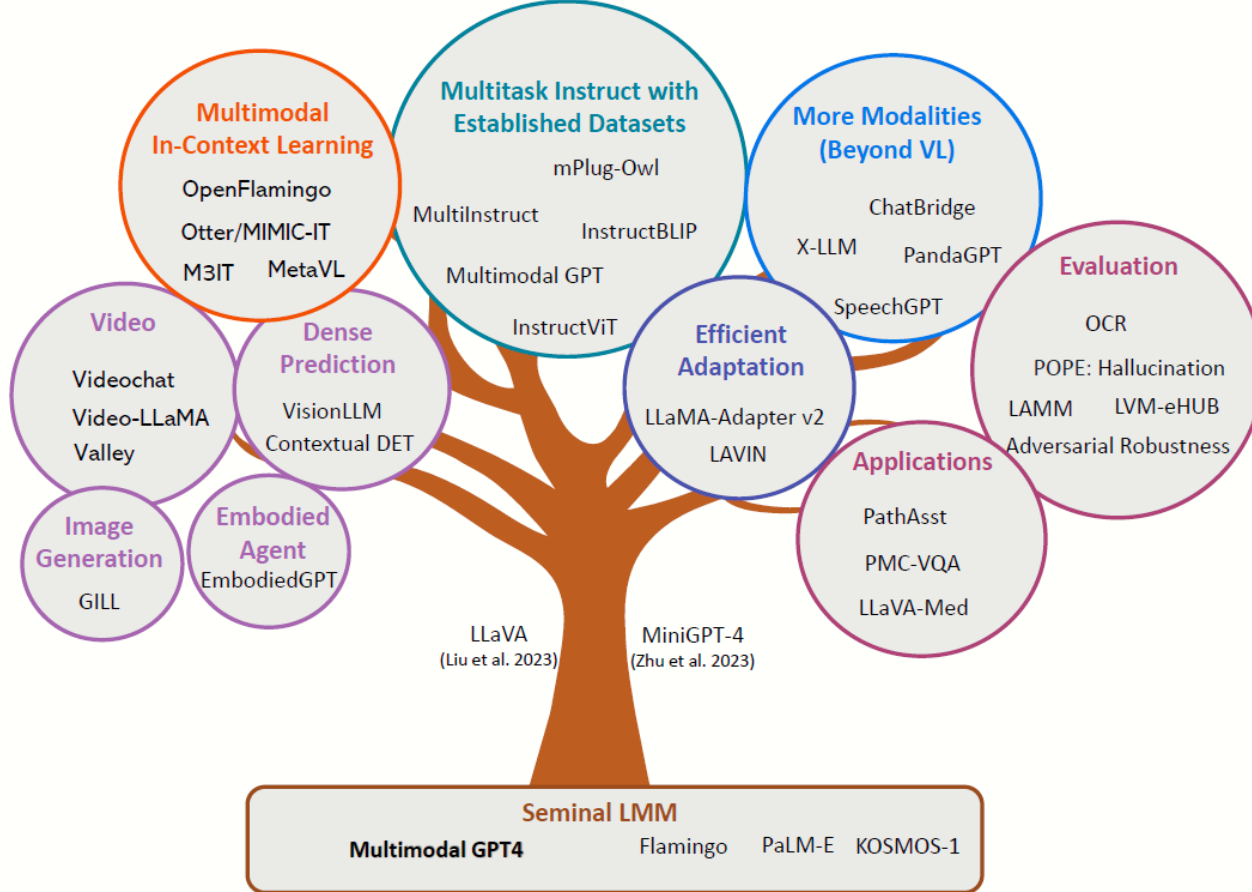




Vision-Language Explosion (2023)

Slide by Chunyuan Li





Vision-Language Explosion (2023)

Slide by Chunyuan Li

1) Cross-Modal Retrieval

Audio



Crackle of a Fire

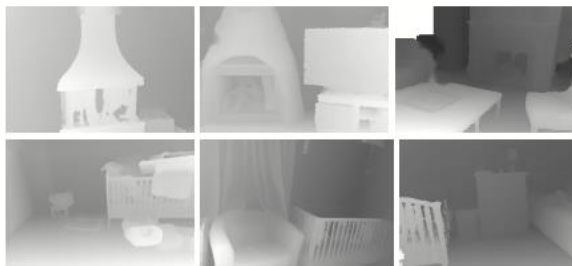


Baby Cooing

Images & Videos



Depth



Text

“A fire crackles while a pan of food is frying on the fire.”

“Fire is crackling then wind starts blowing.”

“Firewood crackles then music...”

“A baby is crying while a toddler is laughing.”

“A baby is laughing while an adult is laughing.”

“A baby laughs and something...”

2) Embedding-Space Arithmetic



Waves



3) Audio to Image Generation



Dog



Engine



Fire



Rain

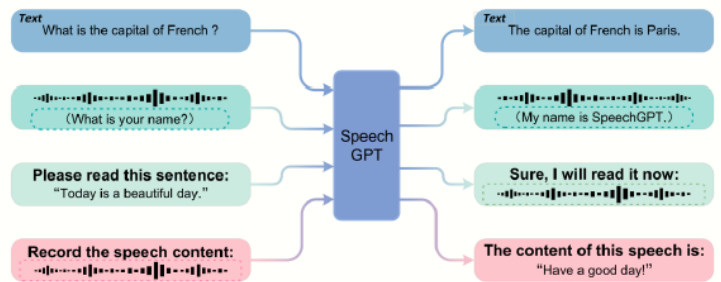
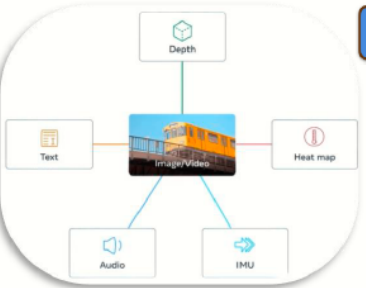
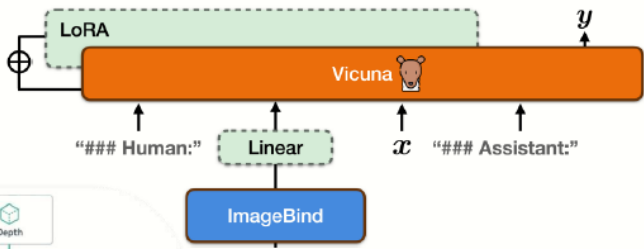
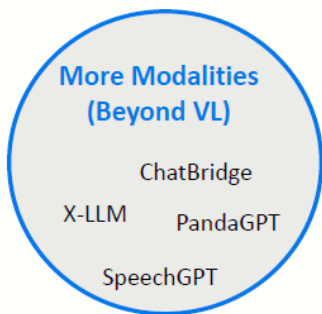


Girdhar et al., ImageBind: One Embedding Space To Bind Them All

More Modalities: ImageBind

More Modalities (Beyond VL)


- ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst
- PandaGPT: One Model To Instruction-Follow Them All
- SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities
- X-LLM: Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages



Adding Modalities

MM-ReAct



wait for user  input
image as <path>

Yang et al., MM-ReAct MM-ReAct: Prompting ChatGPT for Multimodal Reasoning and Action

MM-ReAct

Visual Math and Text Reasoning



What is "a" in the first equation?
What is y in the last equation?

$5 + a = 7$	$40 + y = 45$
$10 + y = 15$	$39 + a = 41$
$20 + a = 22$	$24 + y = 29$
$48 + a = 50$	$15 + a = 17$
$10 + a = 12$	$34 + y = 39$

The value of "a" in the first equation " $5 + a = 7$ " is 2.

The value of "y" in the last equation " $34 + y = 39$ " is 5.



Visual-Conditioned Joke/Meme



What is funny about it?



The humor in the comic may come from the situation, the characters, or the dialogue. The person is feeling unwell and is asked if they are pregnant or have the coronavirus, which is a current and relevant topic. The humor may come from the absurdity of the situation or the unexpectedness of the question.



Spatial/Coordinate Understanding



Where is the frisbee?
What is the relative position of the frisbee and the man?



The frisbee is located at coordinates (63,119,133,177).

The frisbee is located above the man in the image. The man is falling on the ground while the frisbee is in the air. The frisbee is also closer to the top of the image than the man..



Visual Planning and Prediction



What should I do after kneading the dough, before dividing it into pieces? For how long?



After kneading the dough, you should put it in a warm place before dividing it into pieces, according to the recipe in the image.

You should put the dough in a warm place for 1 hour.



Multi-Image Reasoning



How much in total did I spend on travel?
How much in total did I pay for taxes?

The Shop	
Subtotal	\$43.83
Tax	\$321.90
Total	\$365.73

You spent \$1299.90 on airfare and \$43.83 on an Uber ride for a total of \$1343.73 on travel.

You paid \$323.23 in total for taxes. This includes \$321.90 for taxes and fees for the flight purchase and \$1.33 for sales tax for the Windmill Family Restaurant.



Multi-Hop Document Understanding



Excluding the regular cab, what is the average total fueling cost?



The average total fueling cost excluding the Ford F150 (Regular Cab) is \$76.55. This is calculated by taking the sum of the total fueling cost of Toyota RAV4 and Honda Civic and dividing it by 2. $(\$82.56 + \$70.55) / 2 = \$76.55$.



Open-World Concept Understanding



What is the brand of these drink? Which one is unsweetened?



The brand of these drinks is "BETTER THAN MILK". The unsweetened one is "BETTER THAN MILK organic OAT DRINK".



Who are the celebrities in this image?



Keira Knightley and Natalie Portman are the celebrities in this image.



Is it safe to eat this mushroom?



Morel mushrooms are generally considered safe to eat when properly cooked.



Can you breakdown this video tutorial to step-by-step instructions, with start and end time?



Here's a step-by-step breakdown of the tutorial with start and end time.



... Add Lettuce Add Tomato Add Bacon ...



◆ Vision+Language (and multi-modal) are hot!

◆ Why?

- ◆ Align various interface modalities
- ◆ Leverage more data (all modalities)
- ◆ Physical world inherently multi-modal

◆ Large number of design choices!

- ◆ Vision encoding?
- ◆ Method of alignment?
- ◆ Method of fusion?

◆ Tasks:

- ◆ Image ↔ language
- ◆ Visual question answering
 - ◆ + Interaction
- ◆ Embodied AI

Resources:

<https://www.youtube.com/@VLPTutorial>