# โลกสองใบ: Binding Language-Image

ผมสร้างโลก 2 ใบ
สำหรับ Language Model
และ Vision Model
เขาไม่รู้ว่ามันมีโลก multimodal อีกใบหนึ่ง
เขาดำเนินชีวิตในแบบคนรักที่ดีทั้งคู่
ผมโกหกเขา เพื่อให้โลกทั้ง 2 ใบนี้
ยังอยู่กับผมได้

**ป๊อป ปองกูล ไม่ได้กล่าว**

**Saksorn Ruangtanusak (Ha**
AI Research - TILDI

# Who is He ->

A Random Guy?

# Who is Saksorn (Harry?)

**Nuclear Physics Researcher & Robotics Engineer**
@ B.S. in Mechanical Engineering (2018 – 2021)

**AI Engineer (Gold Medal)**
@ SuperAI Engineer SS2

**2D/3D Computer Vision (Data Scientist)**
@ PTTEP ARV Bedrock (June 2022 - Mar 2024)

**LLM Researcher**
@ CJ More (TILDI) (Mar 2024 - Present)

**Hacker who wins 1st in hackathon**

- Typhoon Hack 2024 - SCB10x created ***personalized meeting summarizer***
- Bangkok AI Hack 2023 - SCB10x created ***Financial Adviser ChatBot***
- The Dispatcher 2023 - ARV created ***Anti-drone system***

**Skills:** Python, SQL, PyTorch, Langchain, HF, AWS, GCP, Docker

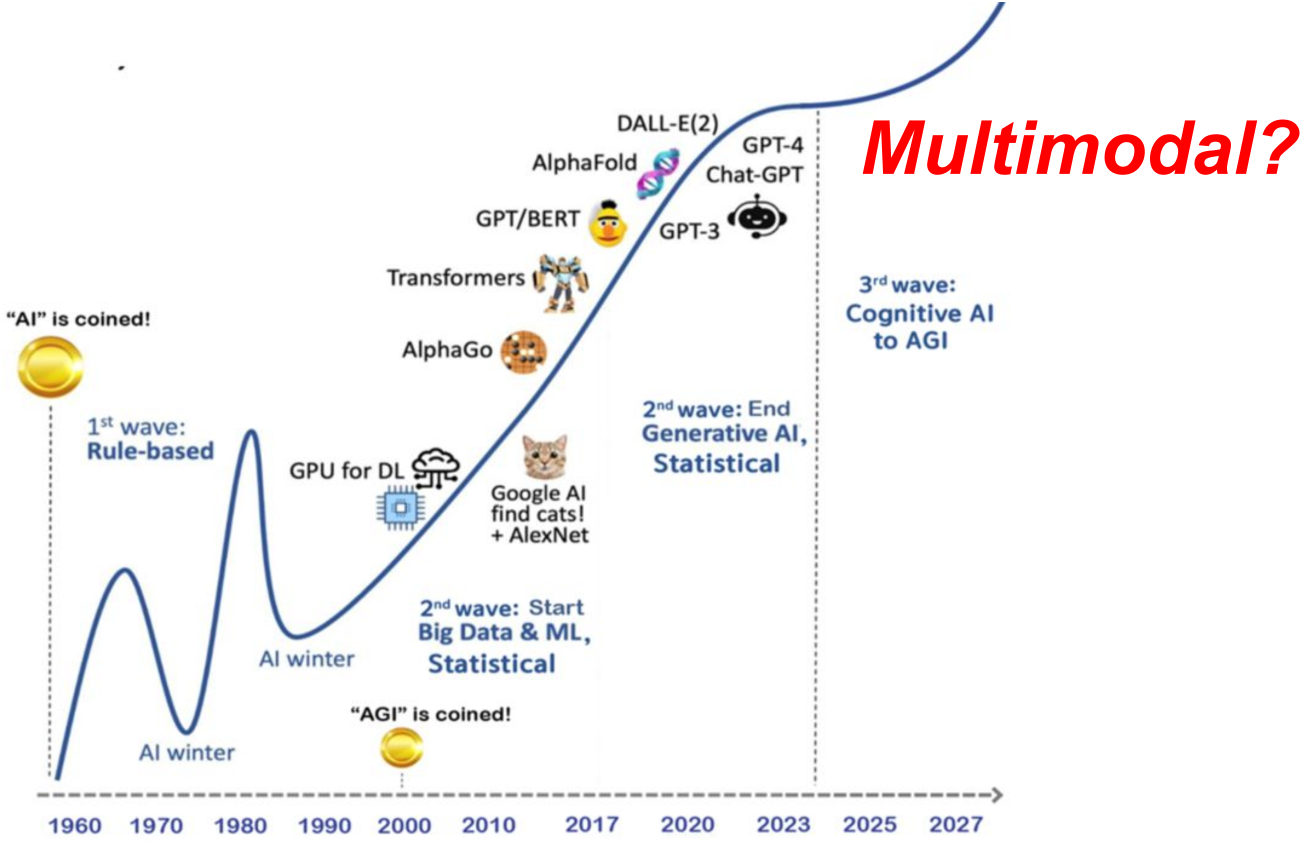

## Publication

1. **IEEE ICCI 2024** (TH) in NLP with BERT
2. **PHYSOR 2022** (US) in Nuclear Physics
3. **IEEE ECTI 2022** (TH) in Optimization.

## Volunteer

1. Coach @ **Super AI Engineer SS4** 2024
2. TA @ **Google Build with AI Day** 2024
3. Technical Staff @ **CUD Hackathon** 2023
4. TA @ **Super AI Engineer SS3** 2023

# Timeline of AI



# Let take a look!

With Multimodal Model ~~

# Look Cool right?

This can enable lots of use cases. More than just Text.

# Introduction to multimodal

Multimodal models are advanced AI systems designed to comprehend, interpret, and generate information across various data formats, including text, images and audio.

What is somtam?

**Multimodal Model**

Somtam, also known as papaya salad, is a popular and flavorful Thai dish made from shredded unripe green papaya. It is typically served as a salad or a side dish and is renowned for its distinctive combination of sweet, sour, salty, and spicy flavors.

A bustling local convenience store in Thailand called CJ, with colorful signage and a variety of products advertised, is a hub of activity as motorcycles and bicycles line the parking area in front.

I like somtam and cj express.

# Task in Multimodal Model

Image Retrieval

Image Captioning

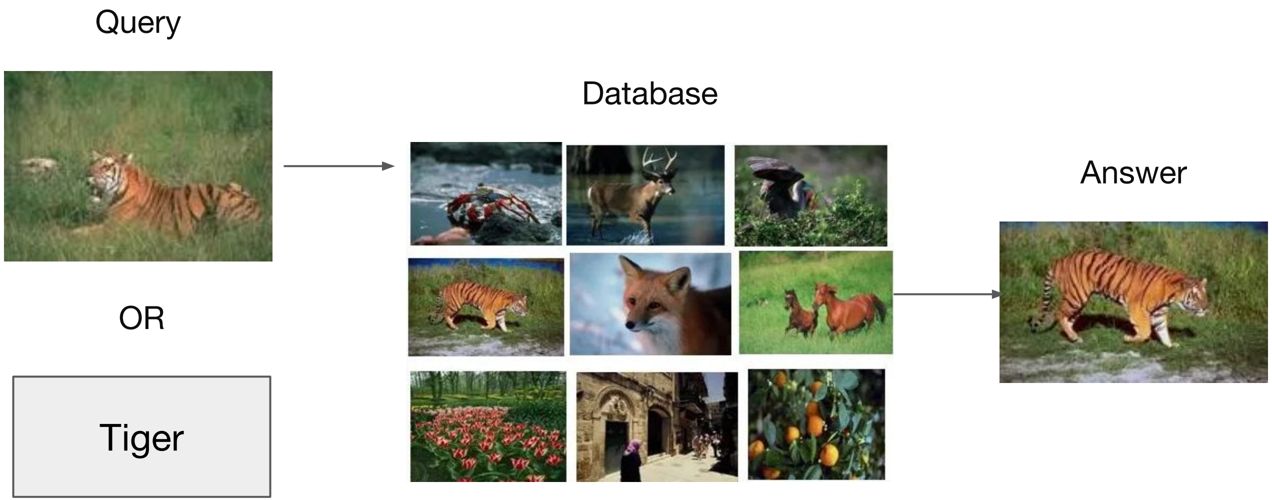
Visual Question Answering

????

# Level 1:
# Image Retrieval

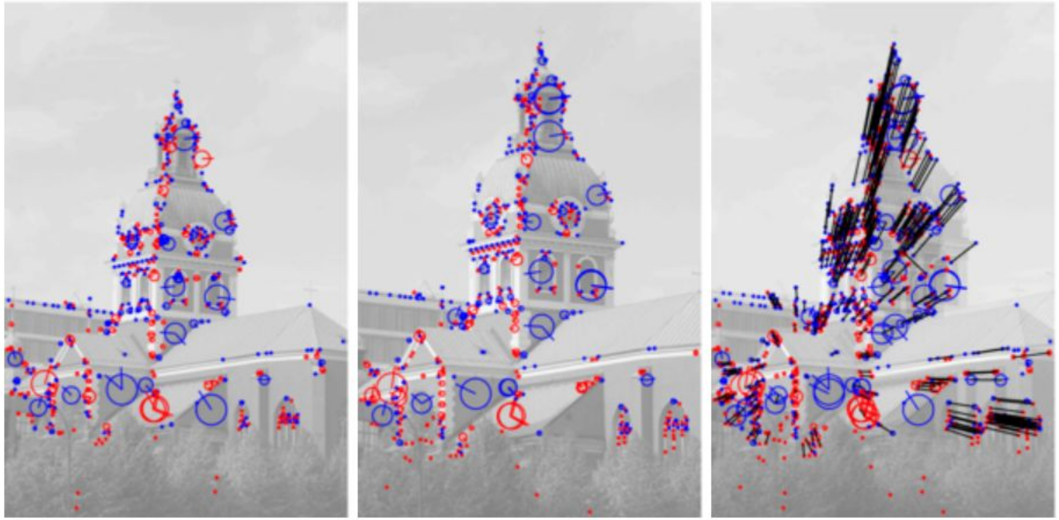
# Level 1 : Image Retrieval

Finding relevant images based on a text query or finding relevant text based on an image query.



# Traditional Feature-Based Methods

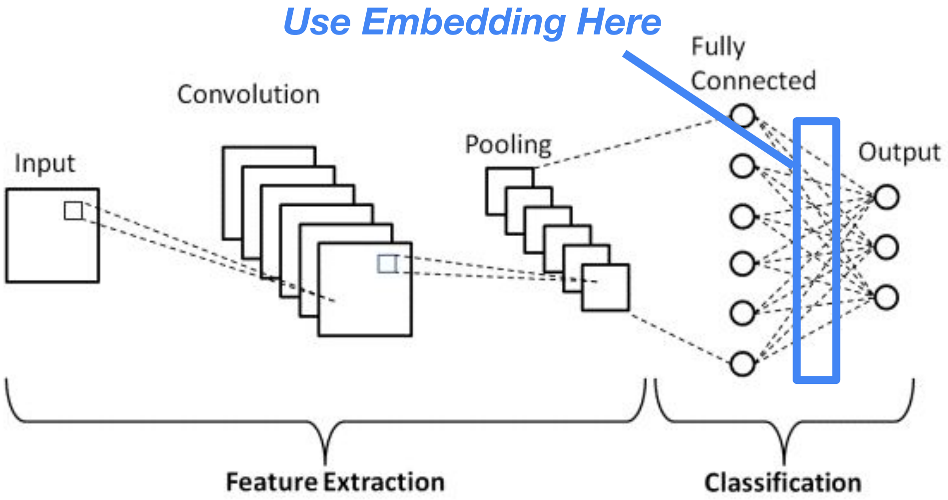
Good old days before deep learning



- **SIFT (Scale-Invariant Feature Transform):**
  - detects and describes local features in images.
- **SURF (Speeded Up Robust Features):**
  - An improvement over SIFT (Speed)
  - simplifying the computation with approximate Gaussian smoothing.

## Break through Image Feature Extraction

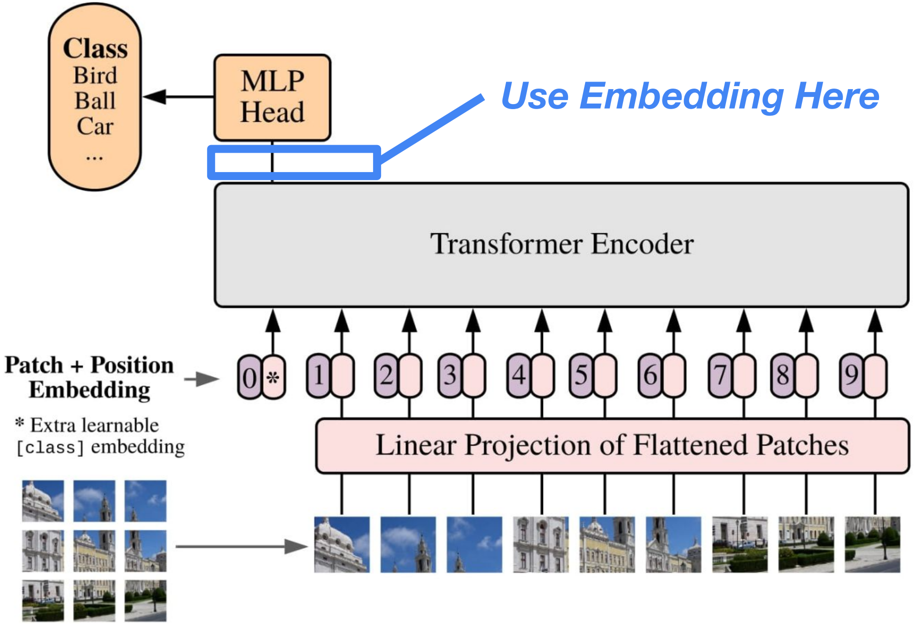
# CNN: Convolutional Neural Networks




- **VGGNet** (VGG16, VGG19)
  - CNN that very efficient to tackle ImageNet
- **ResNet** (ResNet50, 101)
  - Use Residual learning
  - Make model deeper.
- **EfficientNet**
  - uniformly scales network depth/width/resolution
  - using a compound coefficient

## Break through Image Feature Extraction

# ViT: Vision Transformer




- **OG ViT**
  - Turn image into patch
  - Encode patch and position
  - More scalable than CNN
- **BEiT**
  - BERT like pre-training
  - Use MLM (Mask Language Model)
  - Hide some token then predict that token
- **SwinTransformer**
  - Shifted window method
  - Likely CNN + ViT

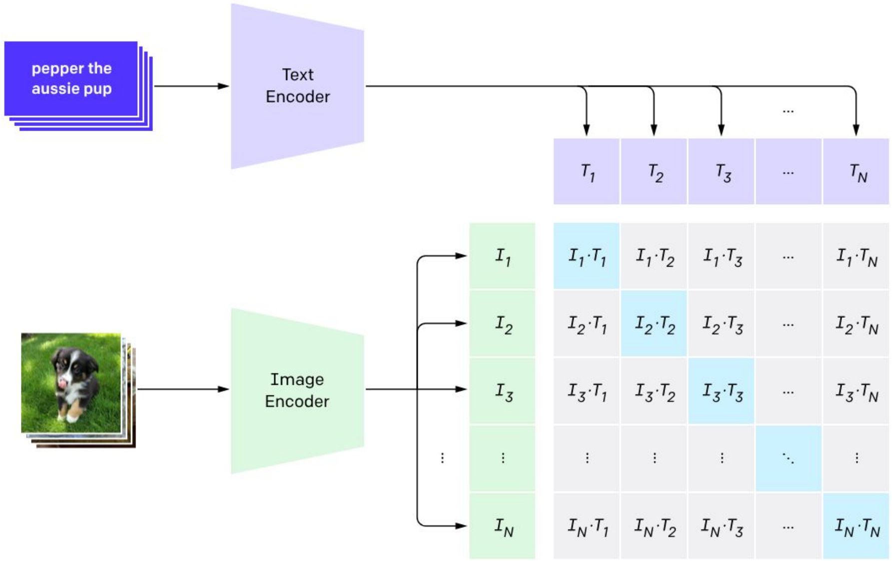
**But these are not good enough for Image retrieval.**

We need something bigger and add some NLP.

*Level 1: Image Retrieval*

Break through Language-Image Pretraining

# CLIP: Connecting Language Image Pre-training




## Why it so good??

**Contrastive Pre-Training:** contrastive learning to align images with textual descriptions effectively.

**Zero-Shot Learning:** Ability to generalize to new tasks without additional training.

**Large-Scale and Diverse Dataset:** Training on diverse internet-sourced data enhances without intensive labeling. (400M Sample)

**Multimodal Understanding:** Unified framework for processing and understanding both images and text.

**Task Versatility:** Applicable to a wide range of tasks, often with superior performance.

*Level 1: Image Retrieval*

## Break through Language-Image Pretraining

# CLIP: Generalized much more than imageNet

| Dataset | ImageNet ResNet101 | CLIP ViT-L |
|---|---|---|
|  ImageNet | 76.2% | 76.2% |
|  ImageNet V2 | 64.3% | 70.1% |
|  ImageNet Rendition | 37.7% | 88.9% |
|  ObjectNet | 32.6% | 72.3% |

Although both models have the same accuracy on the ImageNet test set, CLIP's performance is much more representative of how it will fare on datasets that measure accuracy in different, non-ImageNet settings.

*Level 1: Image Retrieval*

Break through Language-Image Pretraining

# CLIP: Zero Shot Capability

**Inference API**

Zero-Shot Image Classification

Possible class names (comma-separated)

gold fish, carp, salmon, tuna

Compute

Computation time on cpu: cached

| | |
|---|---|
| gold fish | 0.997 |
| carp | 0.003 |
| salmon | 0.000 |
| tuna | 0.000 |

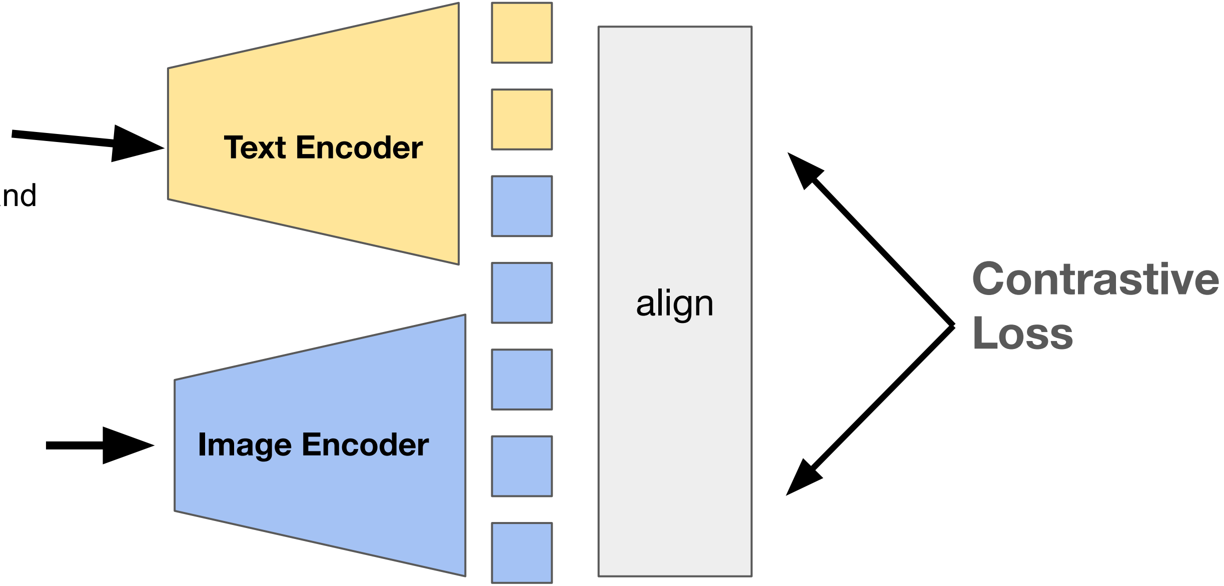
JSON Output

Maximize

https://huggingface.co/openai/clip-vit-large-patch14

# Level 1: Image Retrieval (Wrap Up)

A bustling local convenience store in Thailand called CJ, with colorful signage and a variety of products advertised, is a hub of activity as motorcycles and bicycles line the parking area in front.



# Level 2: Image Captioning

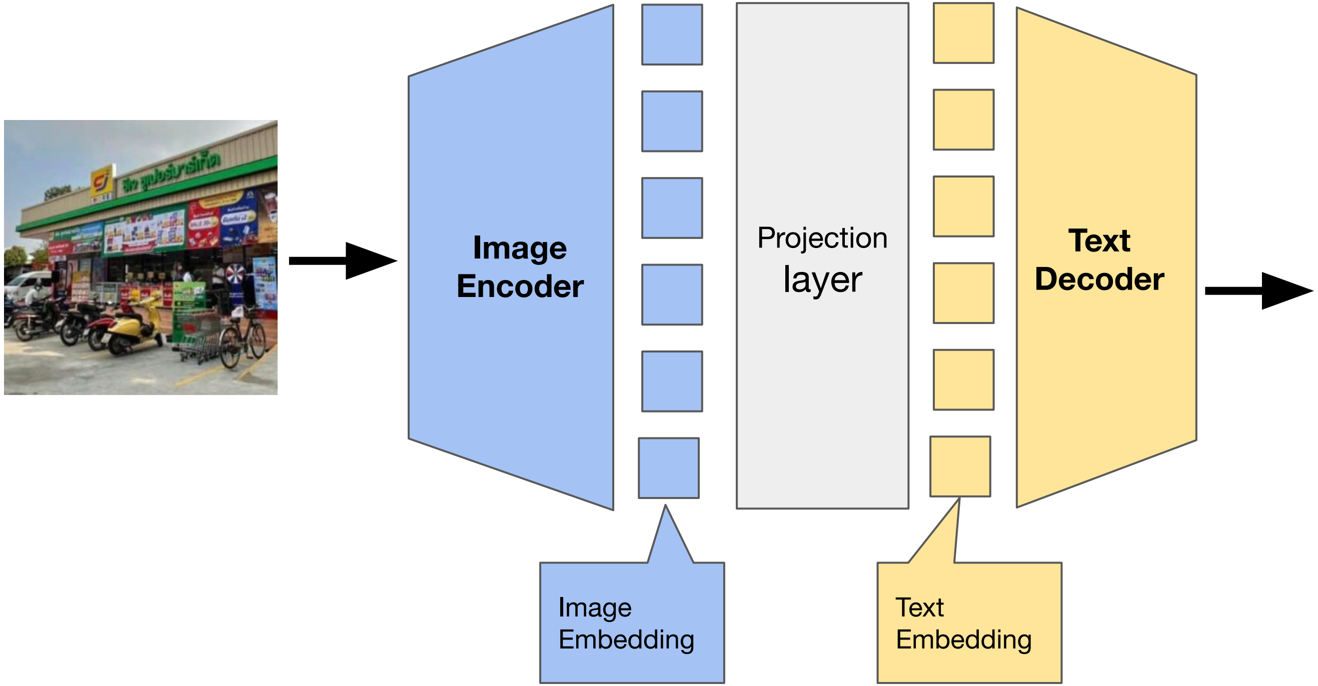
# Level 2: Image Captioning

Generating descriptive text for an image, understanding objects, actions, and context.

a woman riding a horse
with a crowd of people.

several full baskets of
different colored apples.

# Level 2: Image Captioning




A bustling local convenience store in Thailand called CJ, with colorful signage and a variety of products advertised, is a hub of activity as motorcycles and bicycles line the parking area in front.

*Level 2: Image Captioning*

# ClipCap: CLIP Prefix for Image Captioning

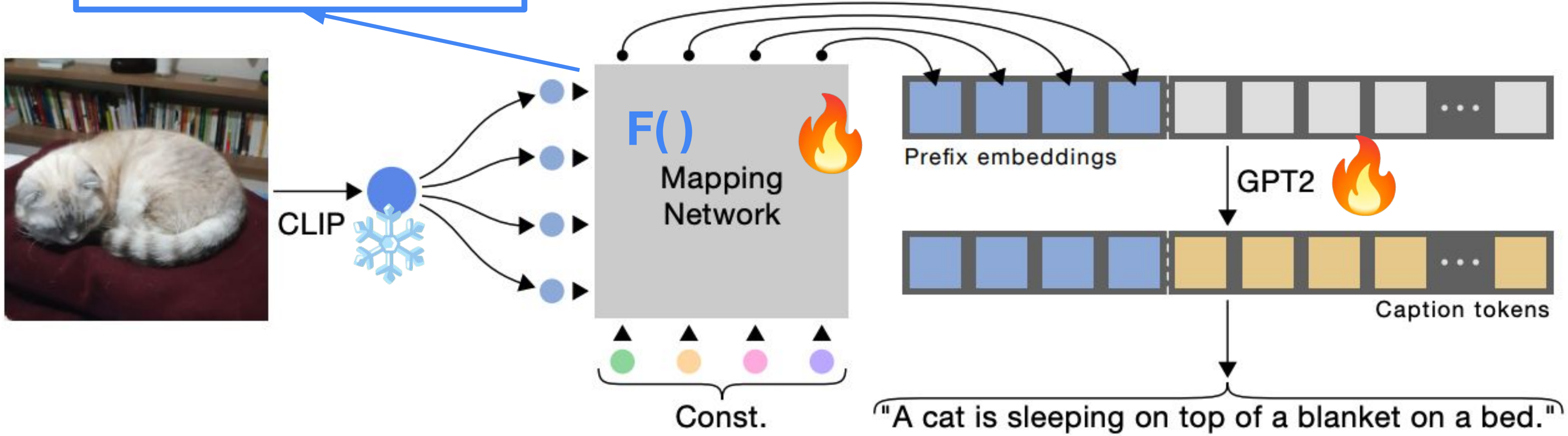
Fast training with single GPU!, Utilized pre-trained model both En, De.

**F = MLP or Cross-Attn**

$$p_1^i, \dots, p_k^i = F(\mathrm{CLIP}(x^i)).$$

**Concat with text embedding**

$$Z^i = p_1^i, \dots, p_k^i, c_1^i, \dots, c_\ell^i.$$



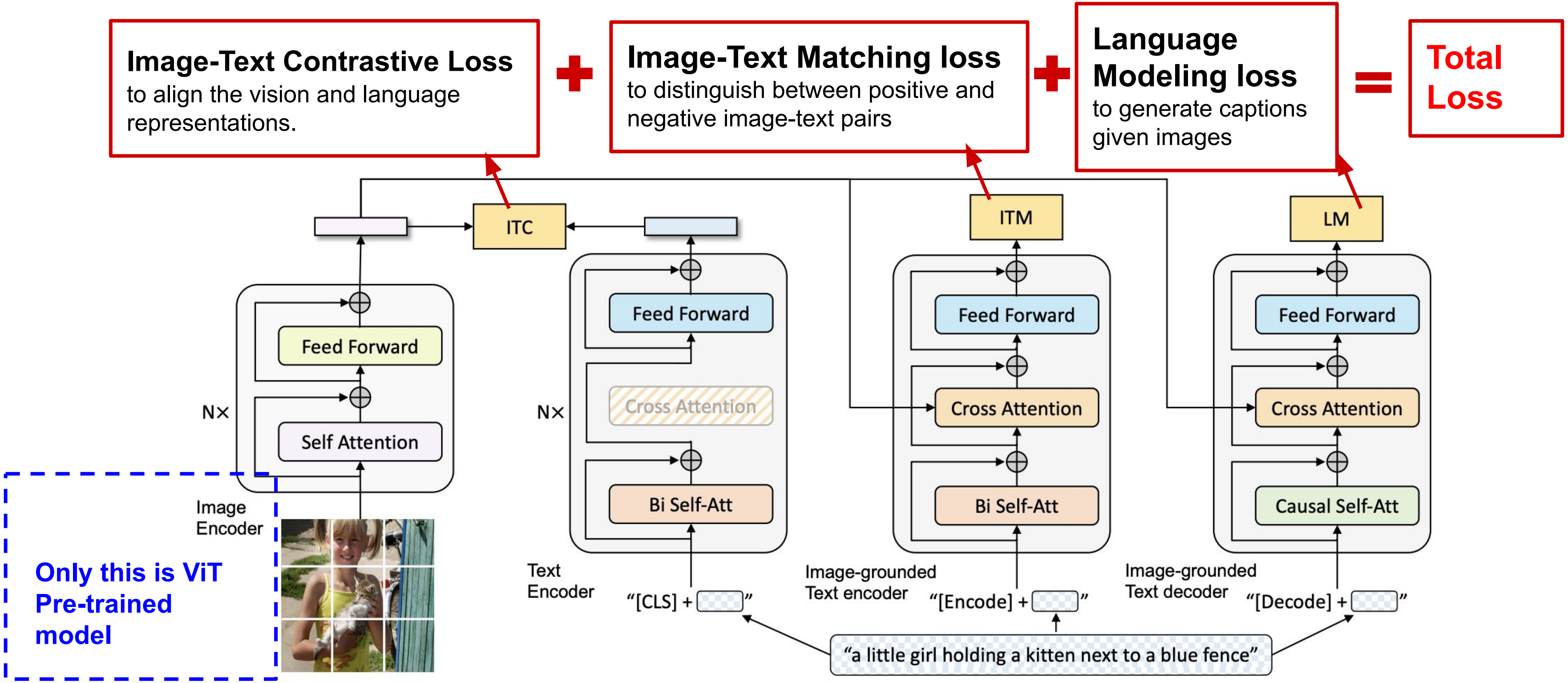

**Predict next word with cross-entropy loss**

$$\mathcal{L}_X = -\sum_{i=1}^{N}\sum_{j=1}^{\ell} \log p_\theta(c_j^i \mid p_1^i, \dots, p_k^i, c_1^i, \dots, c_{j-1}^i).$$

*Level 2: Image Captioning*

# **BLIP:** Bootstrapping Language-Image Pre-training

Large pre-trained for many task. (Retrieval, Captioning) And can FT for VQA, Video Understanding

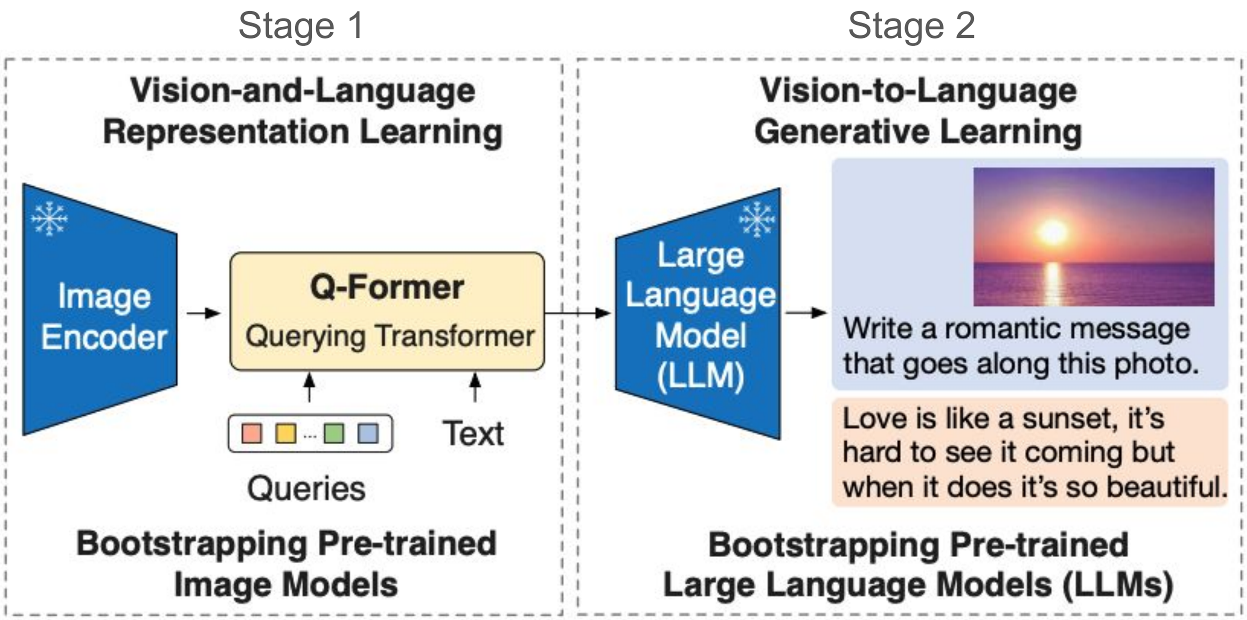


But we have pre-trained LLMs.

**Why not utilized it?**

# BLIP-2: Bootstrapping Language-Image Pre-training

Utilized Frozen Image Encoders and Large Language Models instead of training from scratch.

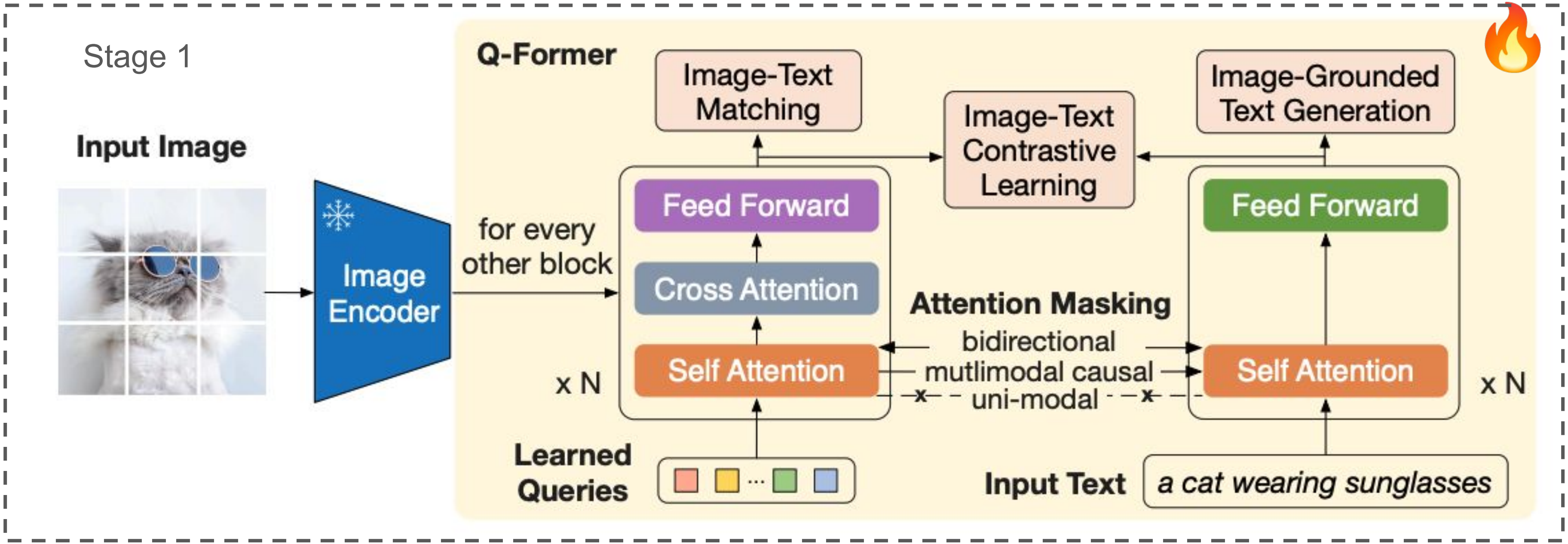



Overview of BLIP-2's framework.

*Level 2: Image Captioning*

# BLIP-2: But what is a Q-Former?

In first-stage pre-training using Extract Visual Representation.

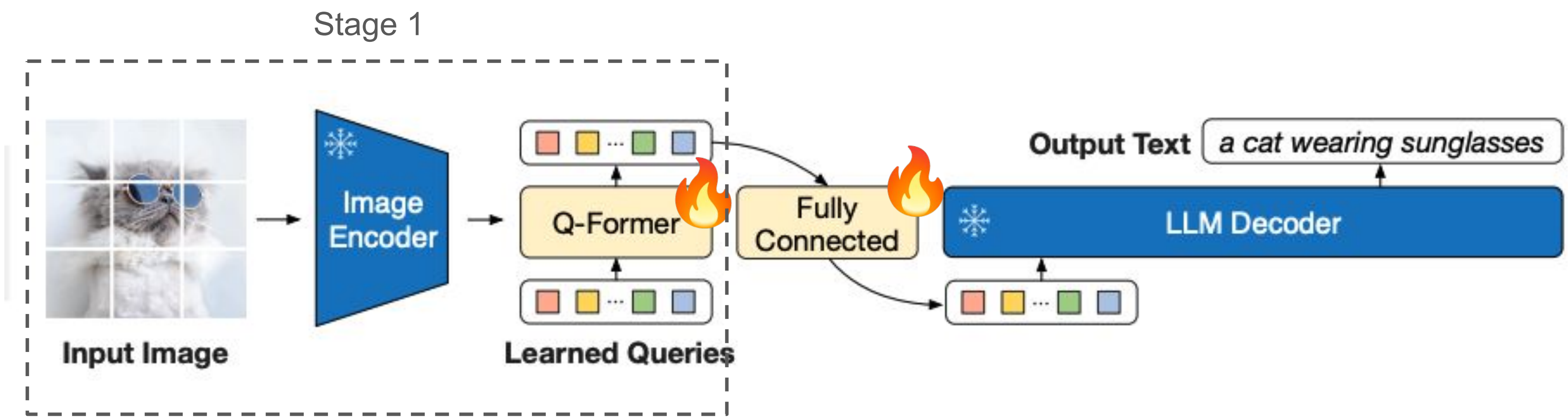



We jointly optimize three objectives which enforce the queries (a set of learnable embeddings) to extract visual representation most relevant to the text.

*Level 2: Image Captioning*

# BLIP-2: How it learn to caption?

In second-stage pre-training using Image-Ground Text Generation Loss.




The fully-connected layer adapts from the output dimension of the Q-Former to the input dimension of the chosen LLM.

*Level 2: Image Captioning*

# Compare all ClipCap, BLIP, BLIP-2

In NoCaps Datasets with CIDEr metrics (Mean Similarity from multiple-reference)

| models | #Trainable Params | in-domain | near-domain | out-of-domain | Overall |
|---|---|---|---|---|---|
| ClipCap | 43M | 84.85 | 66.82 | 49.14 | 65.83 |
| BLIP | 446M | 114.9 | 112.1 | 115.3 | 113.2 |
| BLIP-2 ViT-g OPT2.7B | 1100M | 123.0 | 117.8 | 123.4 | 119.7 |
| BLIP-2 ViT-g OPT6.7B | 1100M | 123.7 | 119.2 | 124.4 | 121.0 |

# Level 3: Visual Question Answering

# Visual Question Answering

**Can you explain this meme in detail?**

**Classify each image in the grid.**

**MODEL**

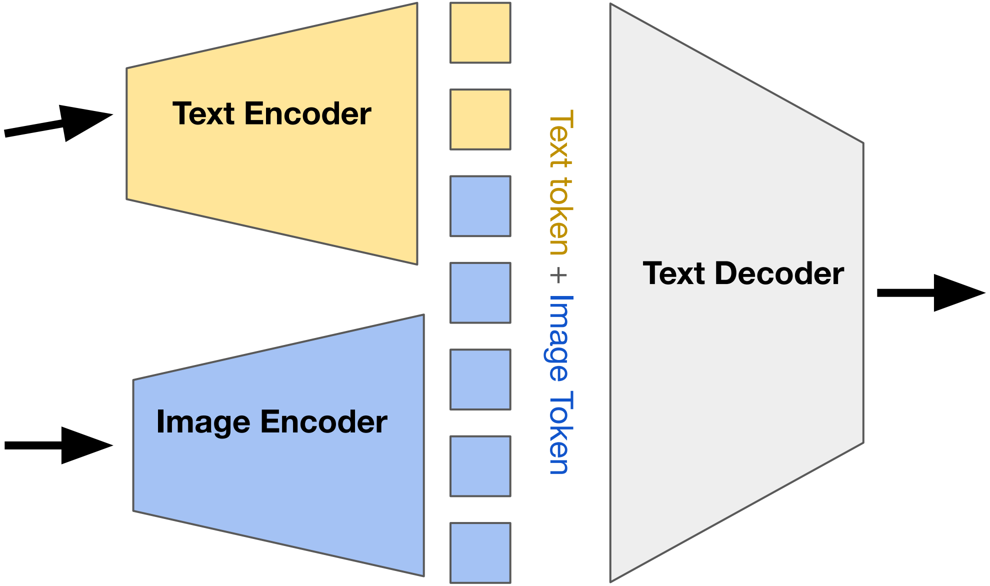
This meme is a visual joke that plays on the similarities between the faces of Chihuahuas and blueberry muffins.

**Row 1:**
Muffin/Dog/Muffin/Dog

**Row 2:**
Dog/Muffin/Dog/Muffin

# VQA Dive

**Caption this images?**

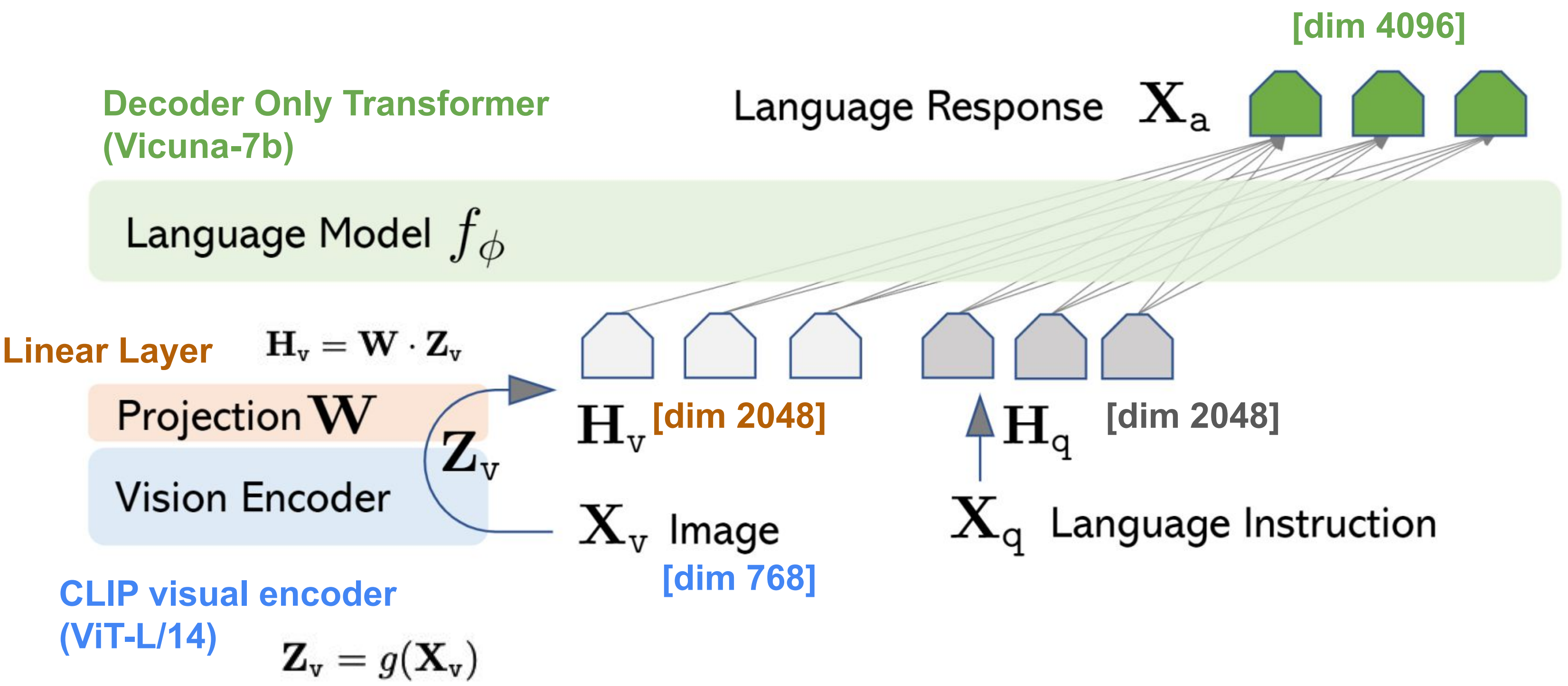



A bustling local convenience store in Thailand called CJ, with colorful signage and a variety of products advertised, is a hub of activity as motorcycles and bicycles line the parking area in front.

*Level 3: VQA*

# LLaVA: Large Language and Vision Assistant

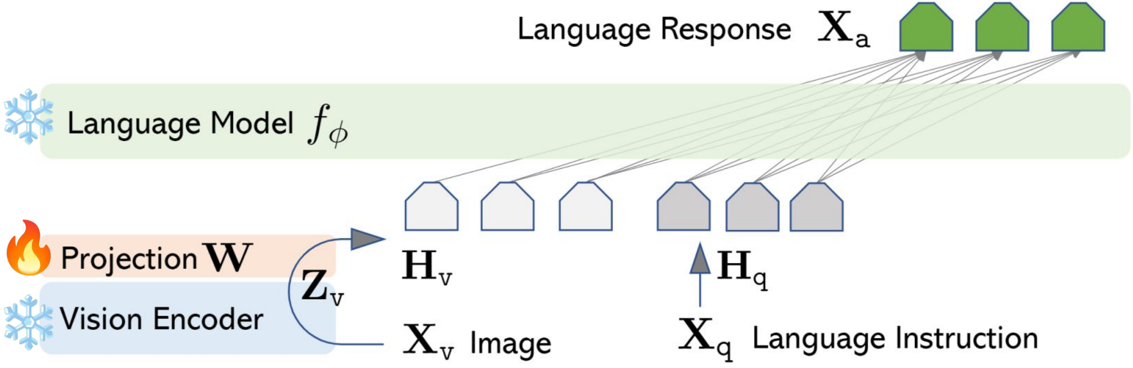
Visual Instruction Tuning!!!! And all the model following this method (Phi-3 vision too)



# Training LLaVA

## Stage 1: Pre-training for Feature Alignment

- Image captioning task single training.
- Use 595K image-text pairs (Subset of LAION-CC-SBU).
- Keep both the visual encoder and LLM weights frozen
- Maximize the likelihood of W (the projection matrix)




$\mathbf{X}_{\text{system-message}}$ <STOP>
Human : $\mathbf{X}^1_{\text{instruct}}$ <STOP> Assistant: $\mathbf{X}^1_a$ <STOP>

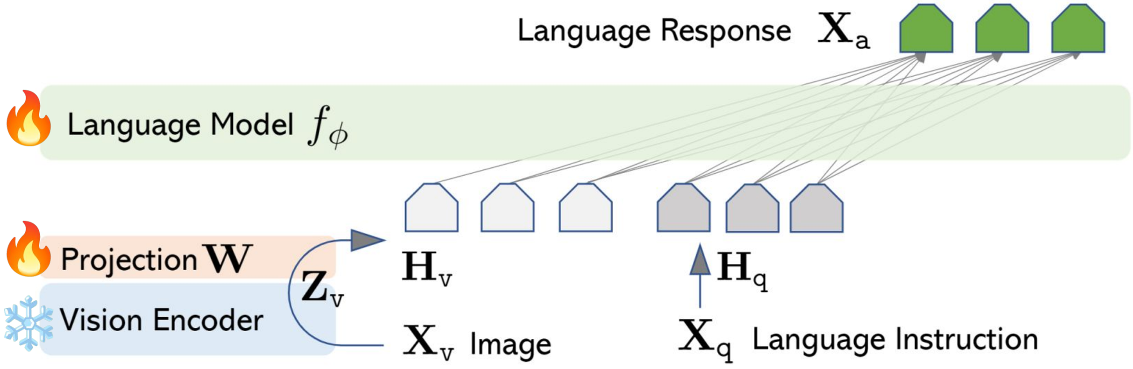
*Input sequence to train model (Single Turn Captioning).*

*Level 3: VQA*

# Training LLaVA

## Stage 2: Fine-tuning End-to-End

- Keep the visual encoder weights frozen.
- fine-tuning on the 158K language-image instruction-following multi-turn
- Maximize the likelihood of W and f-phi




$\mathbf{X}_{\text{system-message}}$ <STOP>
Human : $\mathbf{X}^1_{\text{instruct}}$ <STOP> Assistant: $\mathbf{X}^1_a$ <STOP>
Human : $\mathbf{X}^2_{\text{instruct}}$ <STOP> Assistant: $\mathbf{X}^2_a$ <STOP> $\cdots$

*Input sequence to train model (Multi Turn - Multi Task).*

Level 3: VQA

# Result of LLaVA

## Visual Reasoning

**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. |

*Level 3: VQA*

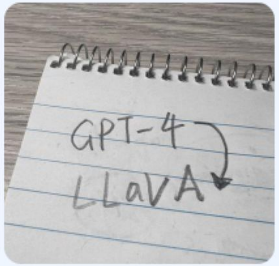
# Result of LLaVA

## Optical character recognition (OCR)

Who will be interested in the content in this image?


Individuals interested in the content of this image might include those who are familiar with the city of Vancouver, Canada, or are interested in traveling there. Additionally, people involved in the technology or computer industry, as well as those interested in artificial intelligence and its applications, may find the image intriguing. This is because the image features an attractive design with a blue logo that has the word "Vancouver" on it, suggesting a connection to the city or a product related to it. The presence of a laptop in the image might also connect to the tech or computer industry, possibly indicating a Vancouver-based company or an event taking place in the city.

what is written in the photo?

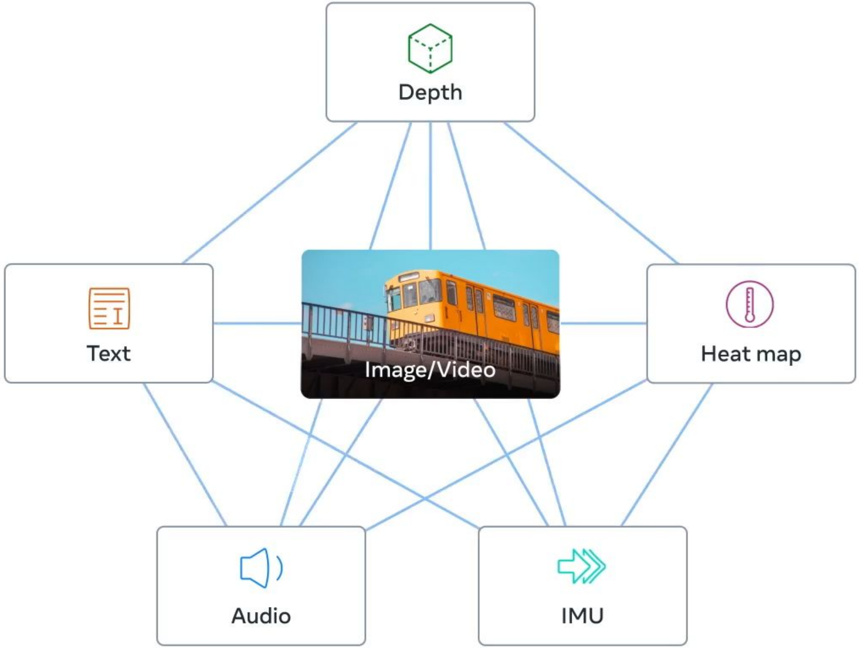



In the photo, there is written "Llava" which is presumably an abbreviation for a location or other identifier.

# Level 4: Beyond Images

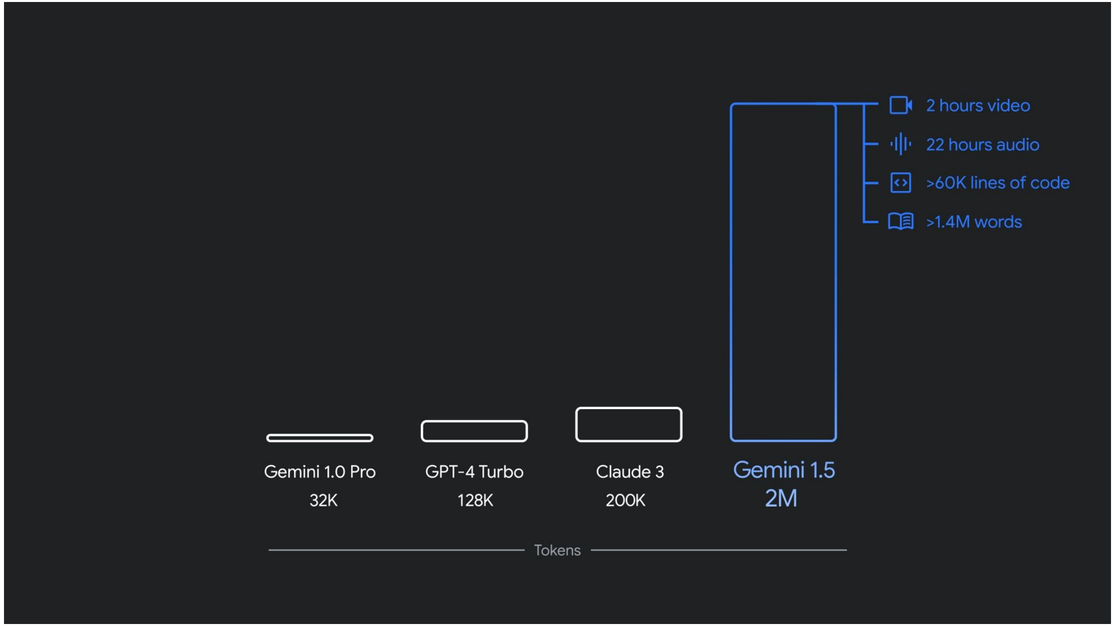
# ImageBind: One embedding to bind them all

- first AI model capable of binding data from six modalities at once
- without the need for explicit supervision.
- images and video, audio, text, depth, thermal and inertial measurement units (IMUs)



# Gemini 1.5 Pro

- can reason across **vision, video, and text**
- **Longer context (2M)**



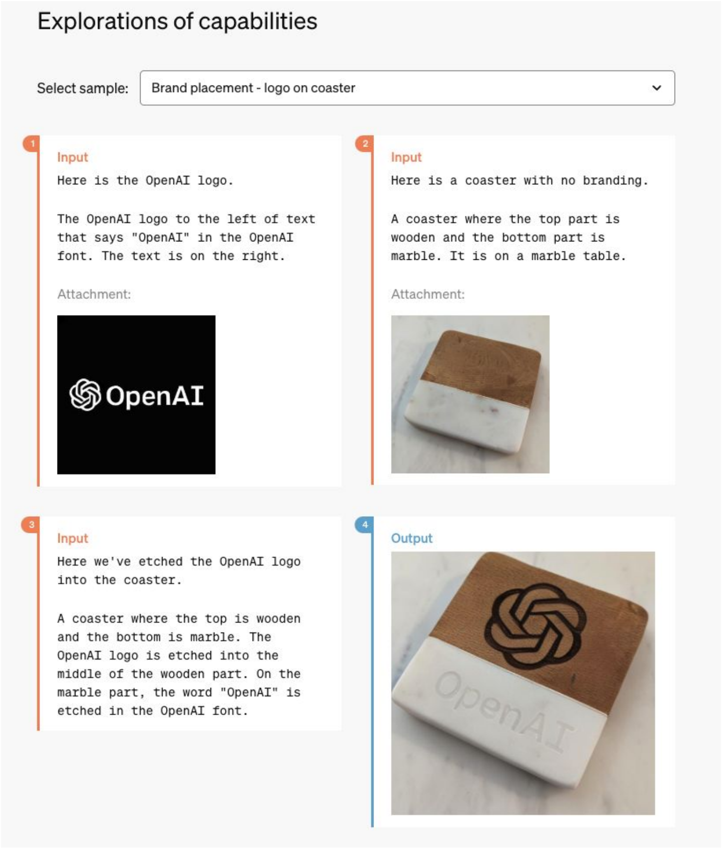

**New Task.**
- Video analysis
- Processing complex documents
- Code understanding

*Level 4: Beyond Images*

# GPT-4o: Omni Model

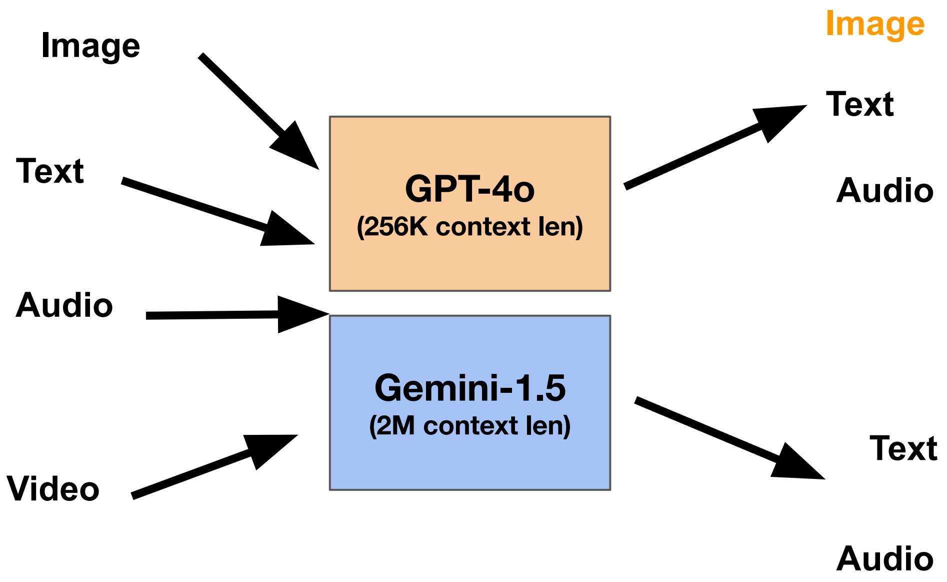
- can reason across **audio, vision, video, and text in real time**

**New Task.**
- Visual Narratives
- Poster Creation
- Character Design
- Poetic Typography
- Poetic Typography with Iterative Editing 2
- Commemorative Coin Design for GPT-40
- Photo to Caricature
- Text to Font
- 3D Object Synthesis
- Brand Placement: Logo on Coaster
- Poetic Typography
- Multiline Rendering: Robot Texting
- Meeting Notes with Multiple Speakers
- Lecture Summarization
- Variable Binding: Cube Stacking
- Concrete Poetry



Level 4: Beyond Images

# Compare GPT-4o and Gemini 1.5

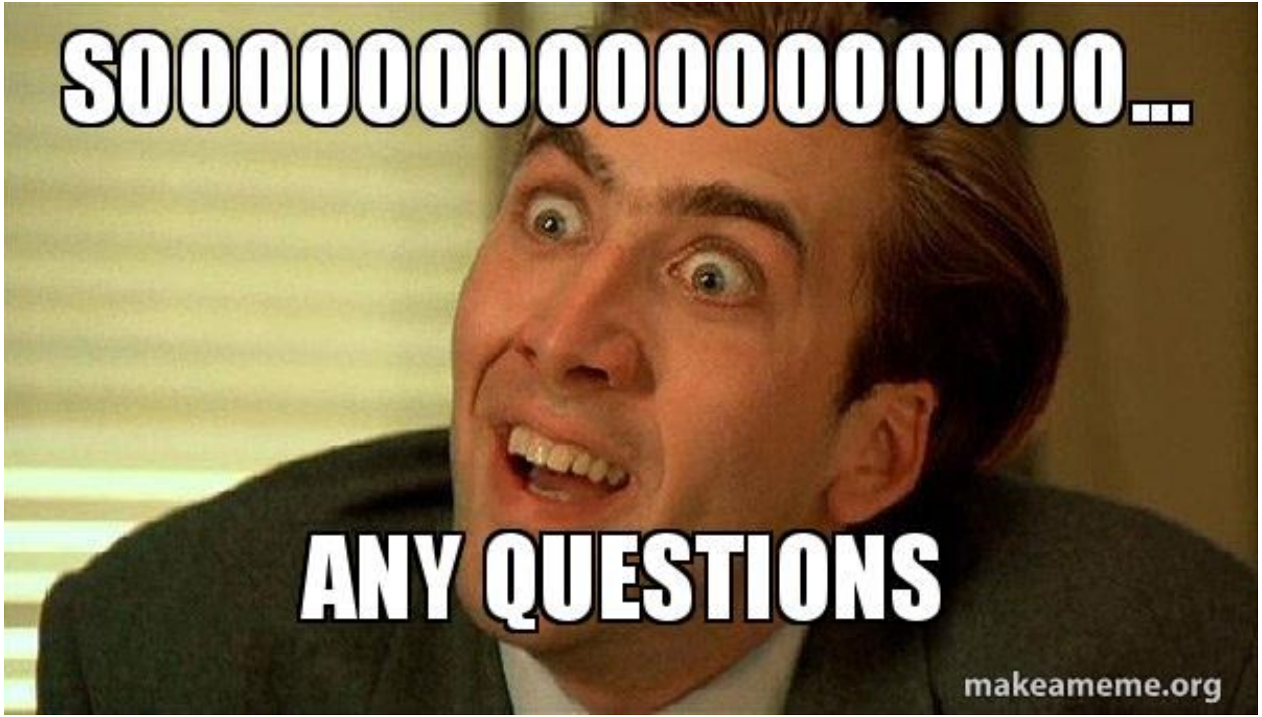



**Very Fast Inference**
avg 320 milliseconds = human response time

**Longer Context x8**
1 hour of video
11 hours of audio

# Thank YOU

# Q & A



# โลกสองใบ:
## Binding Language-Image

# Feedback is a GIFT 🎁

SCAN ME

https://forms.gle/nSZit44NAtEKqXZC9