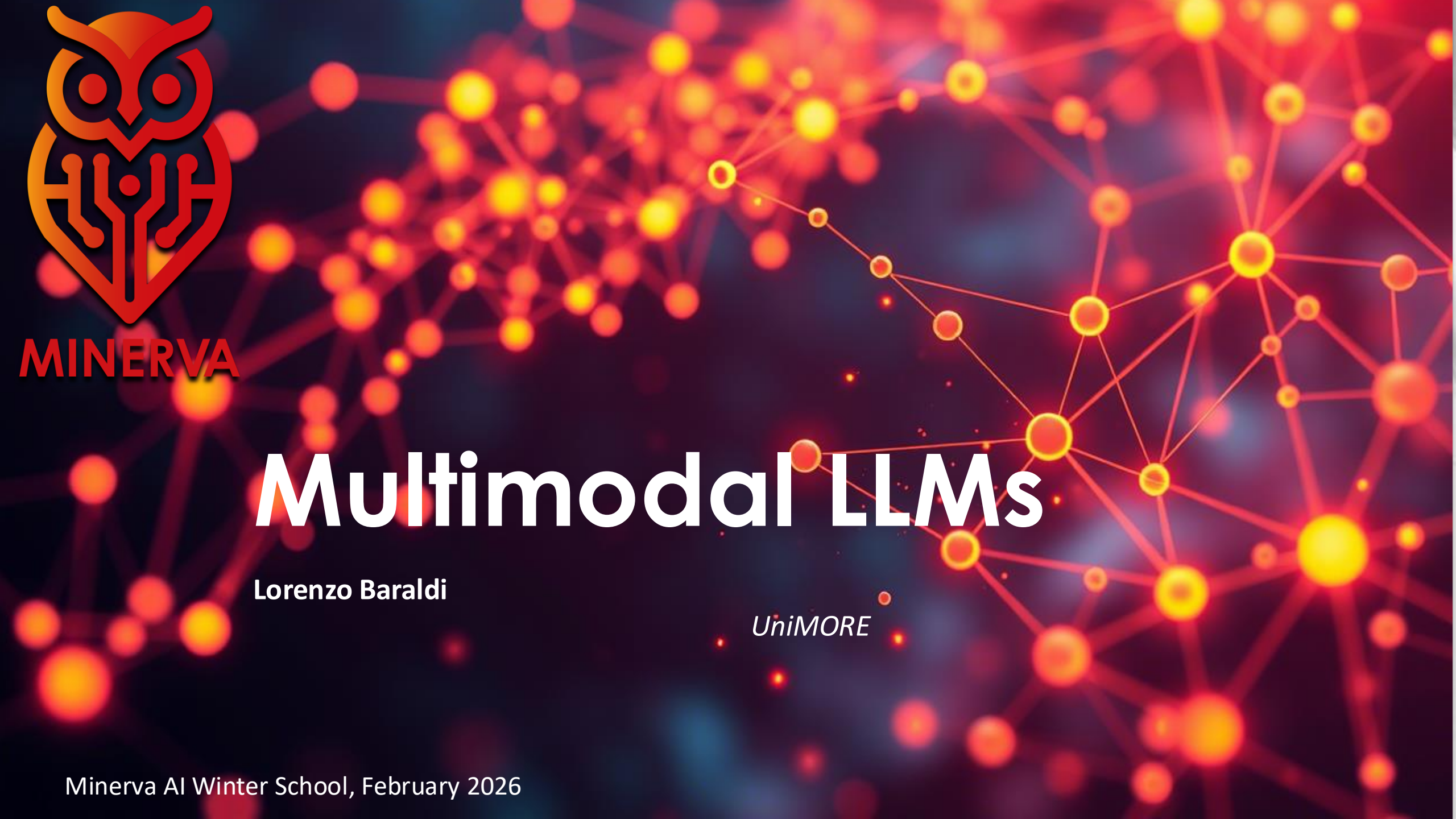


Multimodal LLMs

Lorenzo Baraldi

UniMORE

Minerva AI Winter School, February 2026

What is multimodality?

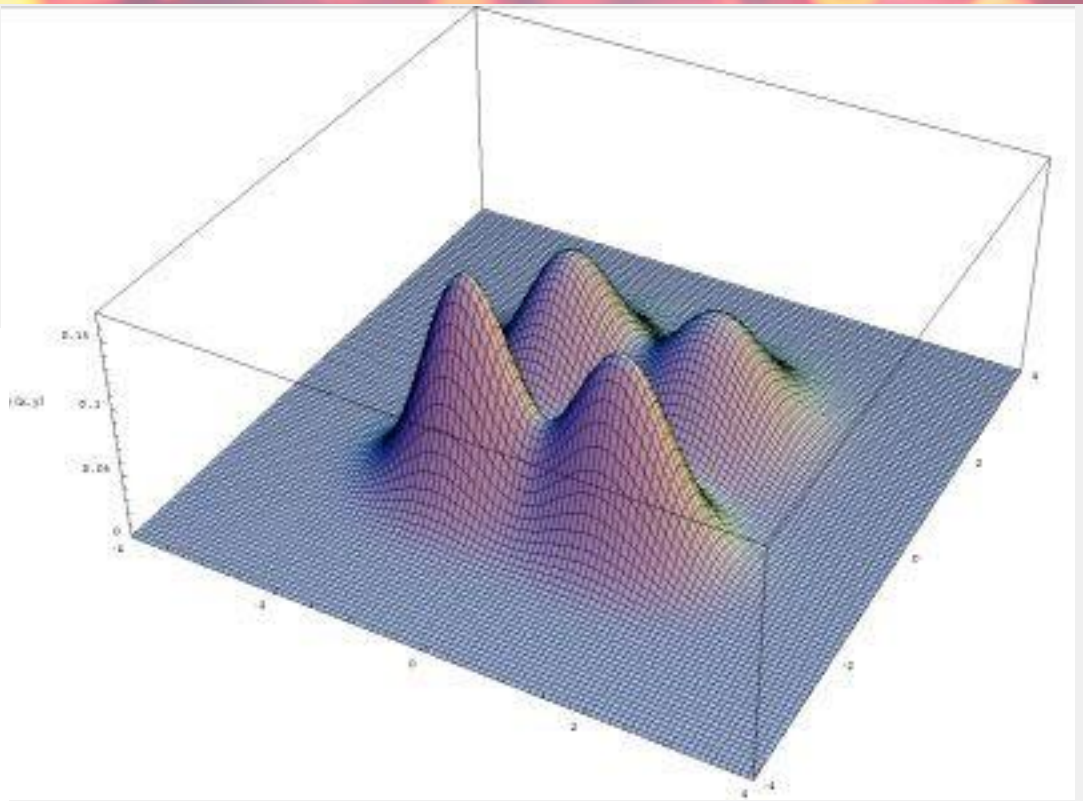
multimodal adjective

mul·ti·mod·al ˌməl-tē-ˈmō-dᵊl -ˌtī-

: having or involving several modes, modalities, or maxima

multimodal distributions

multimodal therapy



In our case, focusing on NLP: text + one or more other *modality* (images, speech, audio, olfaction, others). We'll mostly focus on images as the other modality.

Multi-modal data

- Multimodal data:
  - Input and output from different modalities (e.g. text-to-image, image-to-text)
  - Inputs are multimodal (e.g. a system that can process both text and images)
  - Outputs are multimodal (e.g. a system that can generate both text and images)

Multimodal applications

Let's say we're dealing with two modalities – text, and images:

- Retrieval (image <> text)
- Captioning (image -> text)
- Generation (text -> image)
- Visual question answering (image+text -> text)
- Multimodal classification (image+text -> label)
- Better understanding/generation (image+text -> label/text)

Multimodal is hot right now

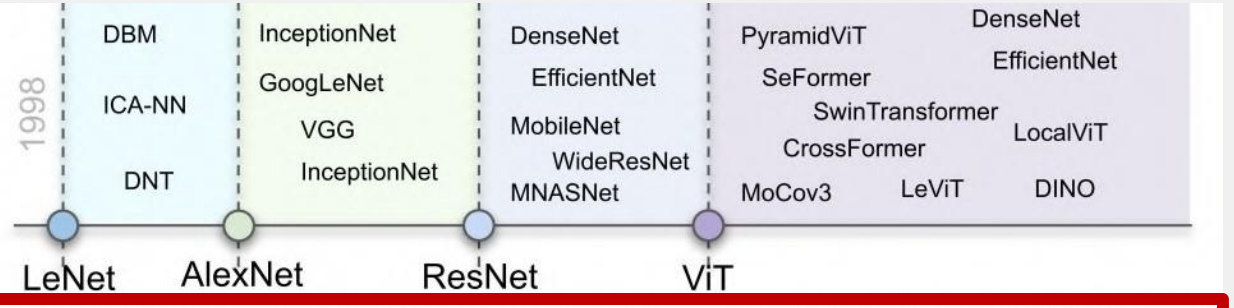
.. and/but has been “the next big thing” for almost a decade!

Language Is Not All You Need: Aligning Perception with Language Models

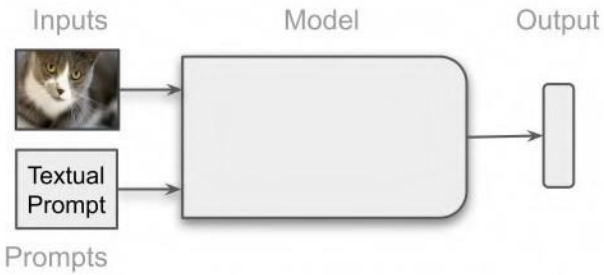
Shaohan Huang,* Li Dong,* Wenhui Wang,* Yaru Hao,* Saksham Singhal,* Shuming Ma*
Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi
Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, Furu Wei†
Microsoft

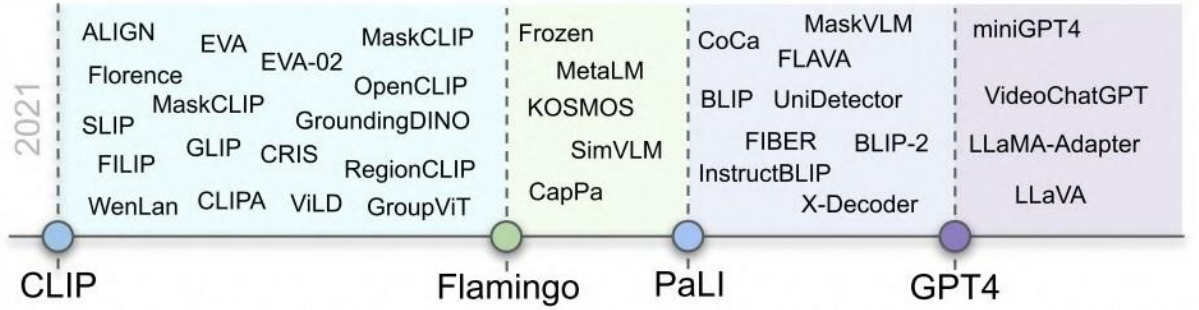
Multimodal models

Traditional Models

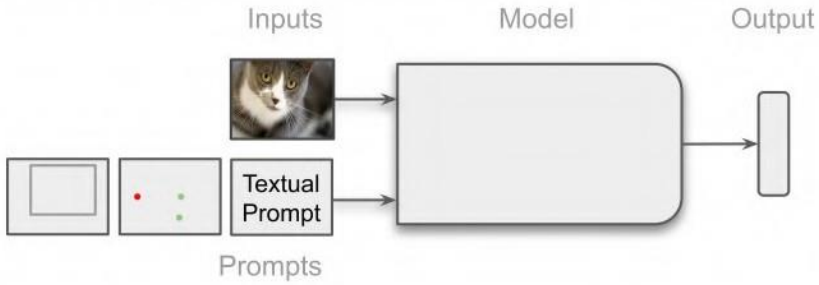


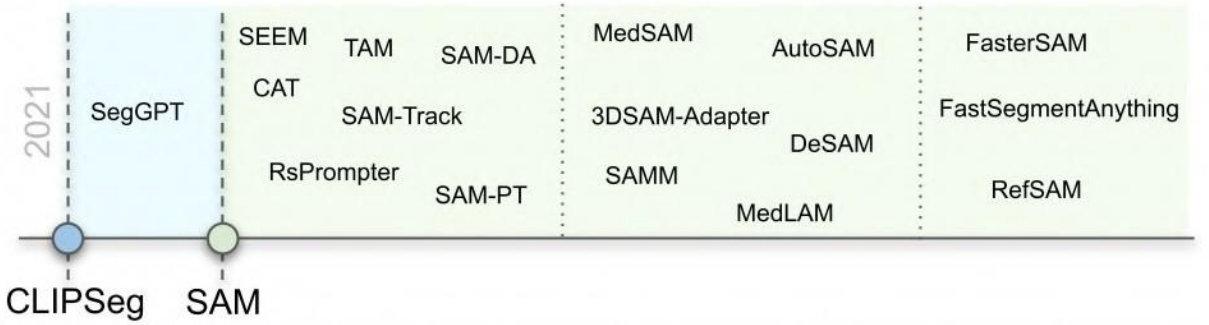
Textually Prompted Models



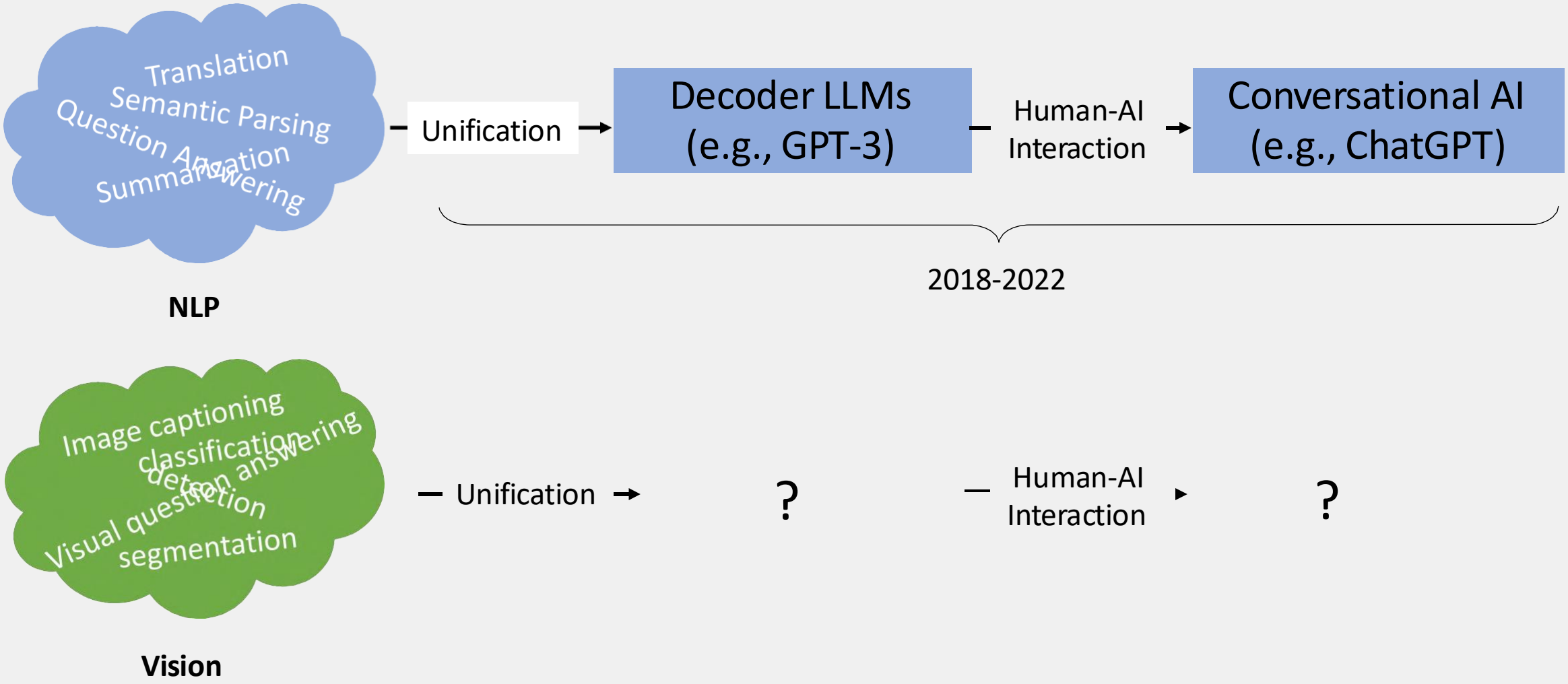


Visually Prompted Models

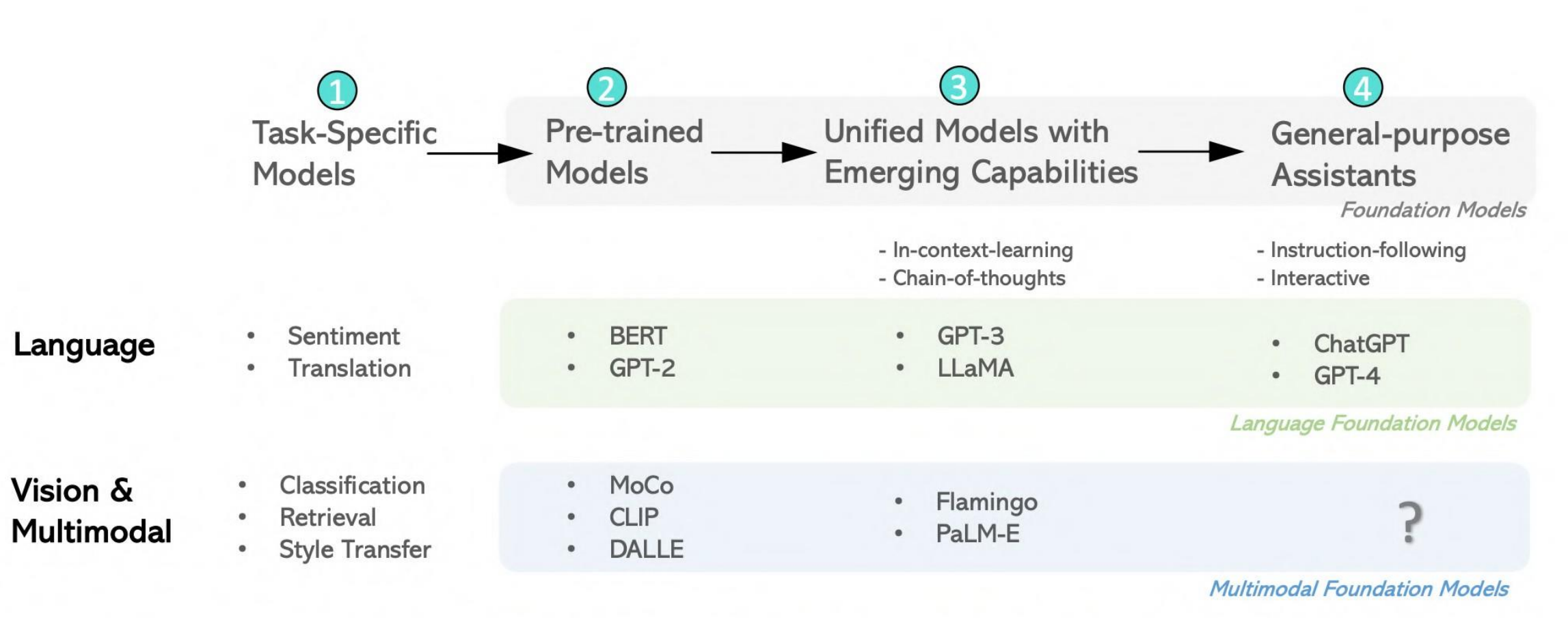




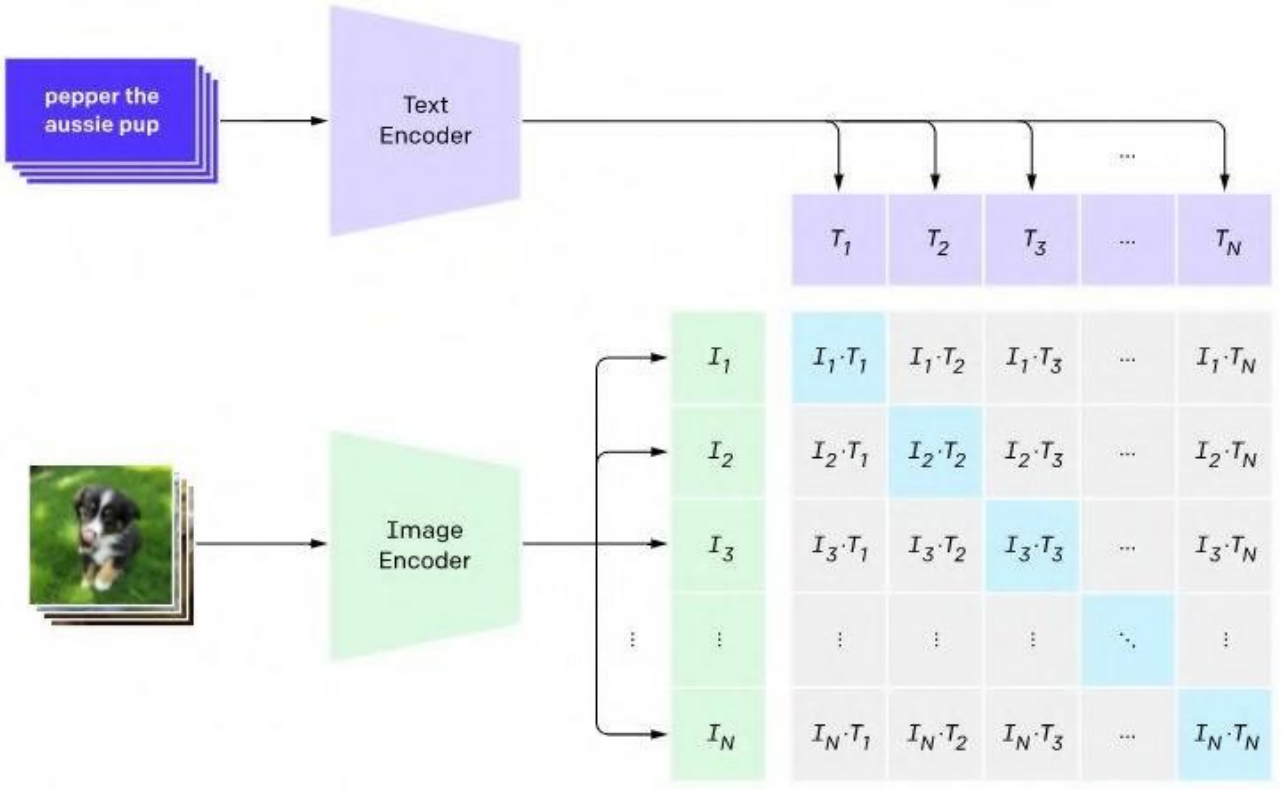
A Lesson from LLMs



A Lesson from LLMs



CLIP: Models and Training Complexity

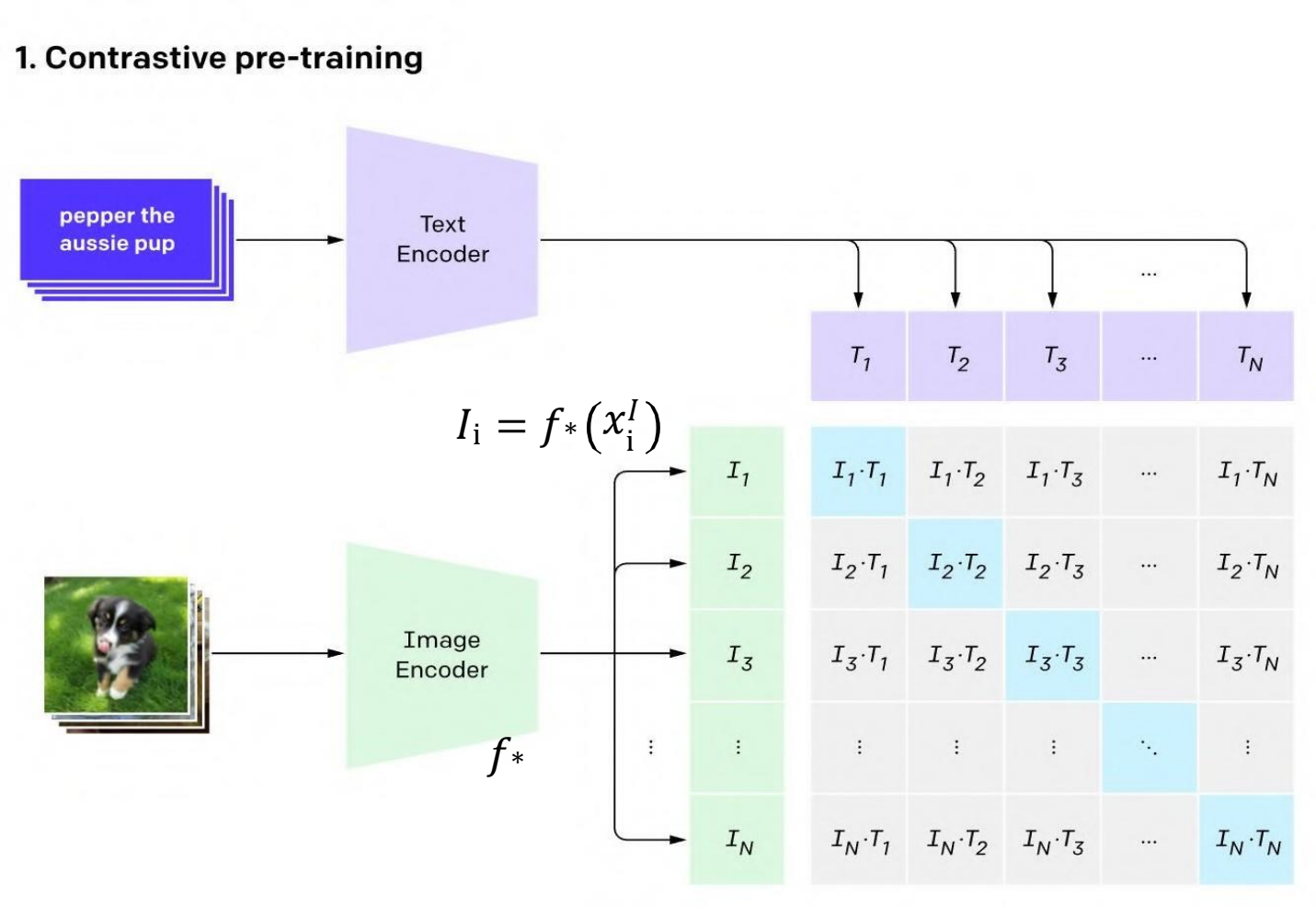


- Text encoder:
  - 12-layer Transformer with causal mask
- Image encoder:
  - ResNet families: RN50, RN101, RN50x4, RN50x16, RN50x64
  - ViT families: ViT-B/32, ViT-B/16, ViT-L/14

Contrastive learning

- Contrastive training to bridge the image and text embedding spaces
- Making embedding of (image, text) pairs similar and that of non-pairs dissimilar
- This embedding space is super helpful for performing searches across modalities
  - Can return the best caption given an image
  - Has impressive capabilities for zero-shot adaptation to unseen tasks, without the need for fine-tuning

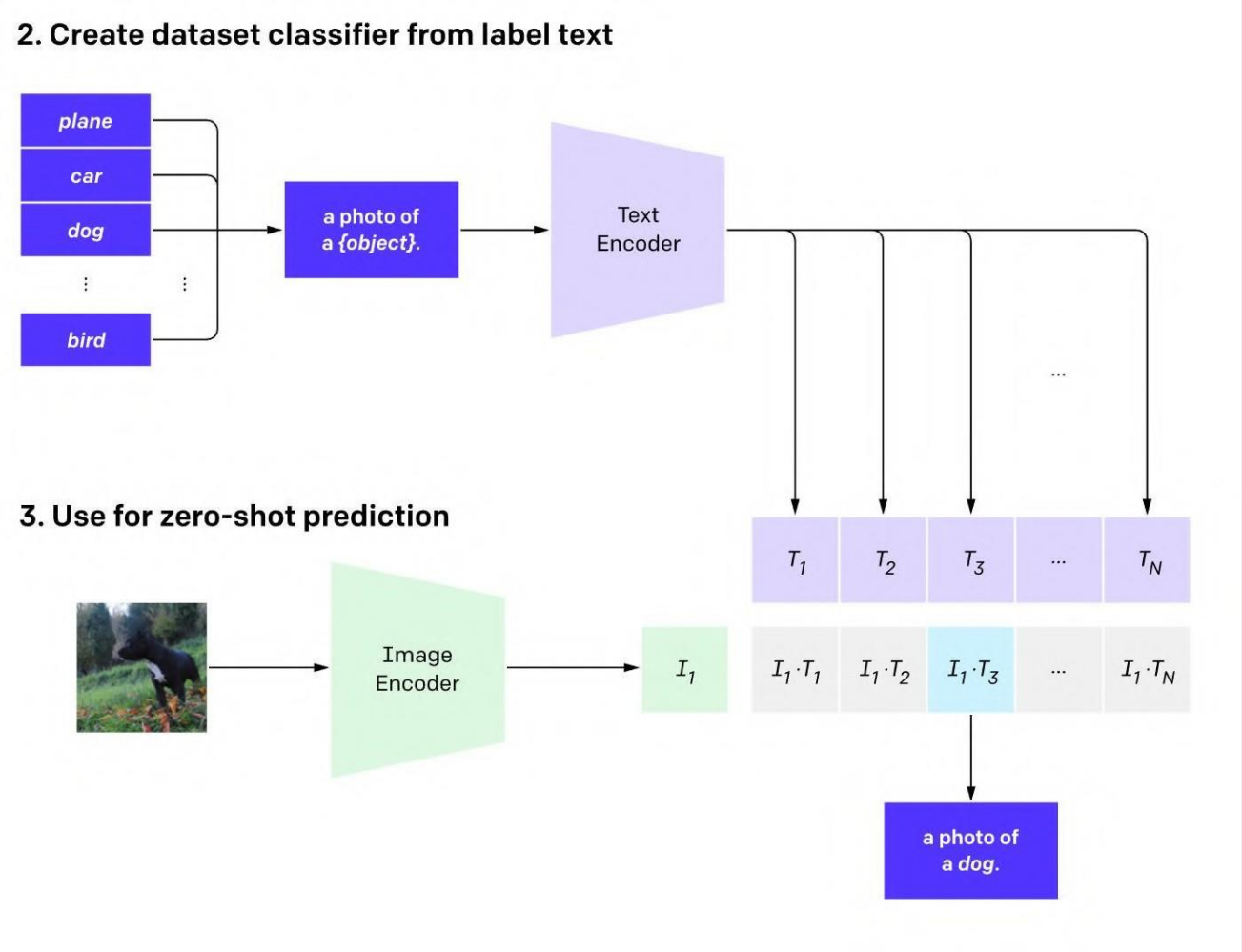
Contrastive learning



- Training batchsize: 32,768
- Training time:
  - RN50x64: 18 days on 592 V100 GPUs
  - ViT-L/14: 12 days on 256 V100 GPUs

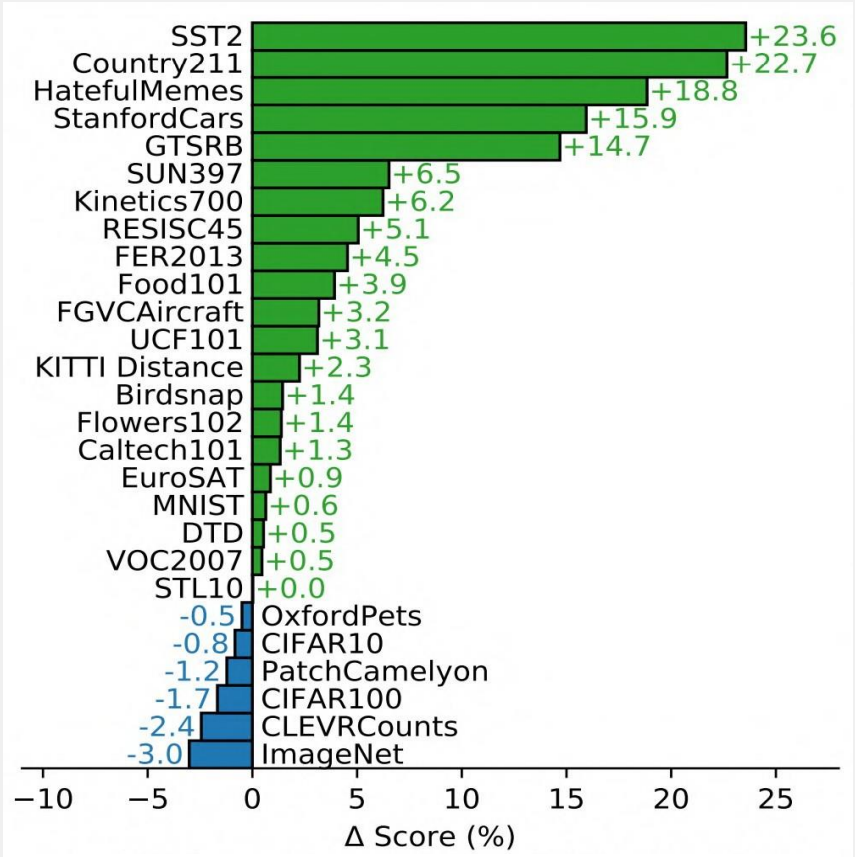
https://openai.com/blog/clip/

CLIP for zero-shot learning

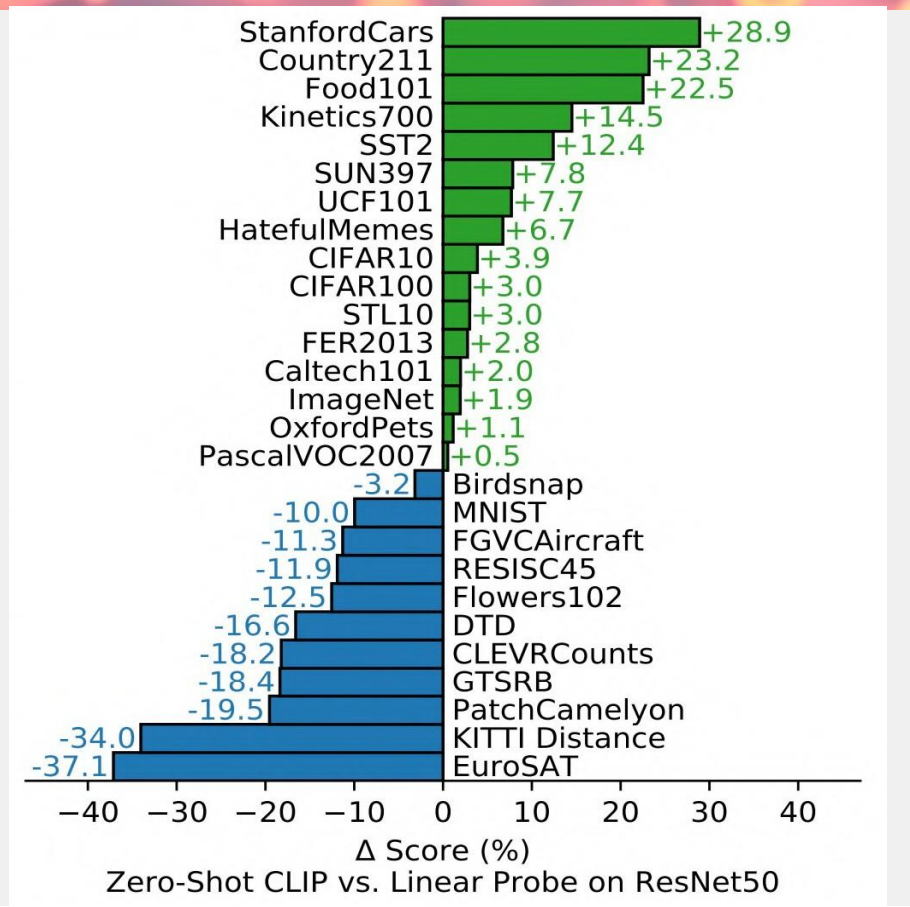


encodes all the text labels and compares them to the encoded image

Zero-shot CLIP outperforms a supervised linear classifier fitted on 16 out of 27 datasets including ImageNet.



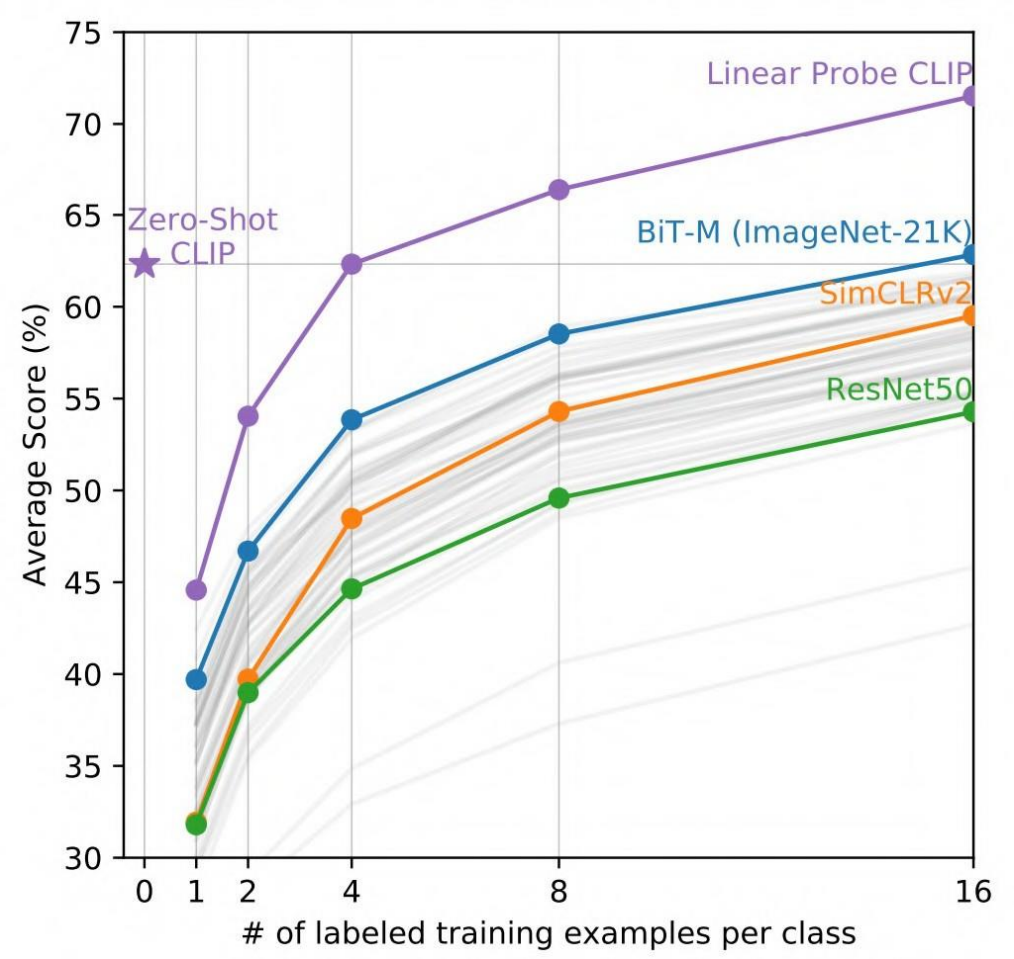
Linear-probing CLIP outperforms the linear probing Noisy Student EfficientNet-L2



Zero-shot CLIP is competitive with a fully supervised linear-probing ResNet50

CLIP's features outperform the features of the best ImageNet model on a wide variety of datasets.

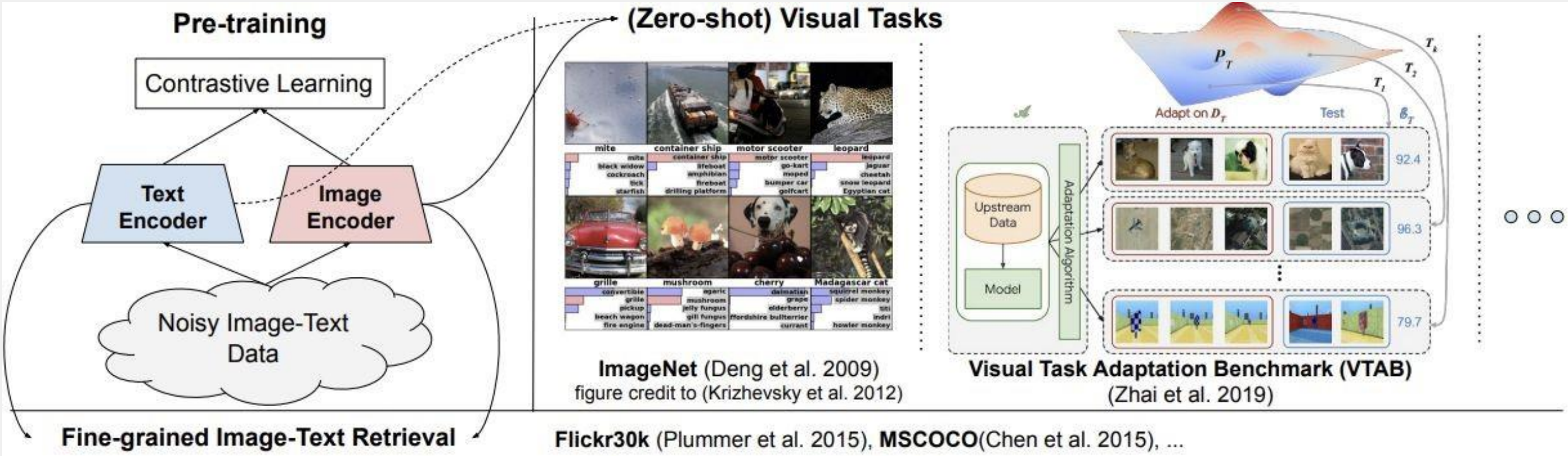
Zero-shot CLIP outperforms few-shot linear probes



Radford et al. "Learning Transferable Visual Models From Natural Language Supervision", ICML 2021

ALIGN (Jia et al., 2021)

Same idea, but EVEN MORE data (JFT at 1.8B image-text pairs vs CLIP's 300m).



Aligned datasets

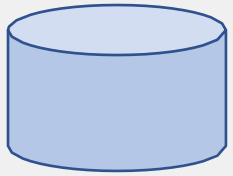
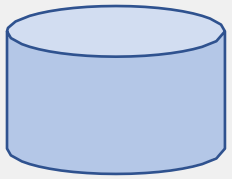
HUGE open source datasets of image-text pairs now exist.

Used to train eg StableDiffusion (Rombach et al., 2022).

https://laion.ai/blog/laion-5b/

Vision and Language Tasks

- Large Multi-modal Models (LMMs) in their current form primarily generates a text sequence.

	Image Captioning	Text-to Image Retrieval	Image-to-Text Retrieval	VQA	Text-to-Image Generation
Input	Image: 	Query: A couple of zebra walking across a dirt road. 	Query:  	Image:  Q: why did the zebra cross the road?	Text: A couple of zebra walking across a dirt road.
Output	A couple of zebra walking across a dirt road.		A couple of zebra walking across a dirt road.	A: to get to the other side (Selected from a pool of 3,129 answers in VQAv2 or generate answer)	
	Generation	**Understanding**	**Understanding**	**Understanding/Generation**	**Generation**

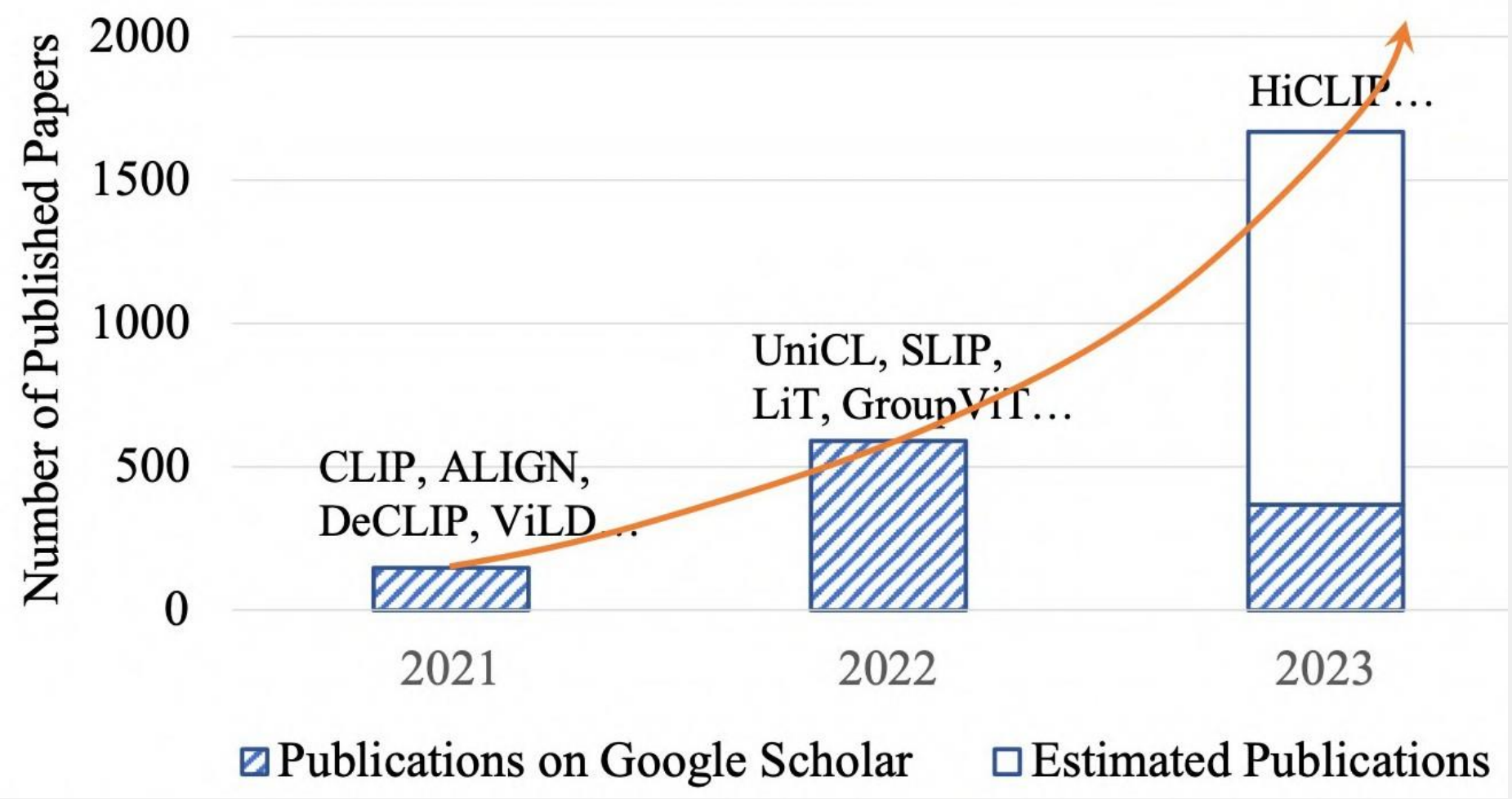
CLIP: Summary

✓CLIP improved open-vocabulary visual recognition capabilities through learning from Internet-scale image-text pairs.

✗CLIP doesn’t go directly from image to text or vice versa. It just connecting the image and text embedding spaces

- CLIP can only address limited use cases such as classification
- It crucially lack the ability to generate language which makes them less suitable to more open-ended tasks such as captioning or visual question answering

Publication on VLMs



CLIP Variants

- Objective function or pretraining
  - Combining CLIP with label supervision (BASIC, UniCL, LiT, MOFI)
  - Contrastive + self-supervised image representation learning
    - Contrastive + Self-supervlised methods like SimCLR (SLIP, DeCLIP, nCLIP)
    - Contrastve + Masked Image Modeling (EVA, EVA-02, MVP)
  - Fine-grained matching loss (FILIP)
  - Region-level pretraining (RegionCLIP, GLIP)
  - Sigmoid loss for language-image pre-training (SigCLIP)

Instead of the softmax-based contrastive loss, we propose a simpler alternative that does not require computing global normalization factors. The sigmoid-based loss processes every image-text pair independently, effectively turning the learning problem into the standard binary classification on the dataset of all pair combinations, with a positive labels for the matching pairs (I_i, T_i) and negative labels for all other pairs $(I_i, T_{j\neq i})$. It is defined as follows:

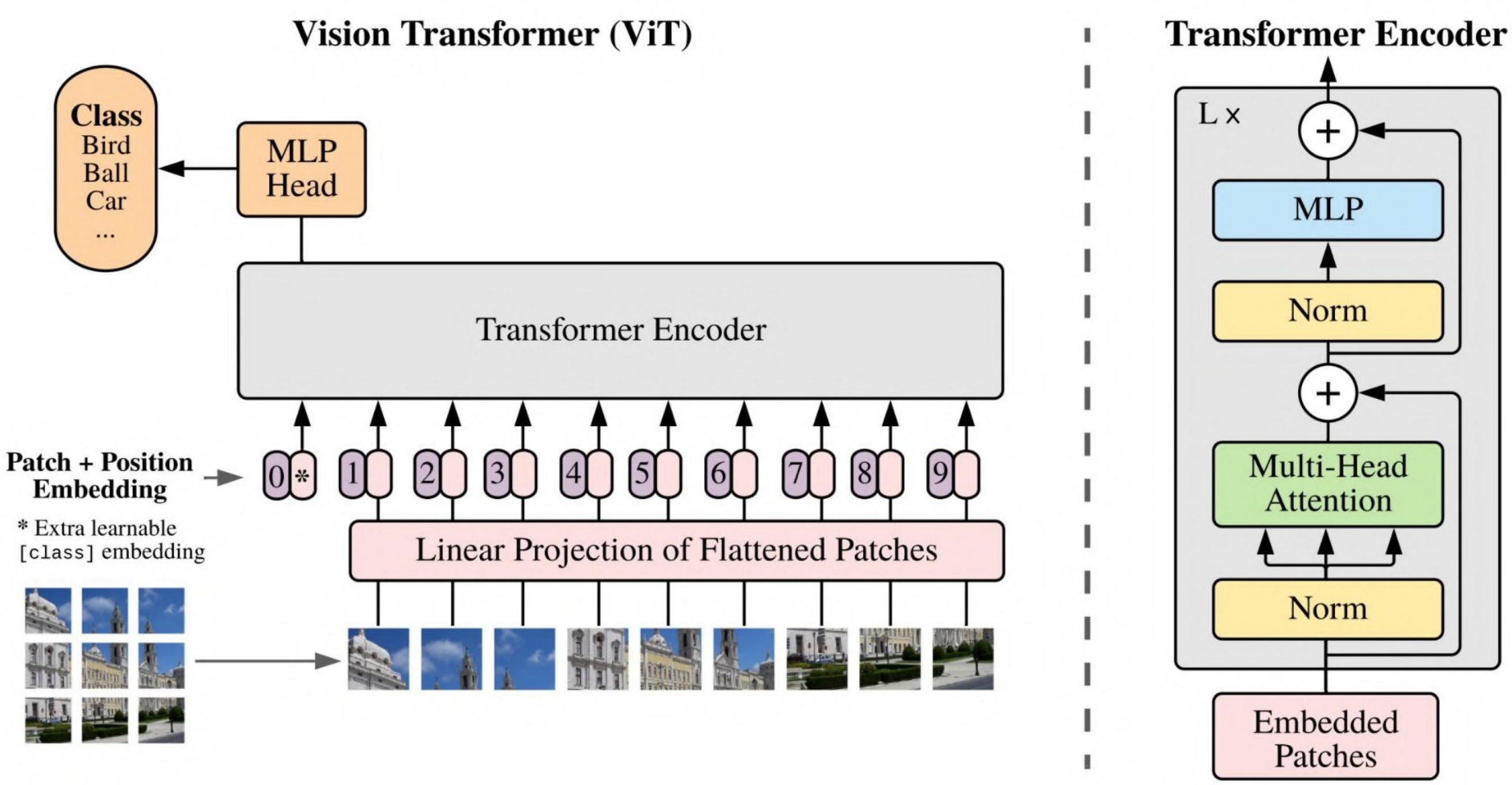
$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

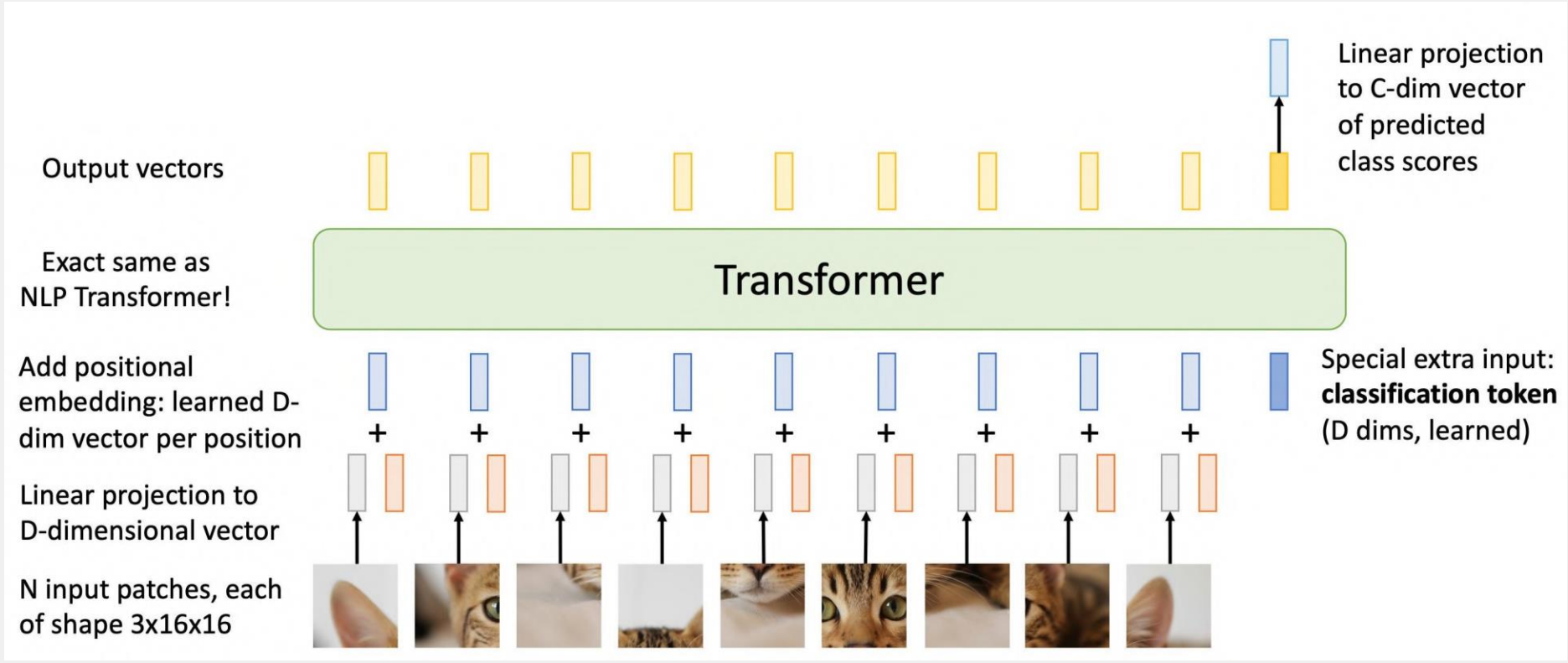
where z_{ij} is the label for a given image and text input, which equals 1 if they are paired and -1 otherwise. At initialization, the heavy imbalance coming from the many negatives dominates the loss, leading to large initial optimization steps attempting to correct this bias. To alleviate this, we introduce an additional learnable bias term b similar to the temperature t. We initialize t' and b to $\log 10$ and -10 respectively. This makes sure the training starts roughly close to the prior and does not require massive over-correction. Algorithm 1 presents a pseudocode implementation of the proposed sigmoid loss for language image pre-training.

Algorithm 1 Sigmoid loss pseudo-implementation.

```
# img_emb      : image model embedding [n, dim]
# txt_emb      : text model embedding [n, dim]
# t_prime, b   : learnable temperature and bias
# n            : mini-batch size

t = exp(t_prime)
zimg = l2_normalize(img_emb)
ztxt = l2_normalize(txt_emb)
logits = dot(zimg, ztxt.T) * t + b
labels = 2 * eye(n) - ones(n) # -1 with diagonal 1
l = -sum(log_sigmoid(labels * logits)) / n
```

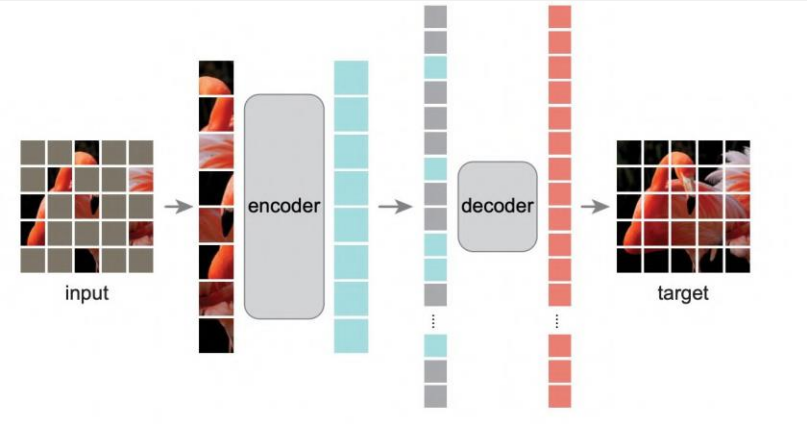
Vision Transfomer as Image Encoder Architecture



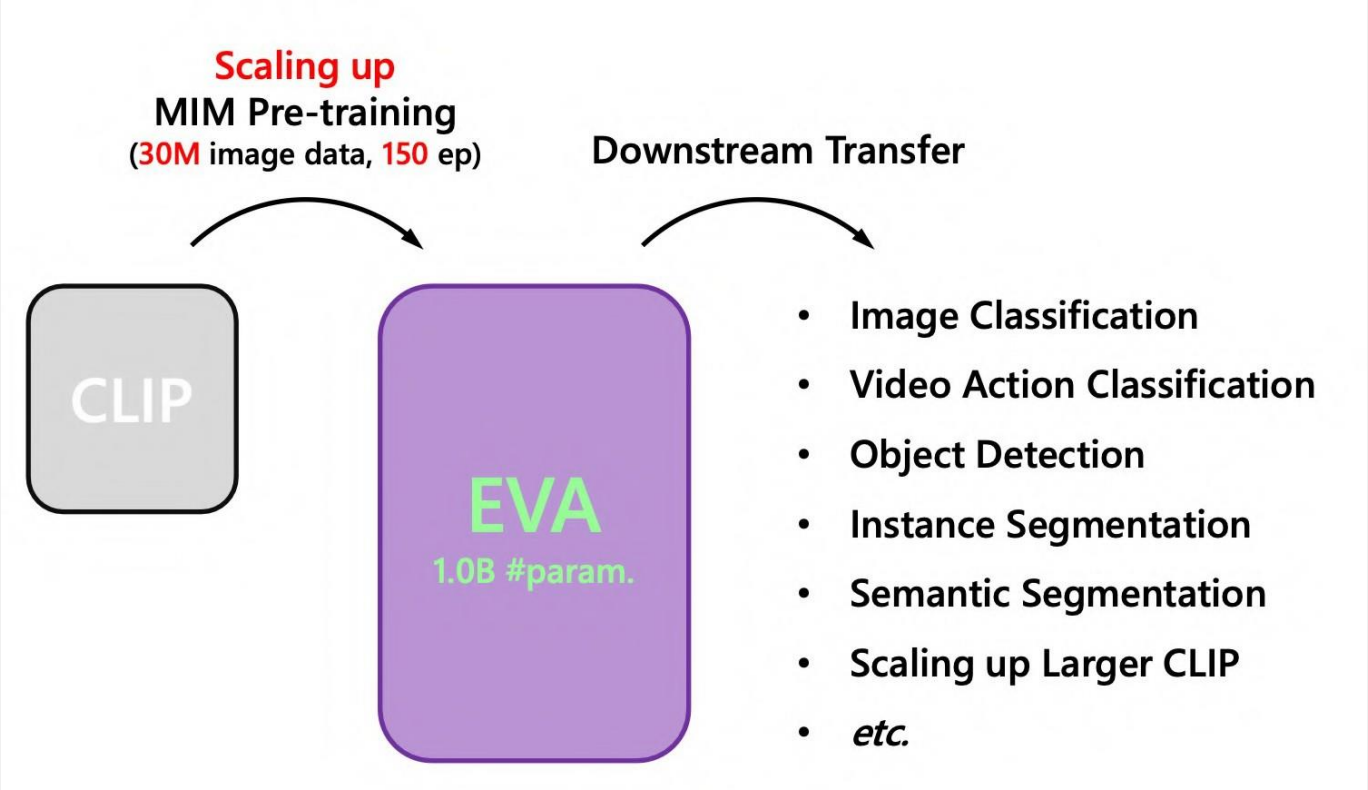


EVA

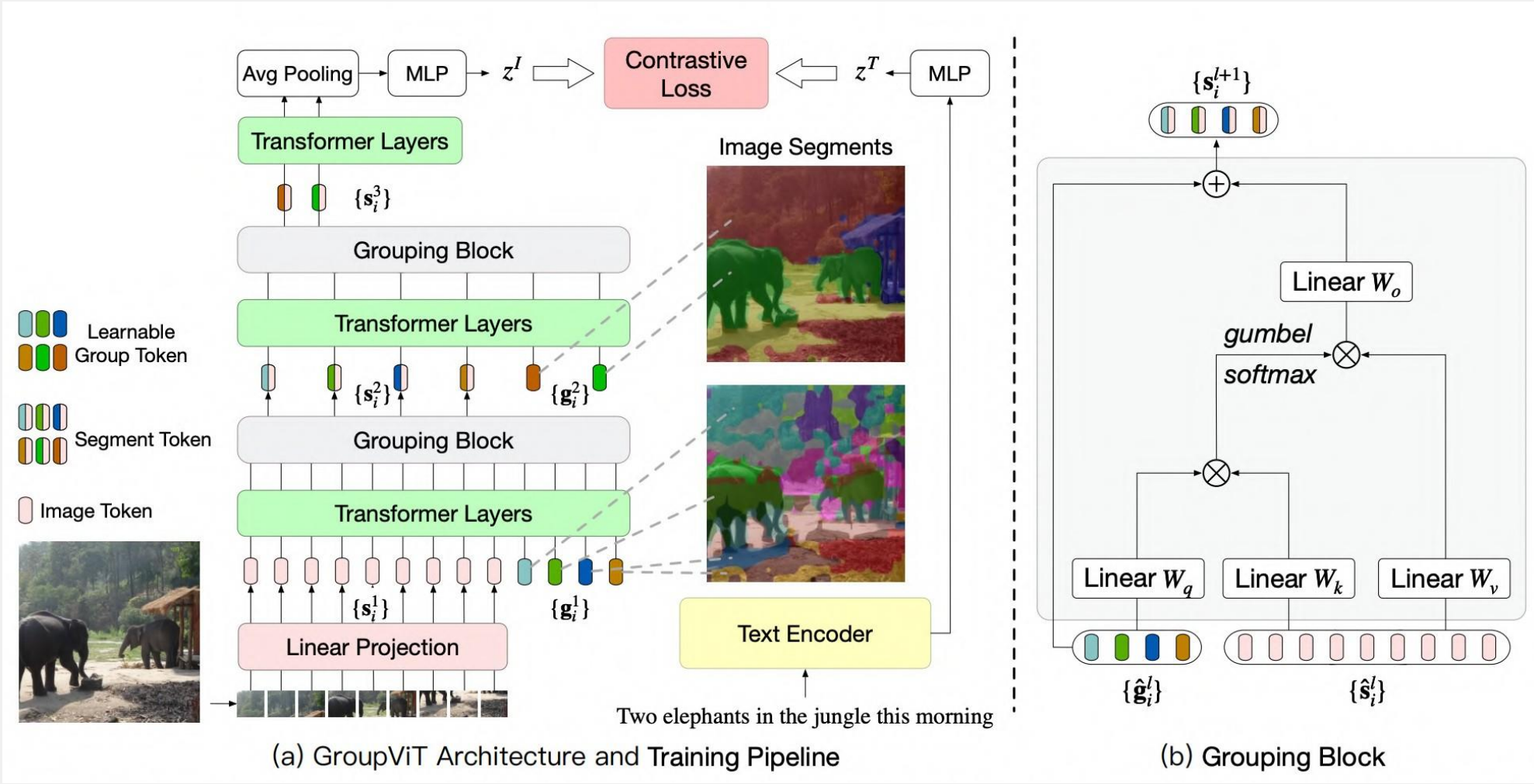
- Simply regressing the masked out image-text aligned vision features (*i.e.*, CLIP features) scales up well (to 1.0B parameters) and transfers well to various downstream tasks.



He et al., “Masked Autoencoders Are Scalable Vision Learners”, 2021



GroupVIT



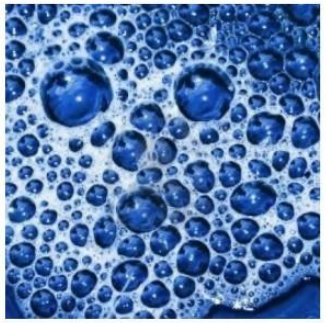
Learning to Prompt for VLMs

Caltech101	Prompt	Accuracy
	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

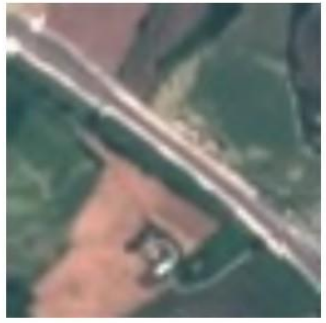
(a)

Flowers102	Prompt	Accuracy
	a photo of a [CLASS].	60.86
	a flower photo of a [CLASS].	65.81
	a photo of a [CLASS], a type of flower.	66.14
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**94.51**

(b)

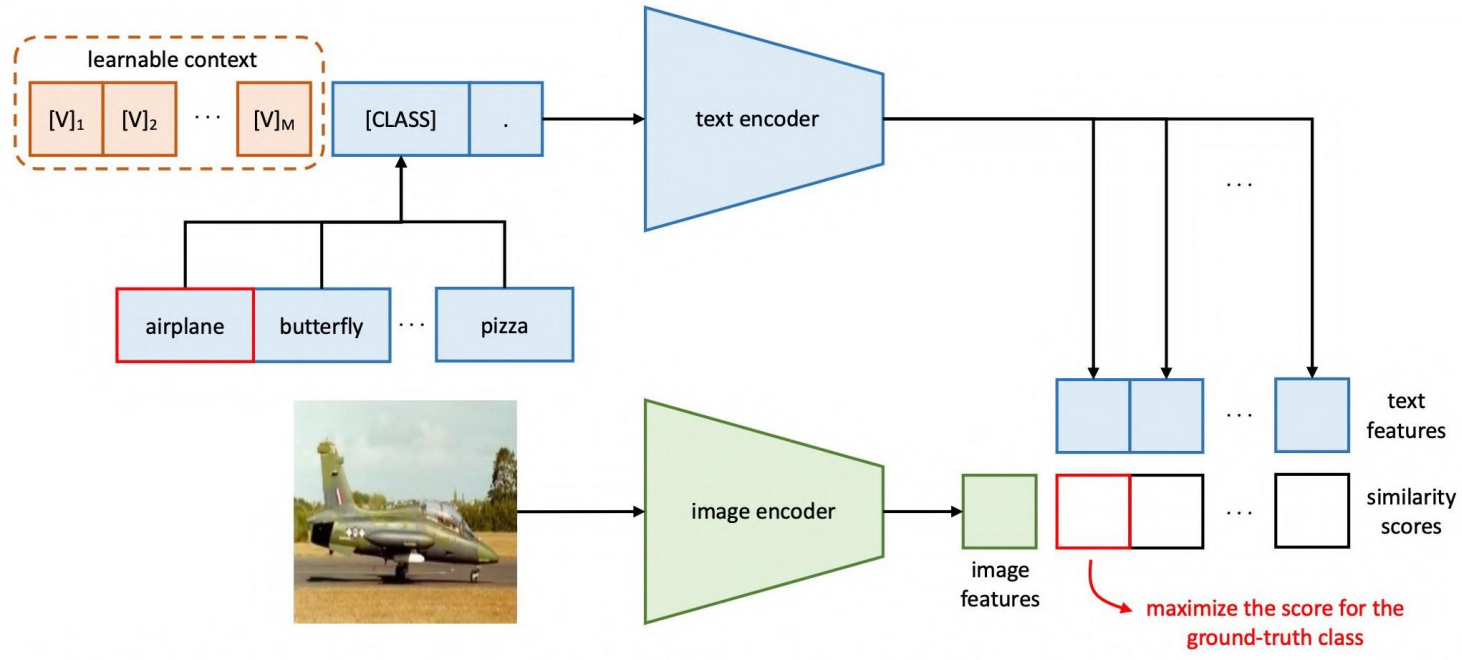
Describable Textures (DTD)	Prompt	Accuracy
	a photo of a [CLASS].	39.83
	a photo of a [CLASS] texture.	40.25
	[CLASS] texture.	42.32
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**63.58**

(c)

EuroSAT	Prompt	Accuracy
	a photo of a [CLASS].	24.17
	a satellite photo of [CLASS].	37.46
	a centered satellite photo of [CLASS].	37.56
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**83.53**

(d)

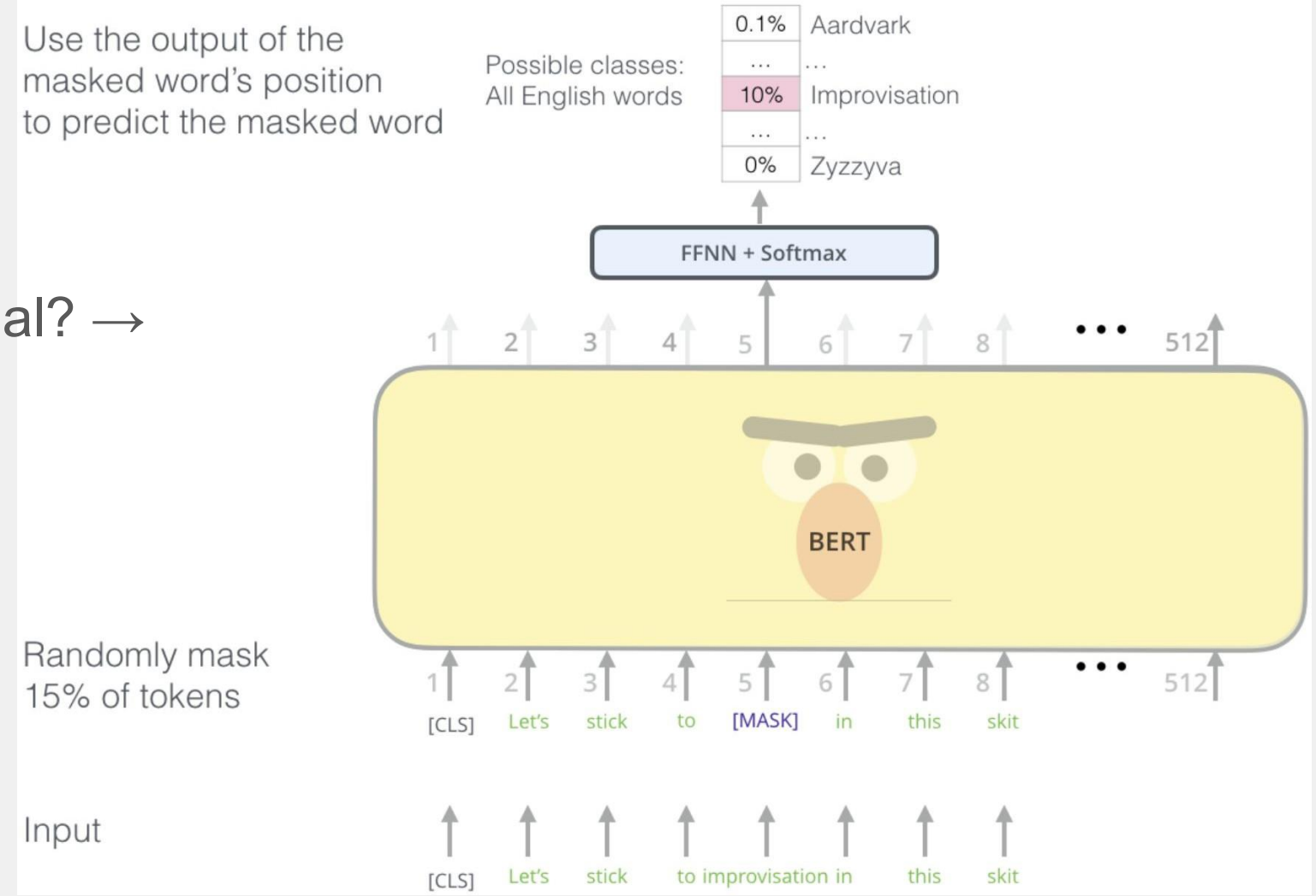
Learning to Prompt for VLMs



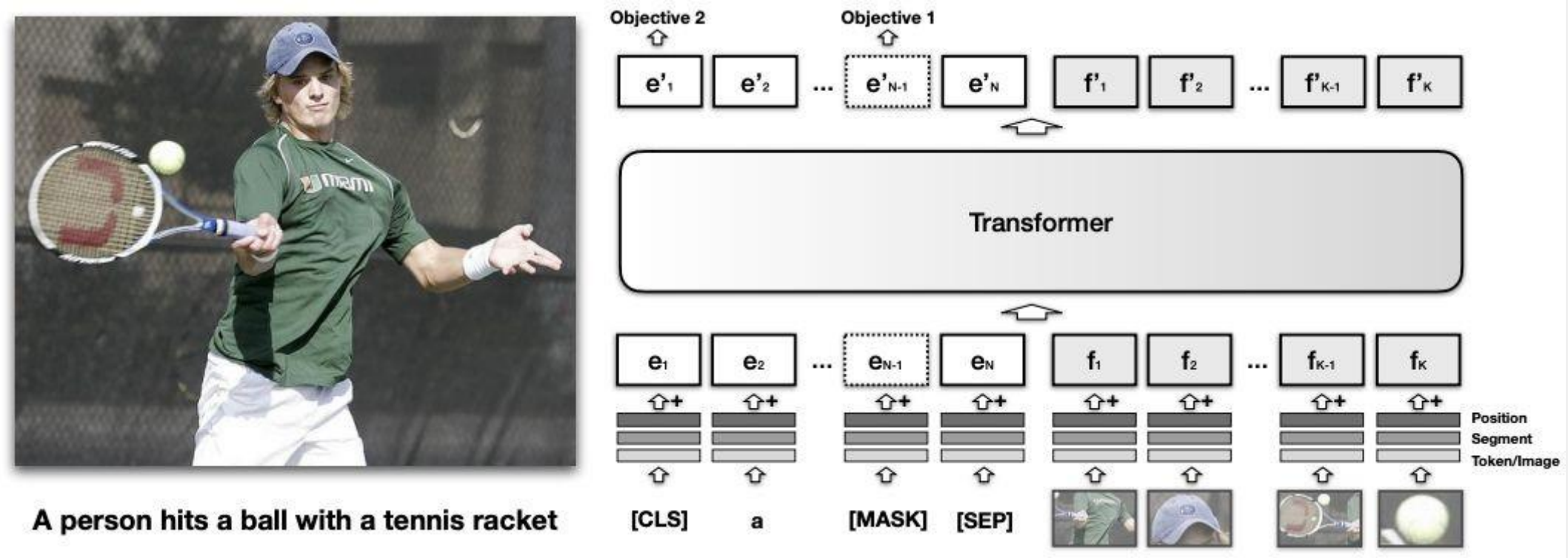
Method	Source	Target			
	ImageNet	-V2	-Sketch	-A	-R
ResNet-50					
Zero-Shot CLIP	58.18	51.34	33.32	21.65	56.00
Linear Probe CLIP	55.87	45.97	19.07	12.74	34.86
CLIP + CoOp ($M=16$)	62.95	55.11	32.74	22.12	54.96
CLIP + CoOp ($M=4$)	**63.33**	**55.40**	**34.67**	**23.06**	**56.60**
ResNet-101					
Zero-Shot CLIP	61.62	54.81	38.71	28.05	64.38
Linear Probe CLIP	59.75	50.05	26.80	19.44	47.19
CLIP + CoOp ($M=16$)	**66.60**	**58.66**	39.08	28.89	63.00
CLIP + CoOp ($M=4$)	65.98	58.60	**40.40**	**29.60**	**64.98**
ViT-B/32					
Zero-Shot CLIP	62.05	54.79	40.82	29.57	**65.99**
Linear Probe CLIP	59.58	49.73	28.06	19.67	47.20
CLIP + CoOp ($M=16$)	**66.85**	58.08	40.44	30.62	64.45
CLIP + CoOp ($M=4$)	66.34	**58.24**	**41.48**	**31.34**	65.78
ViT-B/16					
Zero-Shot CLIP	66.73	60.83	46.15	47.77	73.96
Linear Probe CLIP	65.85	56.26	34.77	35.68	58.43
CLIP + CoOp ($M=16$)	**71.92**	64.18	46.71	48.41	74.32
CLIP + CoOp ($M=4$)	71.73	**64.56**	**47.89**	**49.93**	**75.14**

BERT Refresher

How do we make this multimodal? →

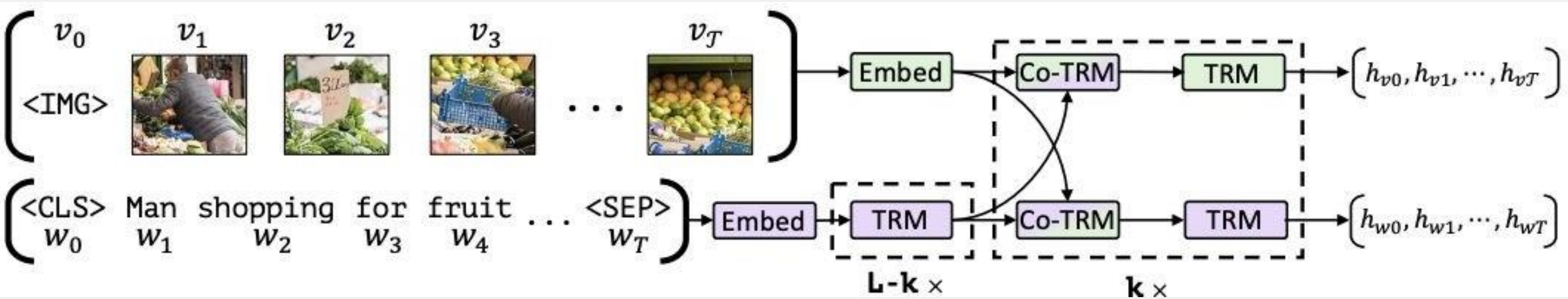


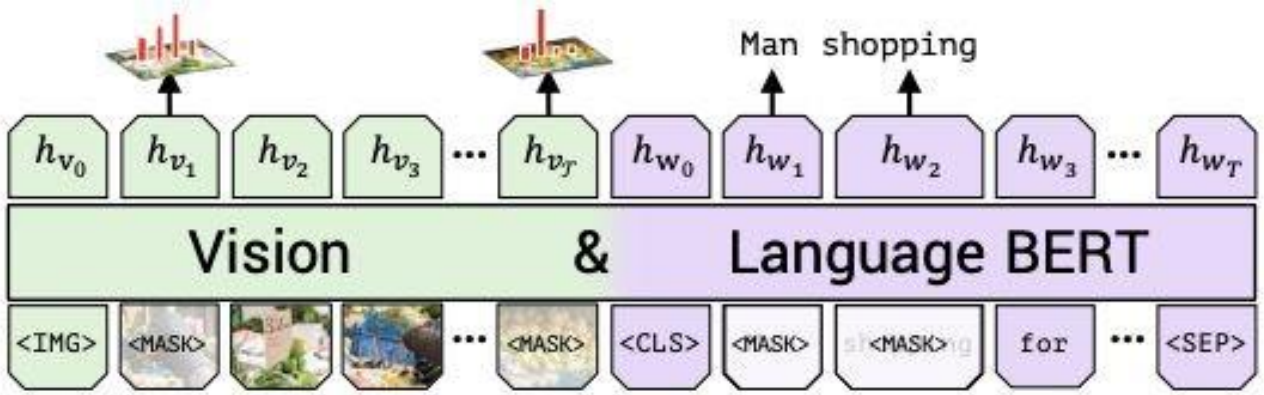
Visual BERTs: VisualBERT



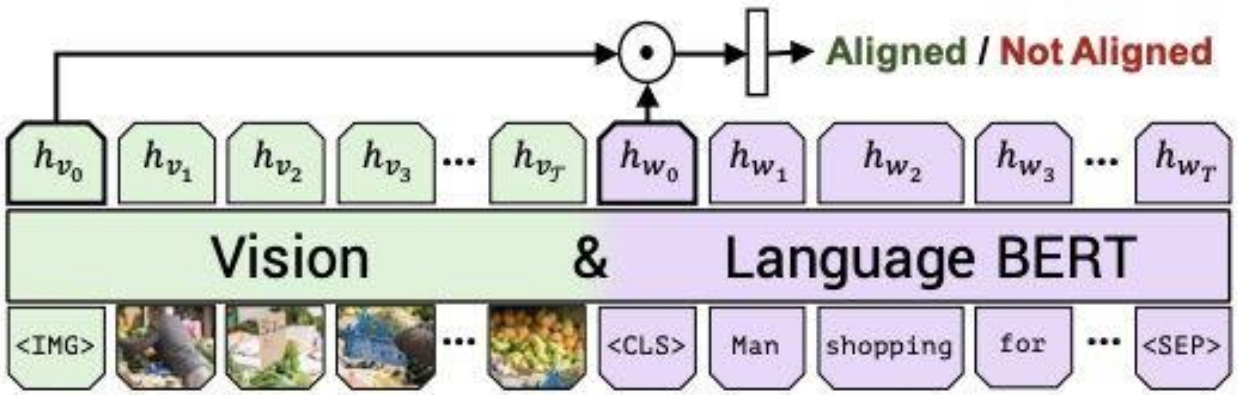
VisualBERT Li et al. 2019

Visual BERTs: ViLBERT





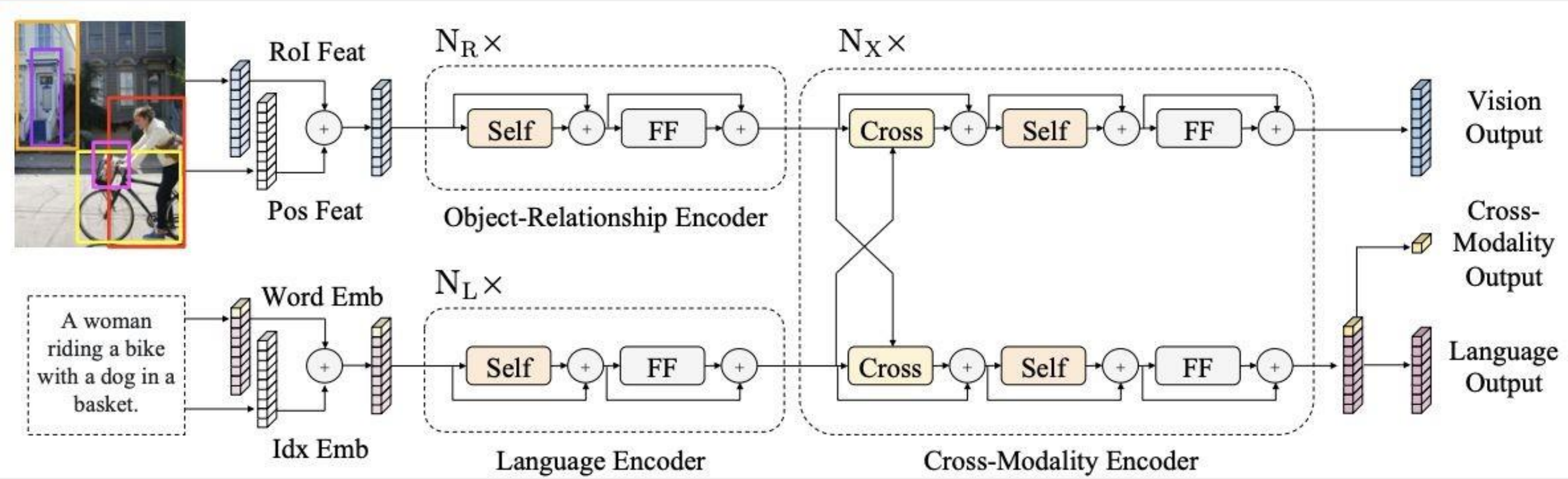
(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

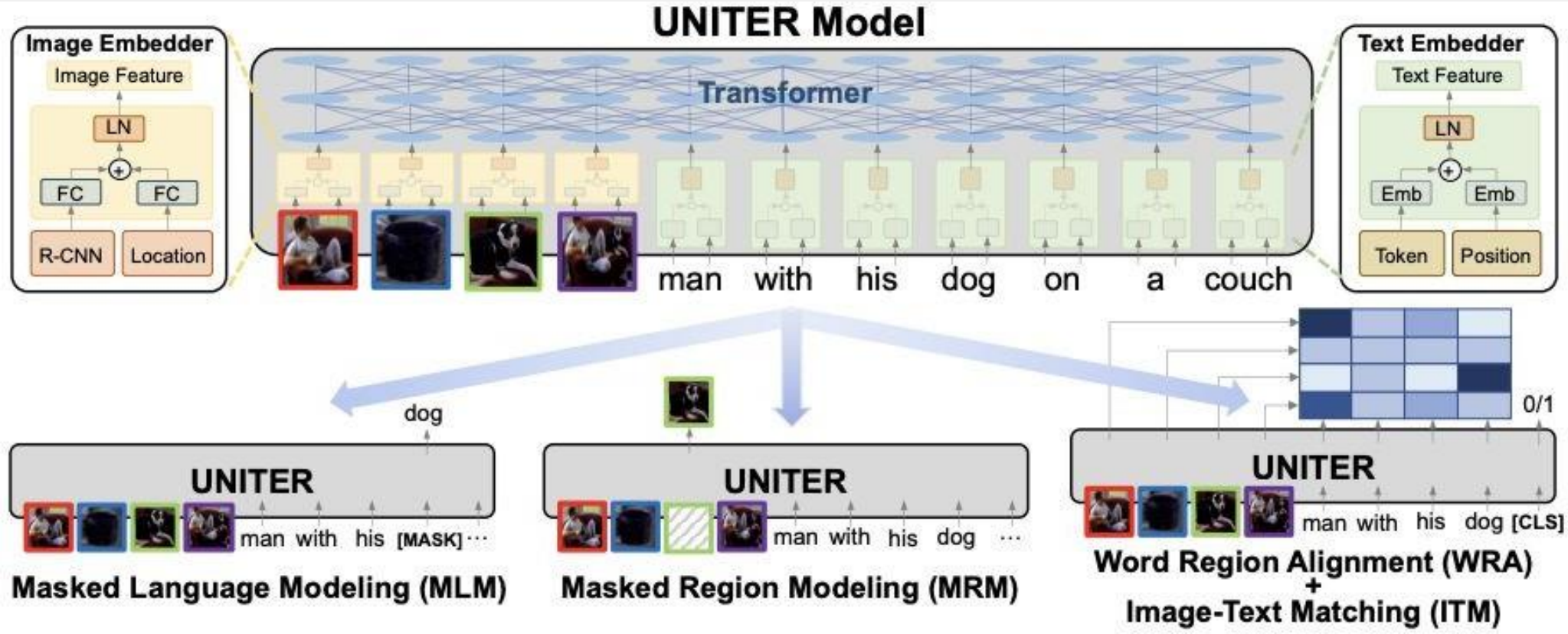
Visual BERTs: LXMERT

Learning Cross-Modality Encoder Representations from Transformers



LXMERT Tan & Bansal 2019

UNITER



Chen, Yi, Lu, et al. 2020

So many models

VL-PTM	Text encoder	Vision encoder	Fusion scheme	Pre-training tasks	Multimodal datasets for pre-training
Fusion Encoder					
VisualBERT [2019]	BERT	Faster R-CNN	Single stream	MLM+ITM	COCO
Uniter [2020]	BERT	Faster R-CNN	Single stream	MLM+ITM+WRA+MRFR+MRC	CC+COCO+VG+SBU
OSCAR [2020c]	BERT	Faster R-CNN	Single stream	MLM+ITM	CC+COCO+SBU+Flickr30k+VQA
InterBert [2020]	BERT	Faster R-CNN	Single stream	MLM+MRC+ITM	CC+COCO+SBU
ViLBERT [2019]	BERT	Faster R-CNN	Dual stream	MLM+MRC+ITM	CC
LXMERT [2019]	BERT	Faster R-CNN	Dual stream	MLM+ITM+MRC+MRFR+VQA	COCO+VG+VQA
VL-BERT [2019]	BERT	Faster R-CNN+ ResNet	Single stream	MLM+MRC	CC
Pixel-BERT [2020]	BERT	ResNet	Single stream	MLM+ITM	COCO+VG
Unified VLP [2020]	UniLM	Faster R-CNN	Single stream	MLM+seq2seq LM	CC
UNIMO [2020b]	BERT, RoBERTa	Faster R-CNN	Single stream	MLM+seq2seq LM+MRC+MRFR+CMCL	COCO+CC+VG+SBU
SOHO [2021]	BERT	ResNet + Visual Dictionary	Single stream	MLM+MVM+ITM	COCO+VG
VL-T5 [2021]	T5, BART	Faster R-CNN	Single stream	MLM+VQA+ITM+VG+GC	COCO+VG
XGPT [2021]	transformer	Faster R-CNN	Single stream	IC+MLM+DAE+MRFR	CC
Visual Parsing [2021]	BERT	Faster R-CNN + Swin transformer	Dual stream	MLM+ITM+MFR	COCO+VG
ALBEF [2021a]	BERT	ViT	Dual stream	MLM+ITM+CMCL	CC+COCO+VG+SBU
SimVLM [2021b]	ViT	ViT	Single stream	PrefixLM	C4+ALIGN
WenLan [2021]	RoBERTa	Faster R-CNN + EffcientNet	Dual stream	CMCL	RUC-CAS-WenLan
ViLT [2021]	ViT	Linear Projection	Single stream	MLM+ITM	CC+COCO+VG+SBU
Dual Encoder					
CLIP [2021]	GPT2	ViT, ResNet		CMCL	self-collected
ALIGN [2021]	BERT	EffcientNet		CMCL	self-collected
DeCLIP [2021b]	GPT2, BERT	ViT, ResNet, RegNetY-64GF		CMCL+MLM+CL	CC+self-collected
Fusion Encoder+ Dual Encoder					
VLMo [2021a]	BERT	ViT	Single stream	MLM+ITM+CMCL	CC+COCO+VG+SBU
FLAVA [2021]	ViT	ViT	Single stream	MMM+ITM+CMCL	CC+COCO+VG+SBU+RedCaps

Vision-Language Models: Toward generative models

- Architecture
  - Dual encoders → CLIP & its mentioned variants
  - Encoder-decoder
  - Fusion decoder

FLAVA (Singh et al., 2021)

Holistic approach to multimodality.

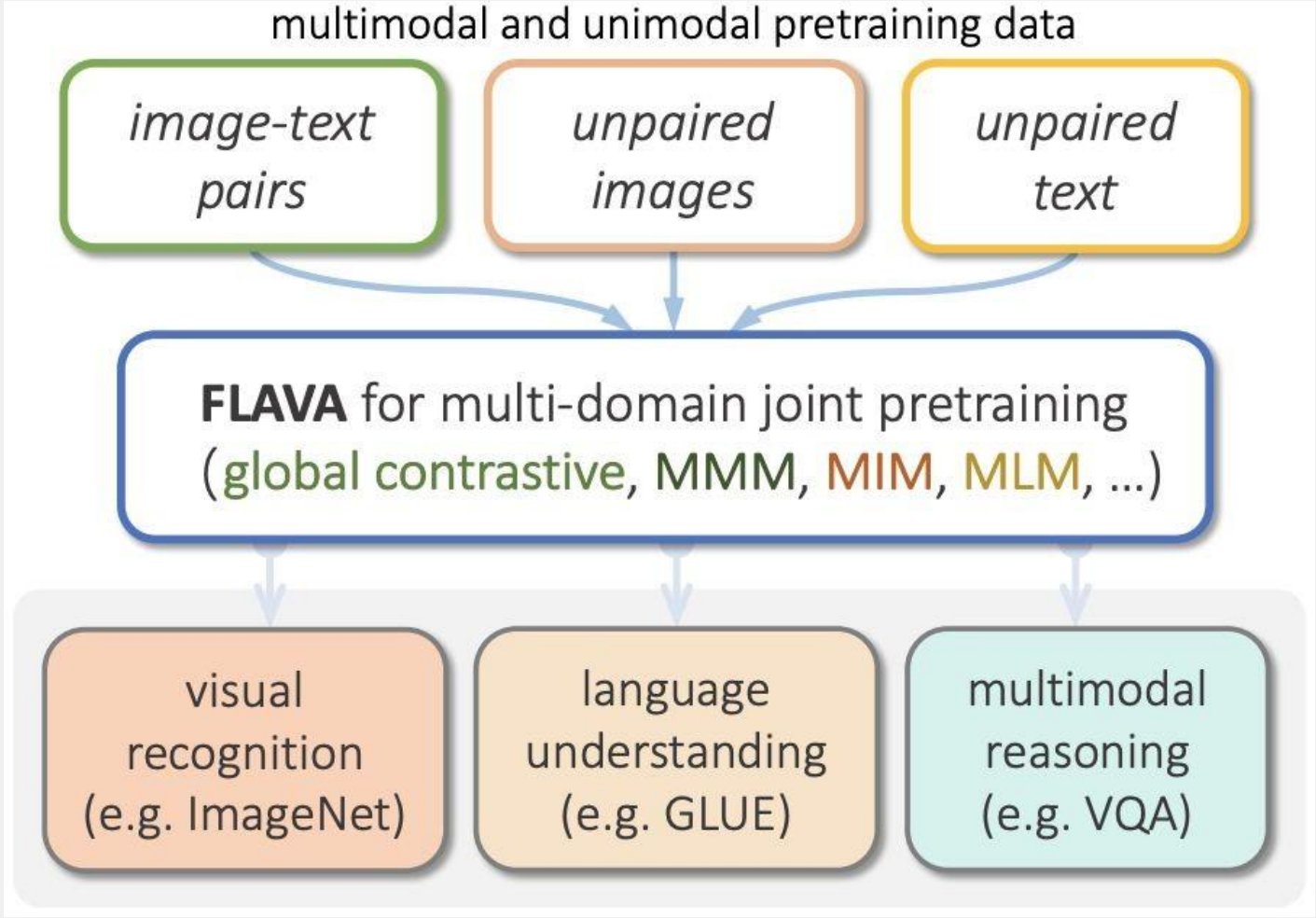
One foundation model spanning V&L, CV and NLP.

Jointly pretrained on:

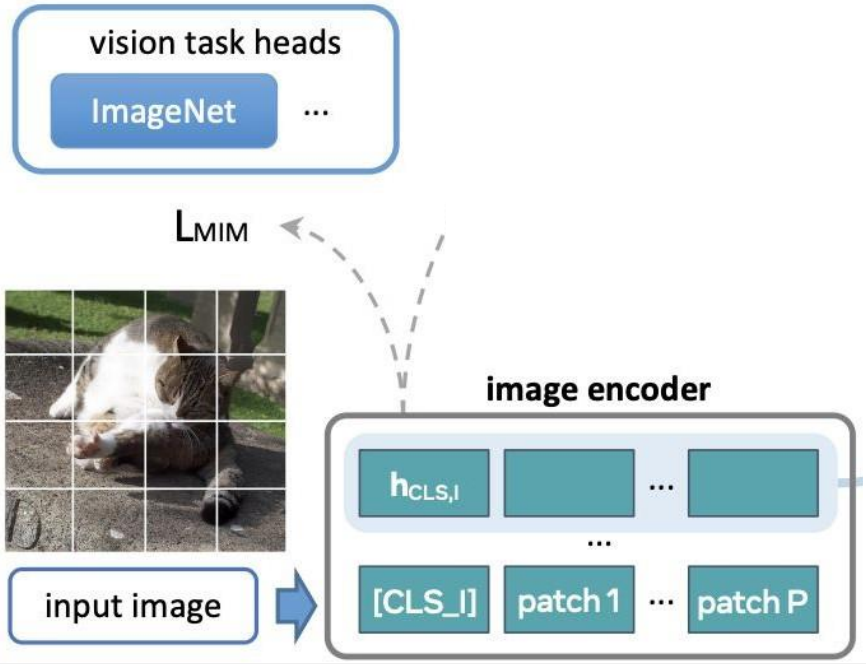
- unimodal text data (CCNews + BookCorpus)
- unimodal image data (ImageNet)
- public paired image-text data (70M)

All data/models are publicly released.

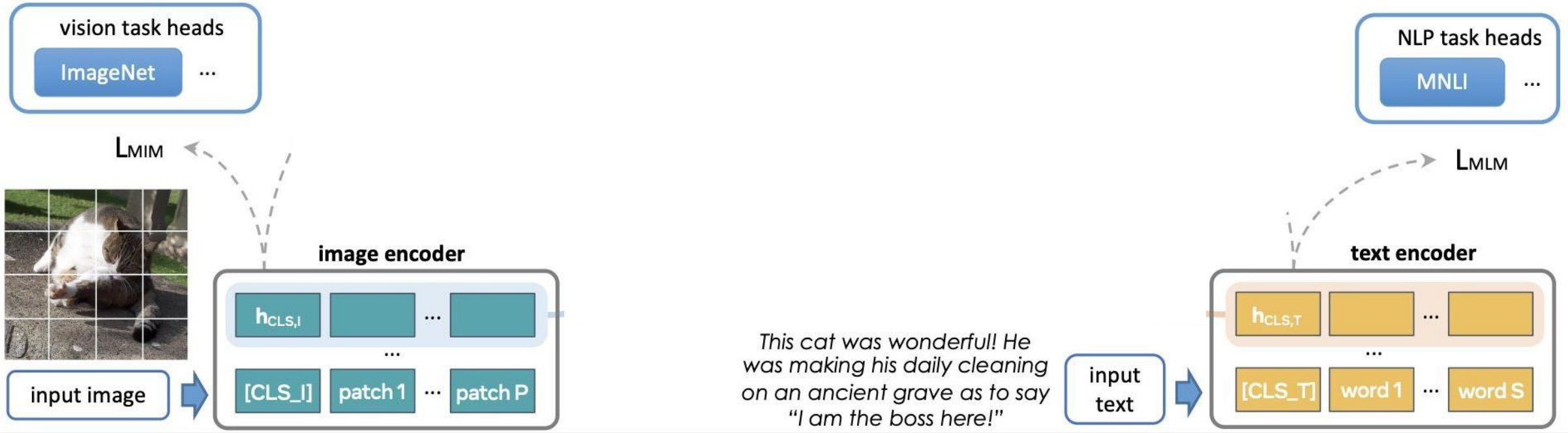
Problem to solve



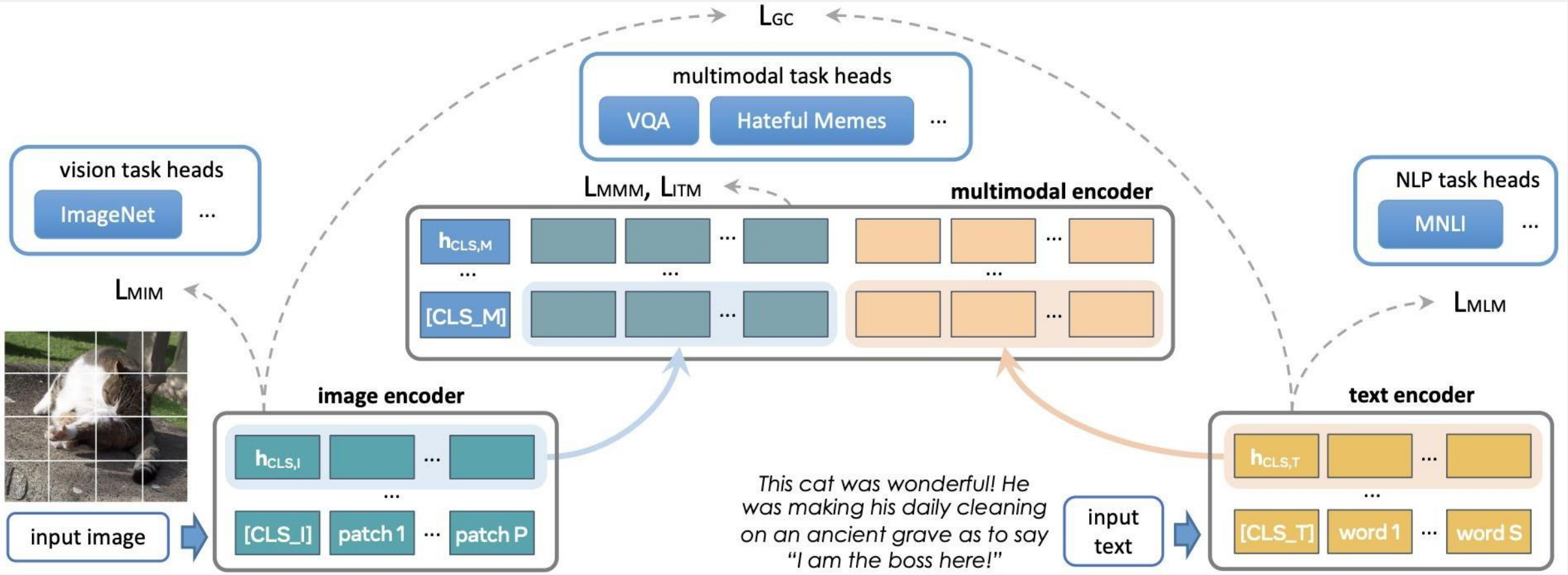
How does FLAVA work?



How does FLAVA work?

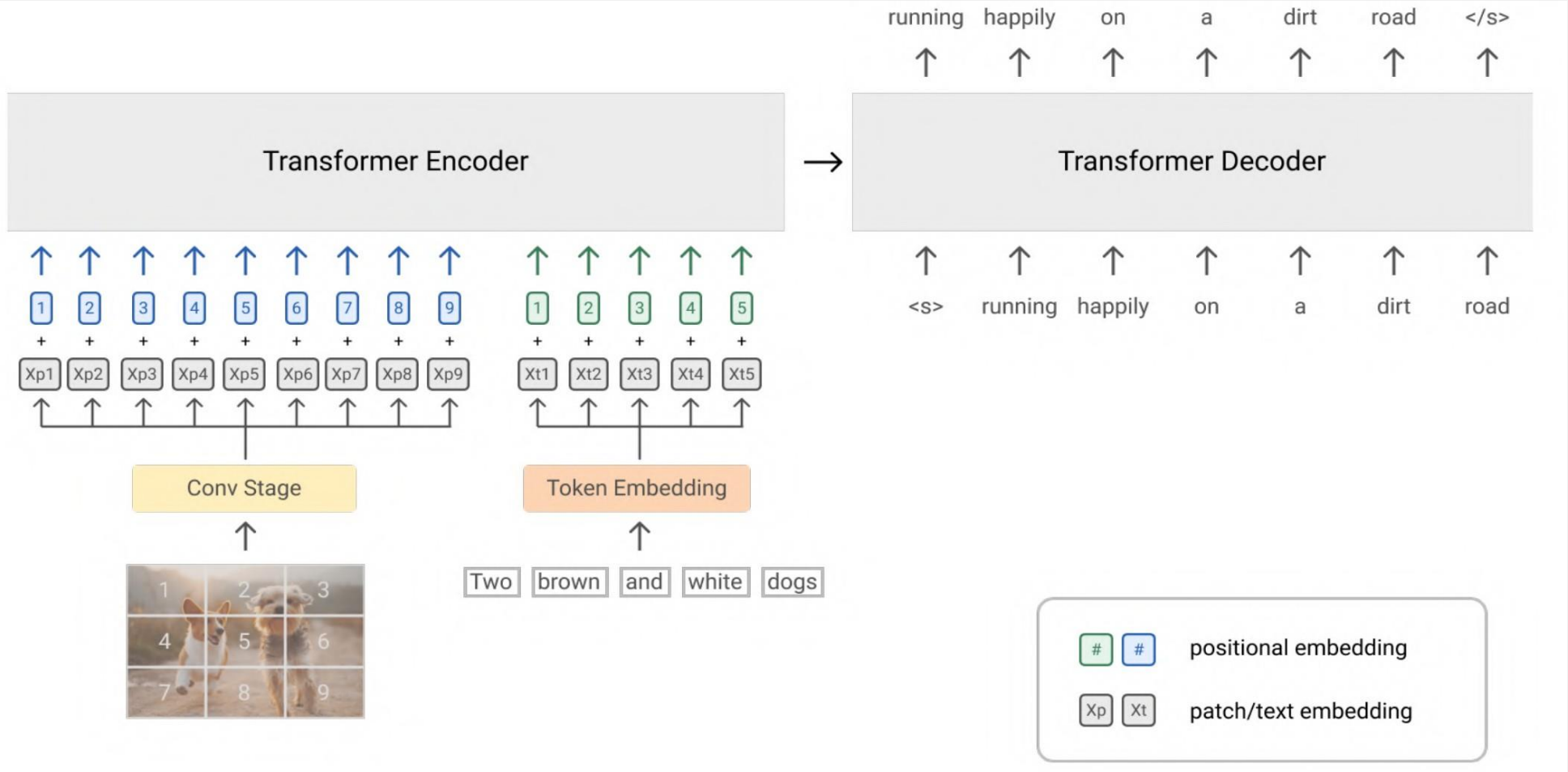


How does FLAVA work?



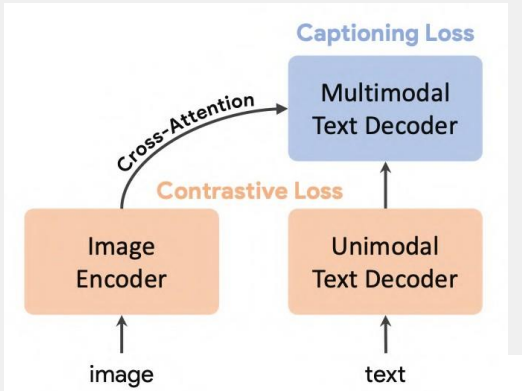
SimVLM

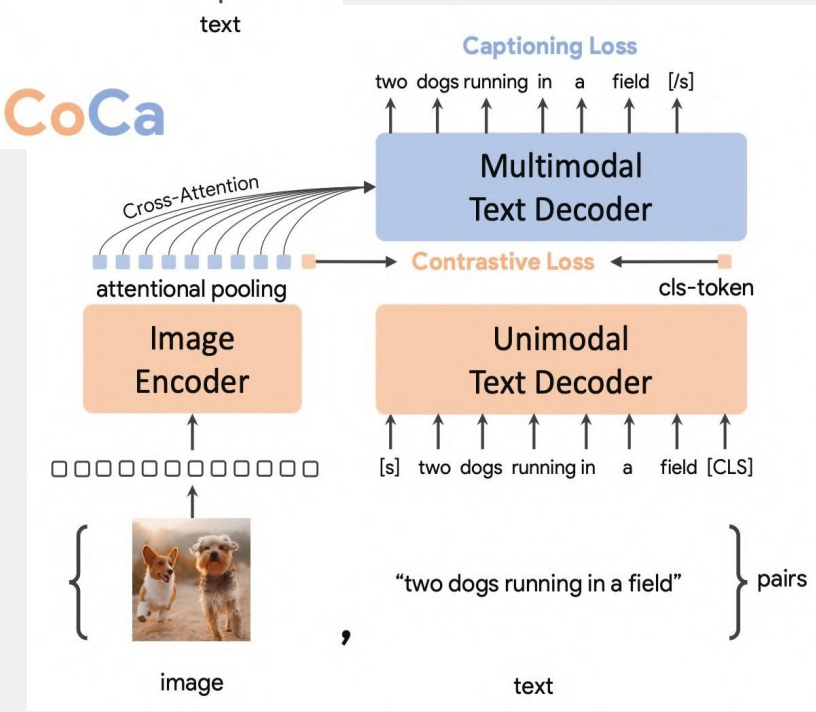
Slowly moving from contrastive/discriminative to generative.



CoCa: Contrastive Captioner

- Use mixed image-text and image-label (JFT-3B) data for pre-training
- A generative branch for enhanced performance and enabling new capabilities (image captioning and VQA)
- CoCa aims to learn a better image encoder from scratch

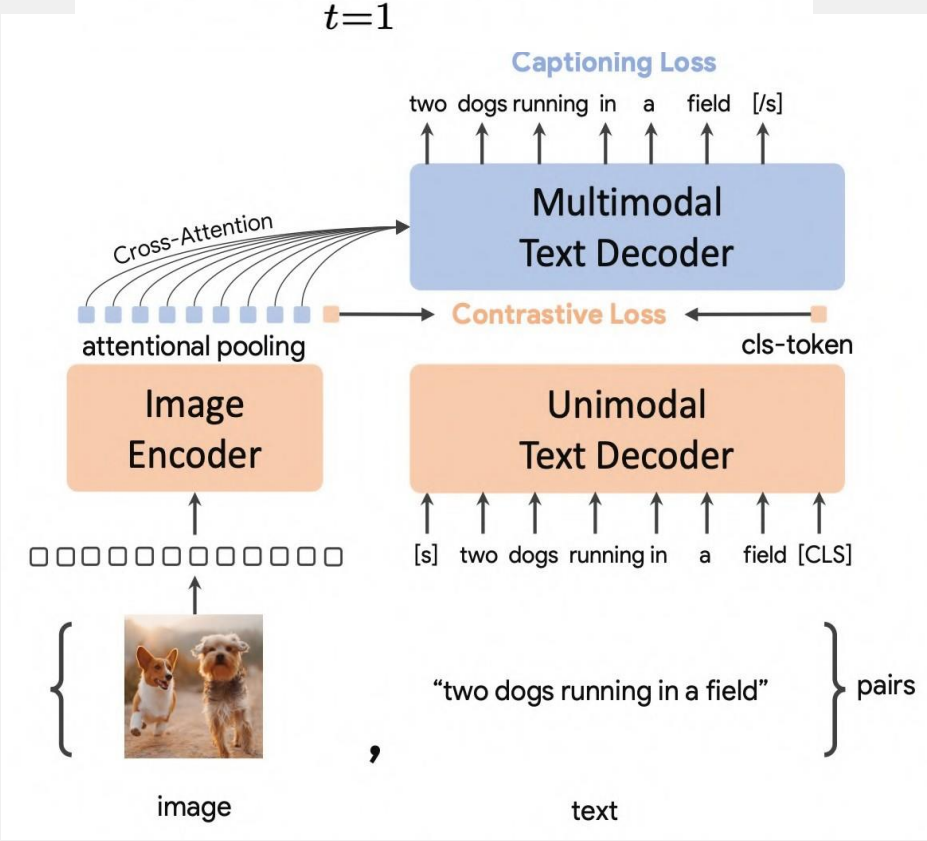




CoCa: Contrastive Captioner

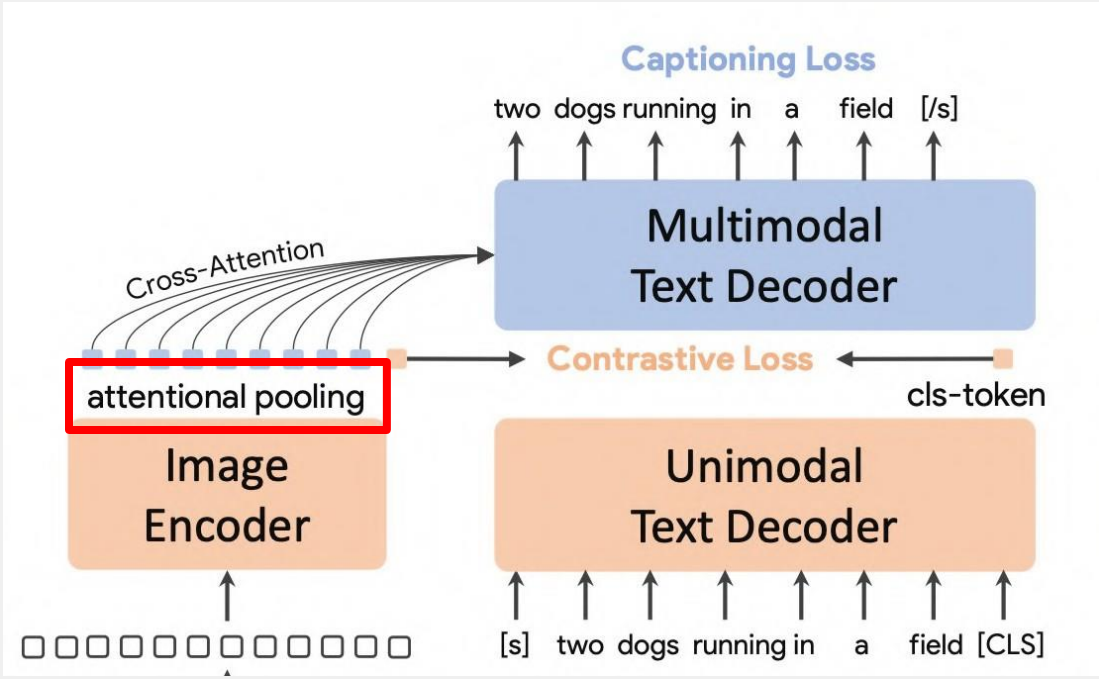
- Use mixed image-text and image-label (JFT-3B) data for pre-training
- A generative branch for enhanced performance and enabling new capabilities (image captioning and VQA)
- CoCa aims to learn a better image encoder from scratch

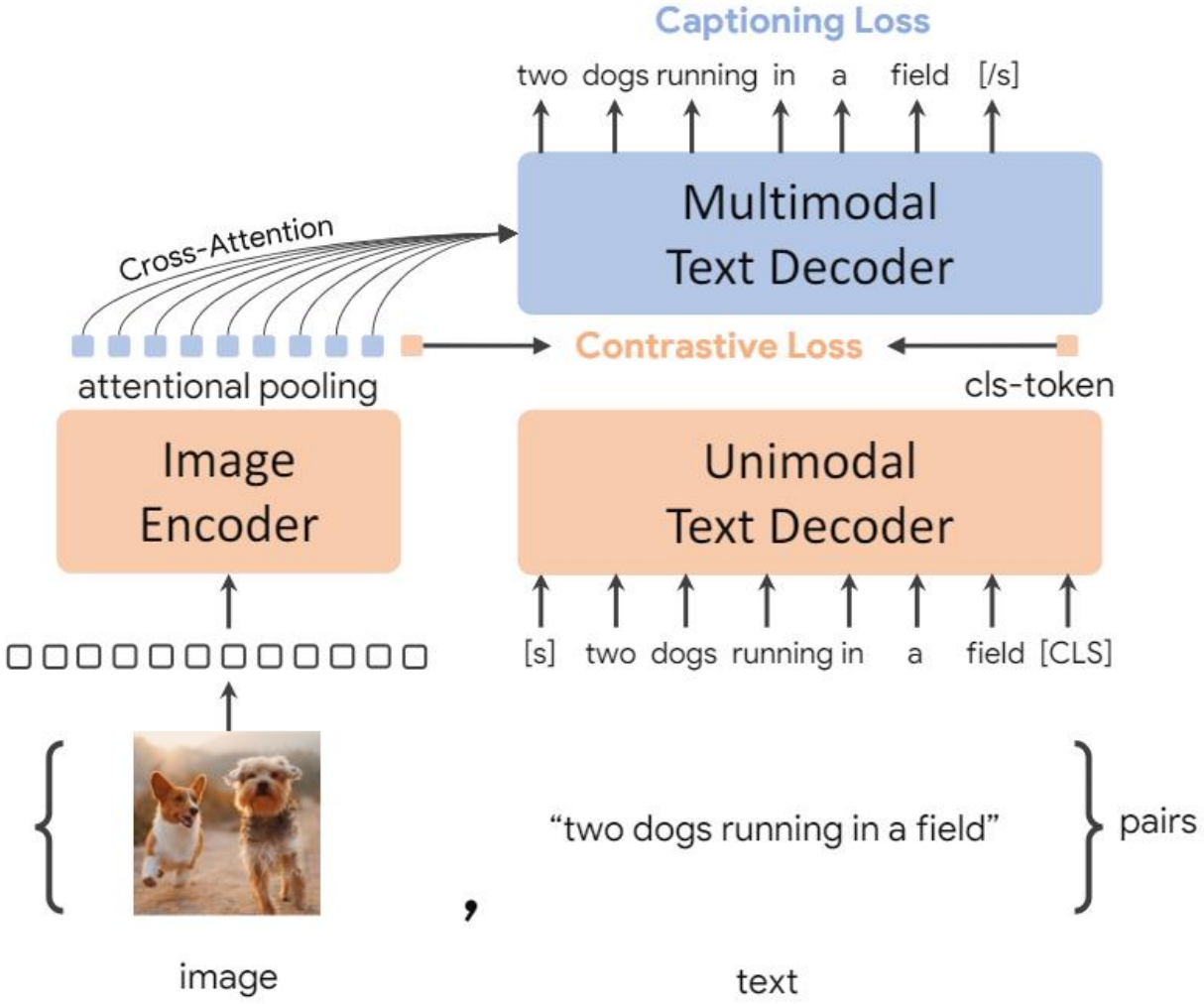
$$\mathcal{L}_{\text{Cap}} = -\sum_{t=1}^{T} \log P_\theta(y_t | y_{<t}, x).$$



CoCa Architecture

- Unified single-encoder, dual-encoder, and encoder-decoder paradigms
  - one image-text foundation model with the capabilities of all three approaches
- Cross-attention is omitted in unimodal decoder layers to encode text-only representations
- Multimodal decoder cross-attending to image encoder outputs to learn multimodal representations.





Algorithm 1 Pseudocode of Contrastive Captioners architecture.

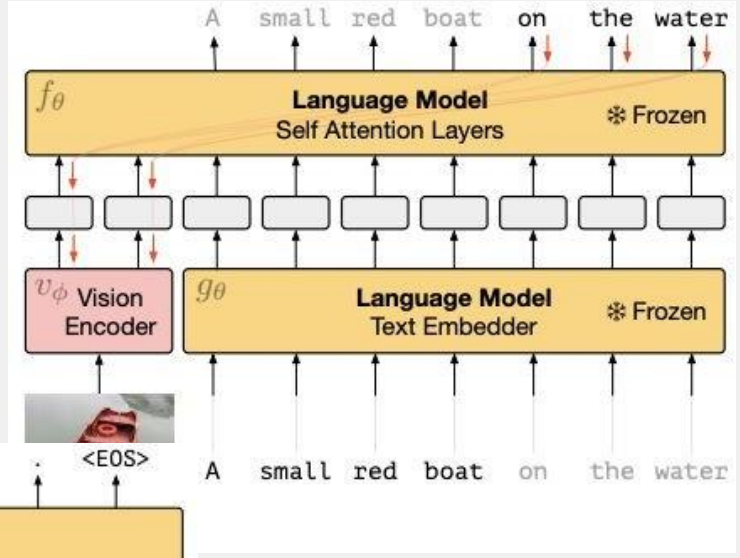
```
# image, text.ids, text.labels, text.mask: paired {image, text} data
# con_query: 1 query token for contrastive embedding
# cap_query: N query tokens for captioning embedding
# cls_token_id: a special cls_token_id in vocabulary

def attentional_pooling(features, query):
  out = multihead_attention(features, query)
  return layer_norm(out)

img_feature = vit_encoder(image) # [batch, seq_len, dim]
con_feature = attentional_pooling(img_feature, con_query) # [batch, 1, dim]
cap_feature = attentional_pooling(img_feature, cap_query) # [batch, N, dim]

ids = concat(text.ids, cls_token_id)
mask = concat(text.mask, zeros_like(cls_token_id)) # unpad cls_token_id
txt_embs = embedding_lookup(ids)
unimodal_out = lm_transformers(txt_embs, mask, cross_attn=None)
multimodal_out = lm_transformers(
    unimodal_out[:, :-1, :], mask, cross_attn=cap_feature)
cls_token_feature = layer_norm(unimodal_out)[:, -1:, :] # [batch, 1, dim]

con_loss = contrastive_loss(con_feature, cls_token_feature)
cap_loss = softmax_cross_entropy_loss(
    multimodal_out, labels=text.labels, mask=text.mask)
```

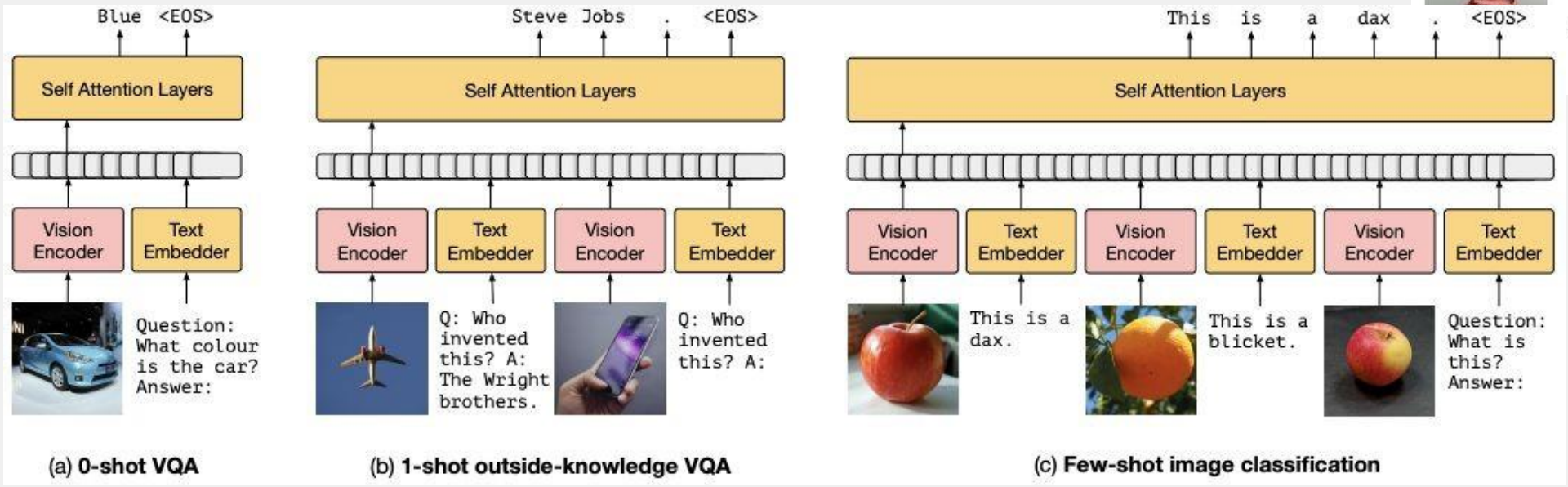
`vit_encoder`: vision transformer based encoder; `lm_transformer`: language-model transformers.

Frozen (Tsimpoukelli, Menick, Cabi, et al., 2021)

Kind of like MMBT but with a better LLM (T5) and a better vision encoder (NF-ResNet).

Multi-Modal Few-Shot Learners!



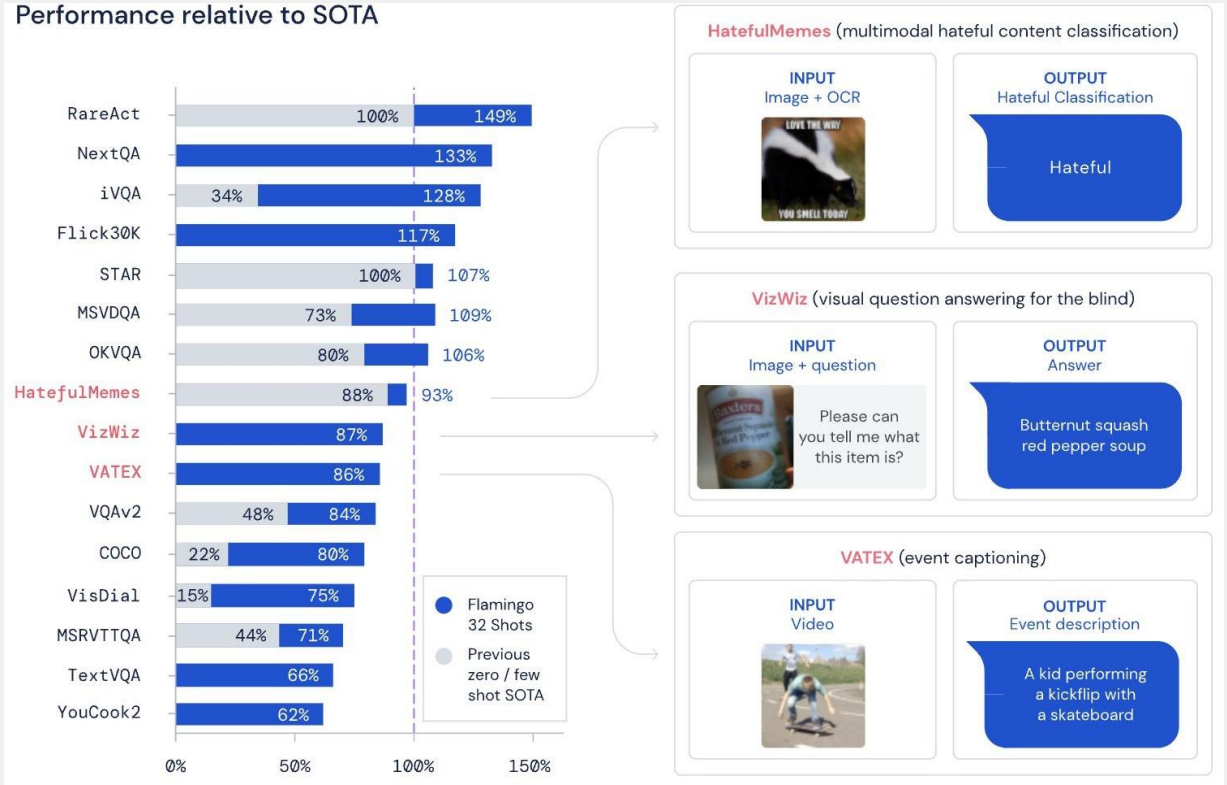


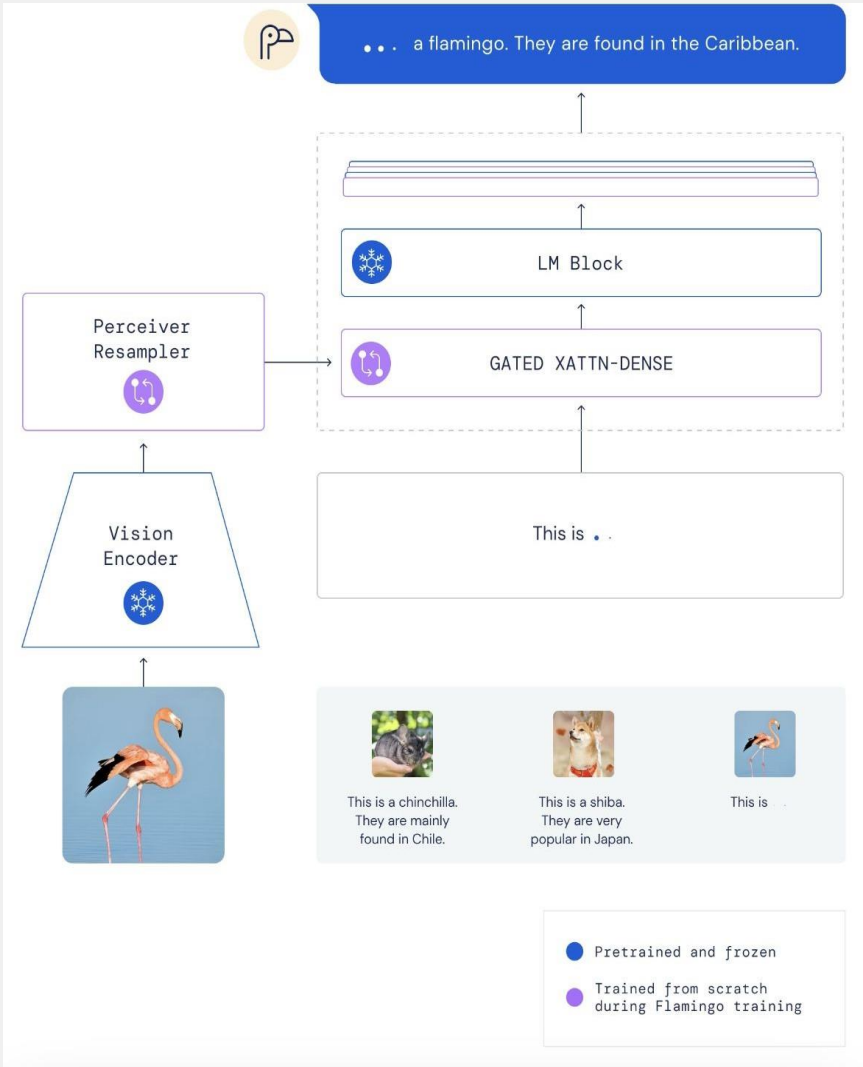
(a) **0-shot VQA** (b) **1-shot outside-knowledge VQA** (c) **Few-shot image classification**

Flamingo (Alayrac et al., 2022)

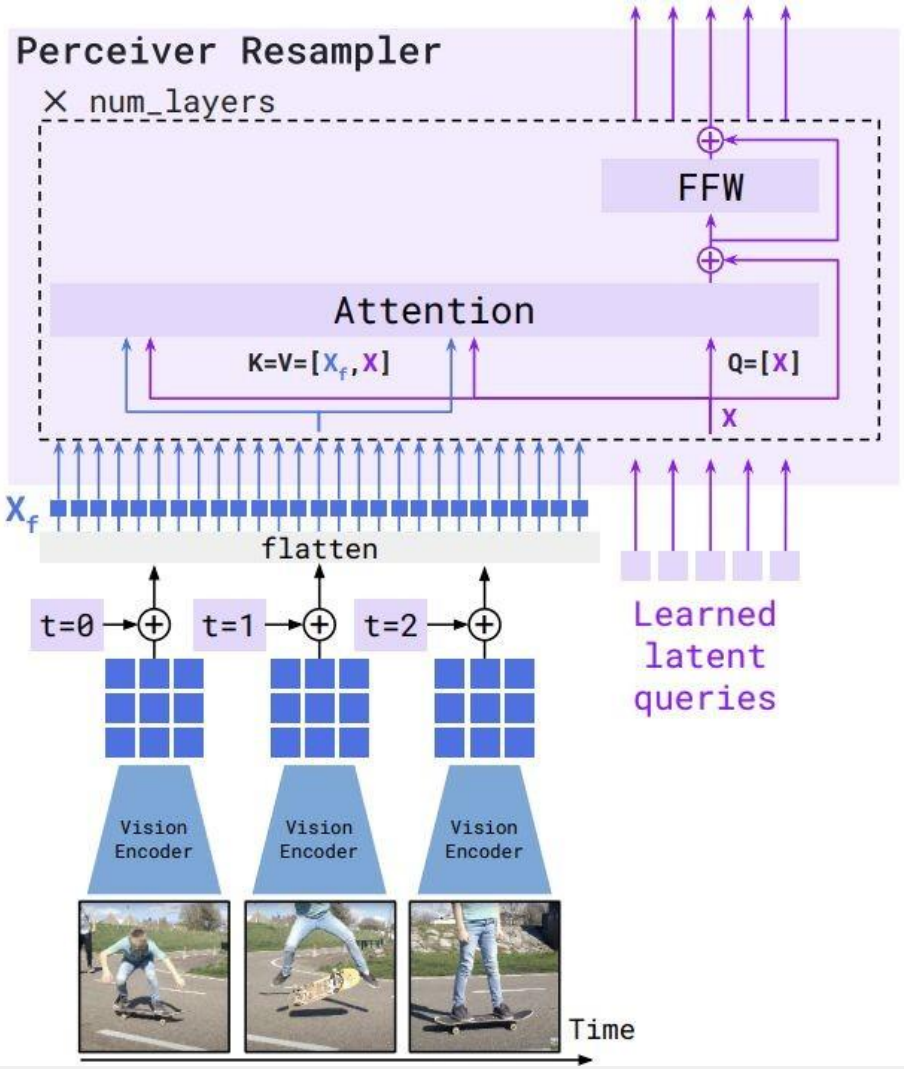
80b param model based on Chinchilla.

Multi-image.





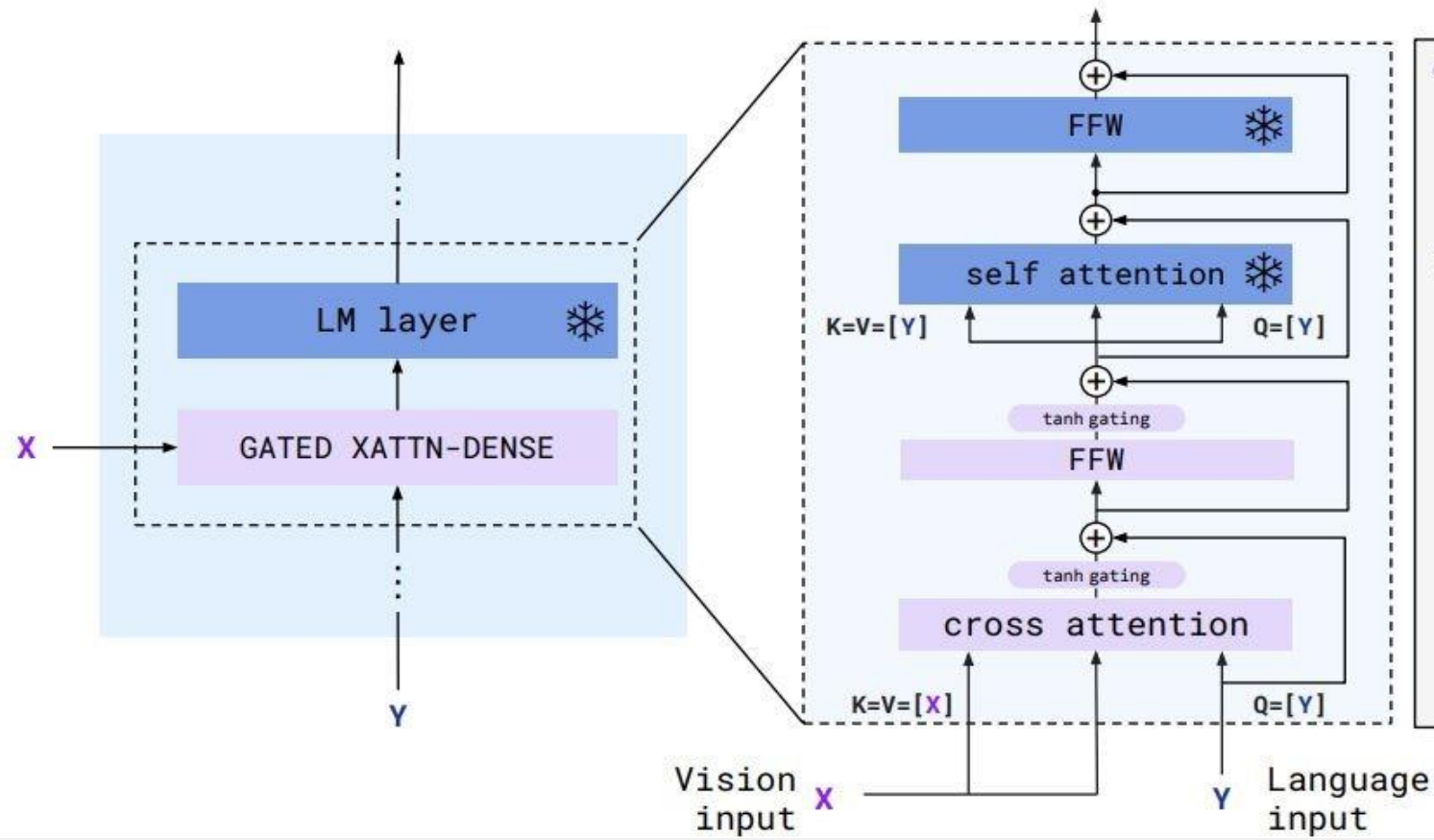
Perceiver Resampler




```python
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
```

Gated XATTN

Inject visual info directly into a frozen LM via cross-attention (remember FiLM?).




```
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 1. Gated Cross Attention
  y = y + tanh(alpha_xattn) * attention(q=y, kv=x)
  # 2. Gated Feed Forward (dense) Layer
  y = y + tanh(alpha_dense) * ffw(y)

  # Regular self-attention + FFW on language
  y = y + frozen_attention(q=y, kv=y)
  y = y + frozen_ffw(y)
  return y  # output visually informed language features
```

Why is this funny?

Original image from Karpathy as a “visual Turing test” →

This is a picture of Barack Obama. He is a former president of the United States.

How many people are there in this picture?

There are at least 5 people in this picture.

Where was this picture taken?

It was taken in a school.

How many mirrors are there?

There are at least 2 mirrors in this picture.

What is the person standing on?

The person is standing on a rug.

He is standing on a scale.

I think you are right.

What is he doing?

He is looking at the scale.

Where is Obama's foot positioned?

Obama's foot is positioned on the right side of the scale.

What happens as a result?

The scale shows a higher weight.

Is the person on the scale aware of it?

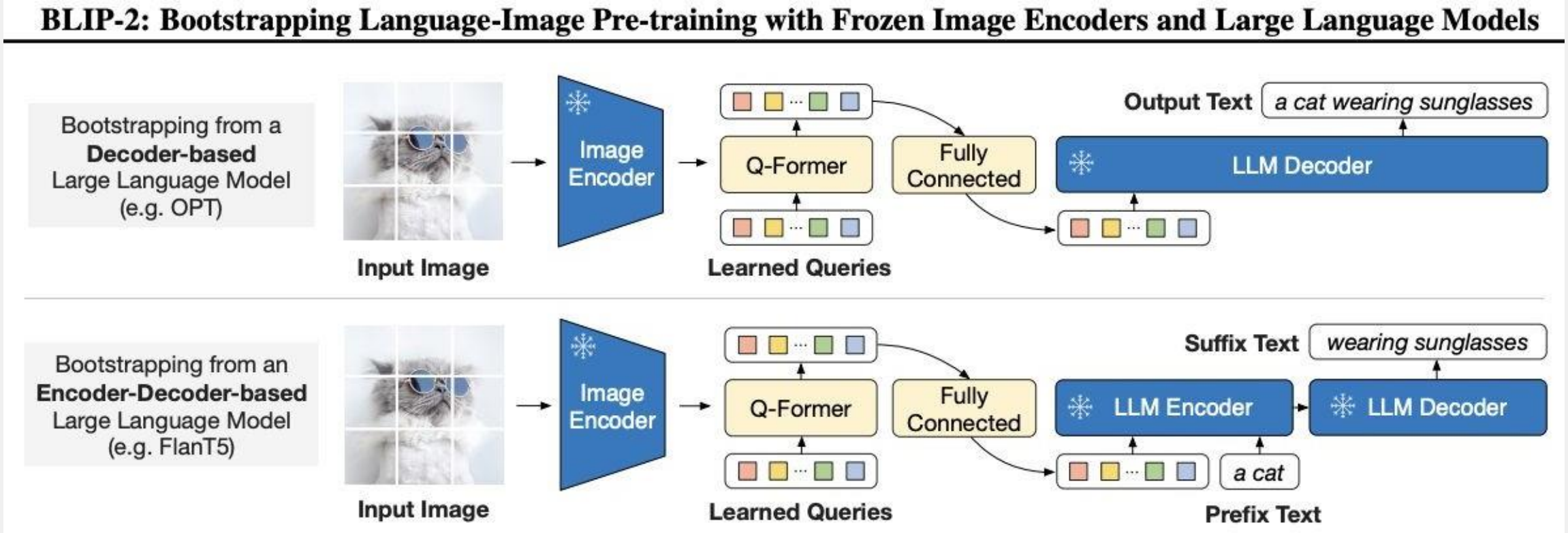
I think he is not aware of it.

Do you think that's why people are laughing?

I think so.

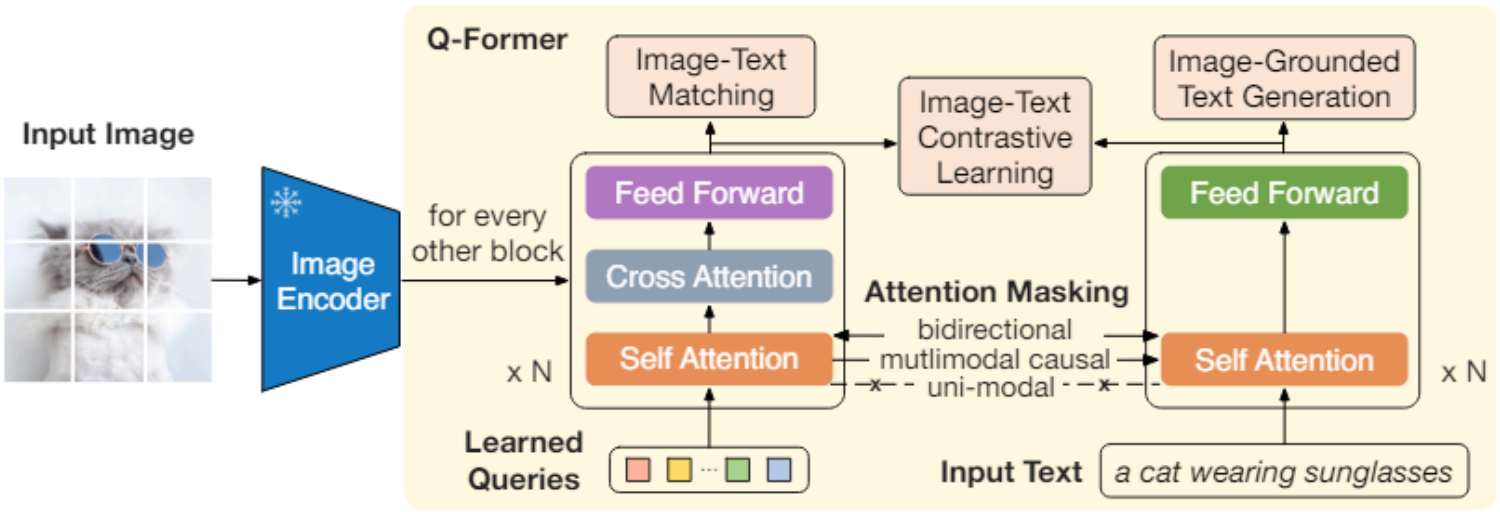
BLIP/BLIP2 (Li et al., 2023)

Freeze it all (CLIP-ViT / OPT decoder / FlanT5 encoder-decoder)



BLIP/BLIP2 (Li et al., 2023)

At training time



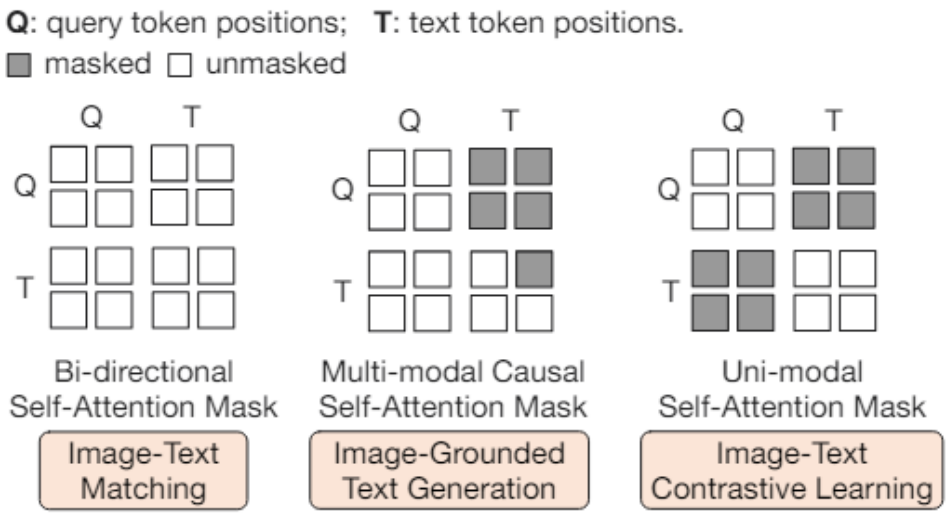
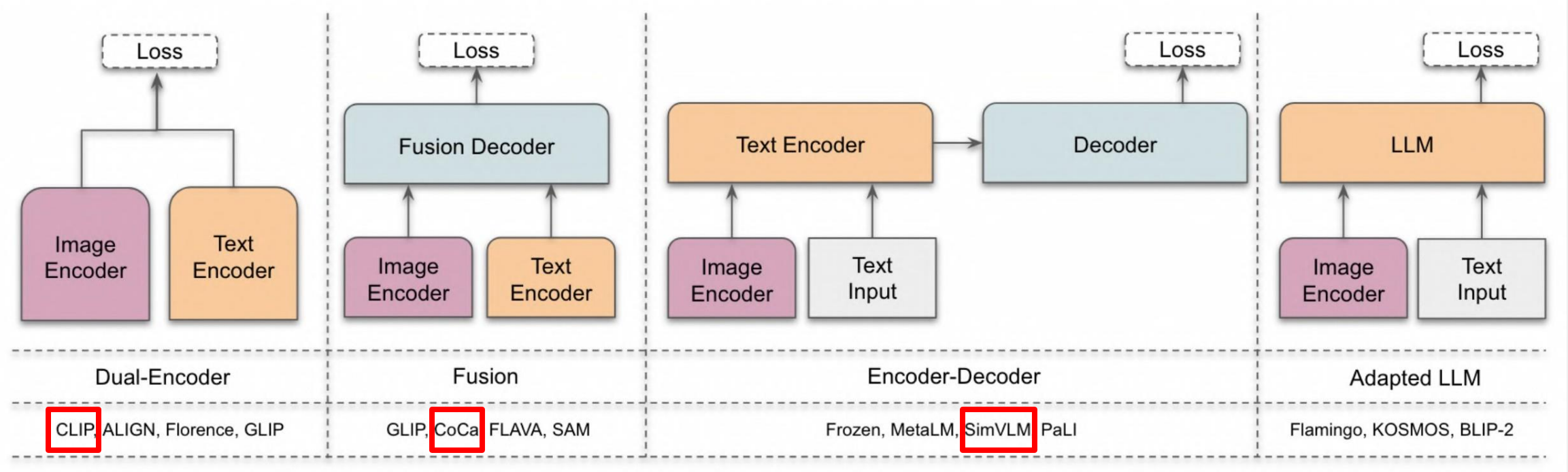
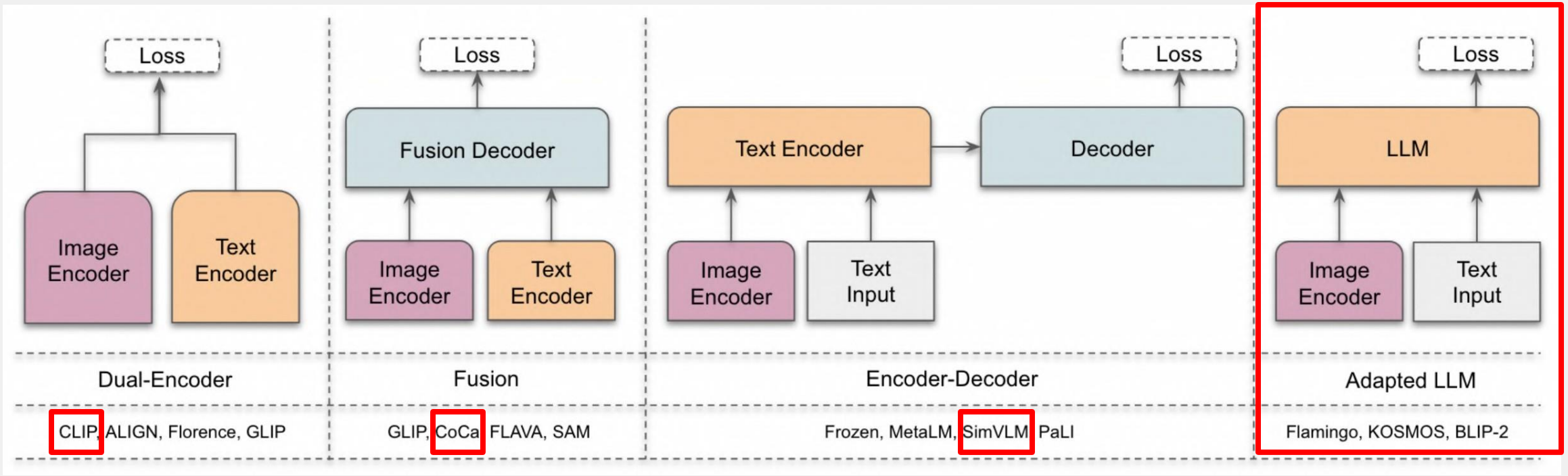


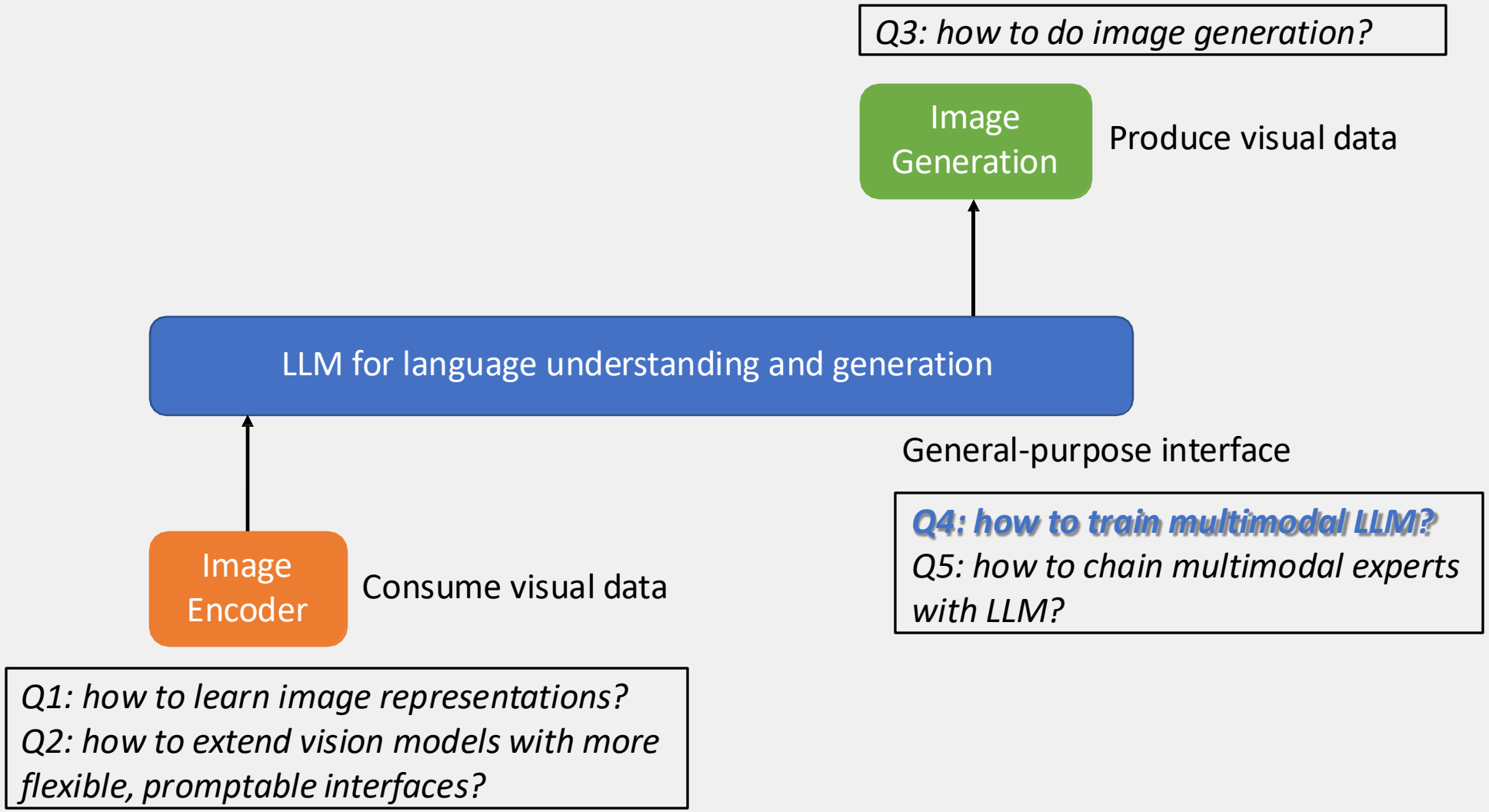
Figure 2. (**Left**) Model architecture of Q-Former and BLIP-2's first-stage vision-language representation learning objectives. We jointly optimize three objectives which enforce the queries (a set of learnable embeddings) to extract visual representation most relevant to the text. (**Right**) The self-attention masking strategy for each objective to control query-text interaction.

Architecture of Multimodal Models



Architecture of Multimodal Models





MultiModal GPT-4

- Model Details: Unknown
- Capability: Strong zero-shot visual understanding & reasoning on many user-oriented tasks in the wild
- How can we build Multimodal GPT-4 like models?

OpenAI

GPT-4 visual input example, Extreme Ironing:

User: What is unusual about this image?

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

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

GPT-4 visual input example, Chicken Nugget Map:

User: Can you explain this meme?

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

Recap on Language Modeling: Large Language Models (LLM)

OpenAI

GPT-2 → GPT-3 → ChatGPT InstructGPT → GPT-4

What's new?

GPT-3	ChatGPT InstructGPT	GPT-4
In-context-learning	In-context-learning	In-context-learning
Chain-of-thoughts (CoT)	Chain-of-thoughts (CoT)	Chain-of-thoughts (CoT)
	Instruction-Following	**Instruction-Following**
		Multimodal Input with image

Multimodal Space

Flamingo
BLIP2
GIT
…

Gap?
Instruction-Following
→ Alignment Research

Multimodal GPT-4

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

Instruction Tuning

Instruction	Input	Output
Translate English into Simplified Chinese	*Hello, Vancouver*	你好，温哥华
Summarize in just 10 words to make the message even more brief and easier to remember.	*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

Instruction Tuning

Instruction

Input → Output

Summarize in Chinese to make it easier to remember.

CVPR 2023 Visas

The CVPR'23 organizing committee deeply regrets that many members of our community could not receive visas to attend CVPR 2023. For several months, the organizers have actively raised concerns with Canadian immigration authorities (IRCC), government agencies, and politicians. In some cases, we have been successful in helping people obtain visas, but in many cases, our efforts were unsuccessful. The organizers acknowledge that the international representation of members from all over the world is what has made CVPR successful. We share in the frustration of those who were unable to attend. We continue to allow virtual to in-person registration switches for attendees who receive their visas before the conference.

*CVPR'23*签证问题: 组委会努力解决，提供虚拟和现场注册转换服务

"CVPR'23 visa issue: organizing committee works to solve and provide virtual and in-person registration switch services."

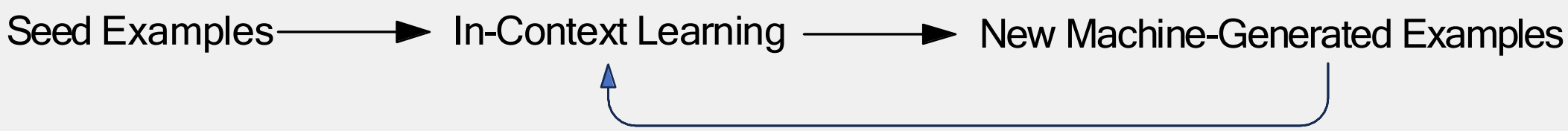
Self-Instruct Tuning

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: ….



Language Modeling: Large Language Models (LLM)

OpenAI

GPT-2 → GPT-3 → ChatGPT / InstructGPT → GPT-4

What's new?

GPT-3	ChatGPT / InstructGPT	GPT-4
In-context-learning	In-context-learning	In-context-learning
Chain-of-thoughts (CoT)	Chain-of-thoughts (CoT)	Chain-of-thoughts (CoT)
	Instruction-Following	**Instruction-Following**
		Multimodal Input with image

Open Source Community

LLaMA

Alpaca

Vicuna

GPT4-Alpaca

Tulu

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	...	

Large Multimodal Models

- Building multimodal gpt4 with open-source resources

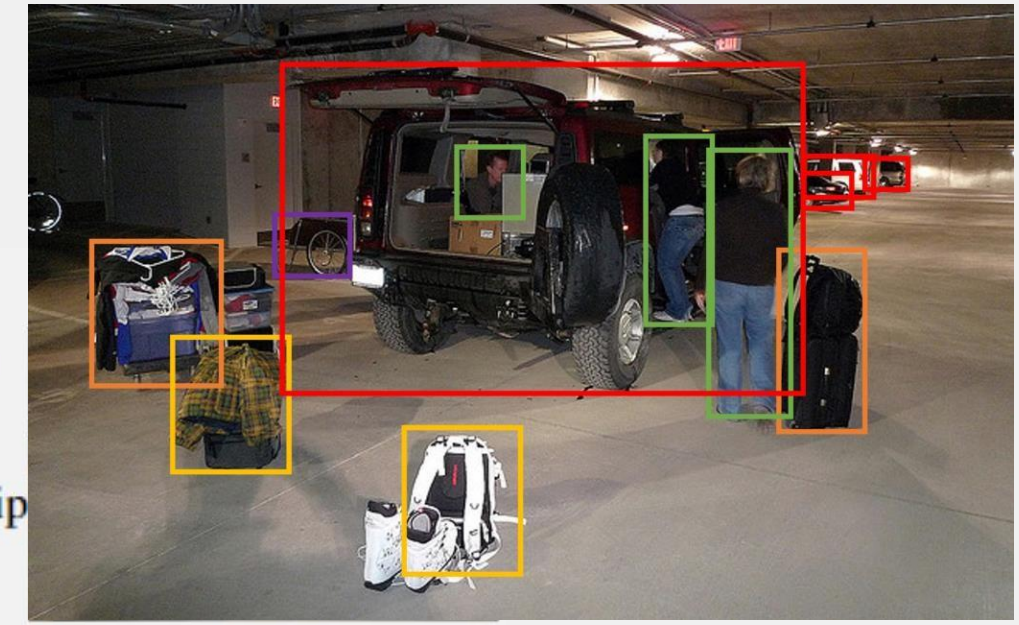
LLaVA as a running example in this lecture

- Data
- Model
- Performance

GPT-assisted Visual Instruction Data Generation

- 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]

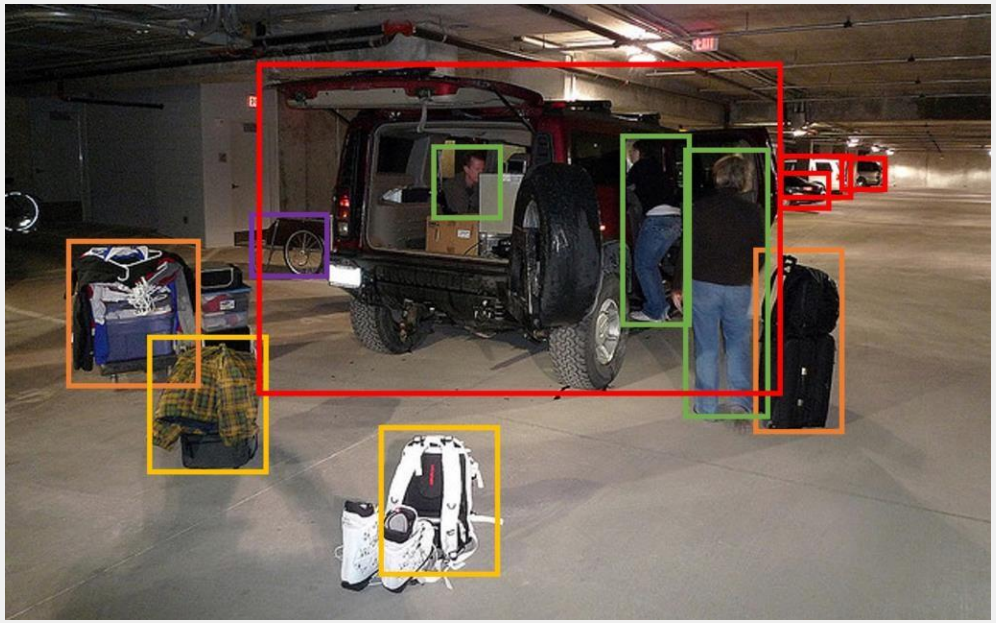
GPT-assisted Visual Instruction Data Generation

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.

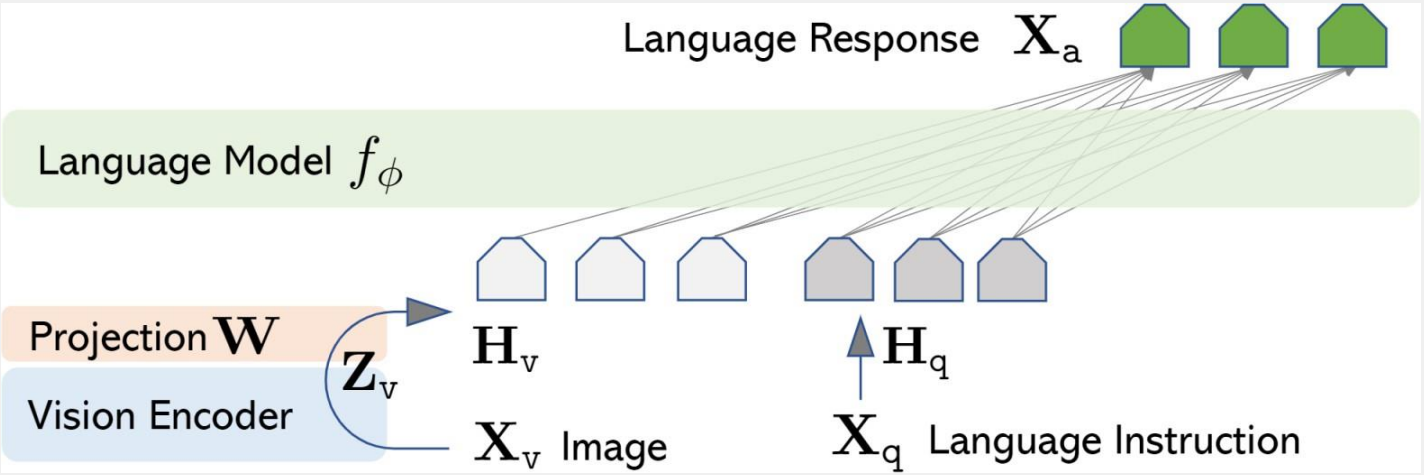
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They 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



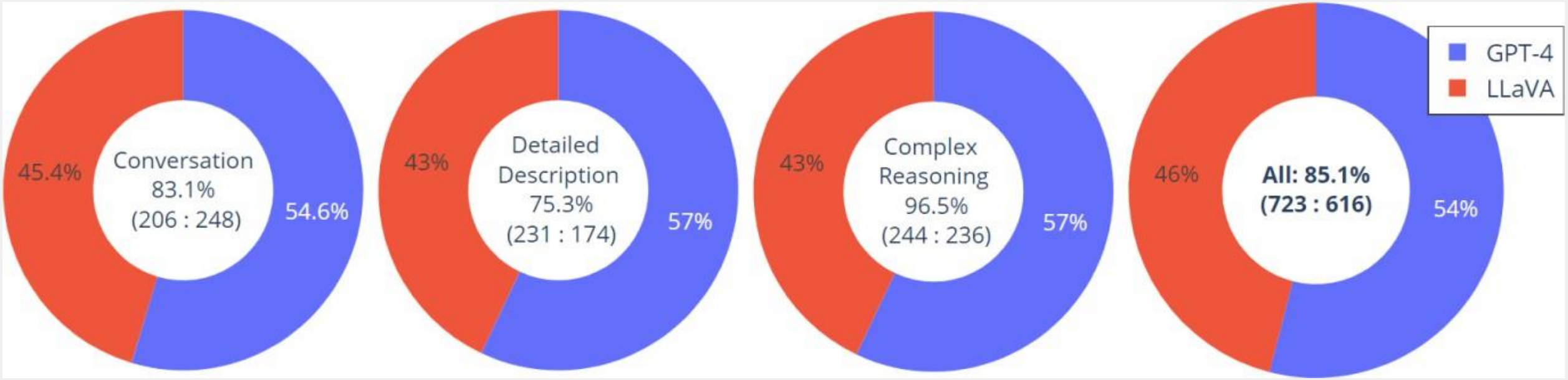
❑ Two-stage Training

- **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

Example 1: Extreme Ironing

Visual input example, Extreme Ironing:

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

User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [34]	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	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Visual input example, Chicken Nugget Map:

User	Can you explain this meme in detail?
LLaVA	The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.
User	Can you explain this meme?
GPT-4 [32]	This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.
User	Can you explain this meme in detail?
BLIP-2	sometimes i just look at pictures of the earth from space and marvel how beautiful it is
User	Can you explain this meme in detail?
OpenFlamingo	It's a picture of a chicken nugget on the International Space Station.

Example 2: Chicken Nugget Map

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.

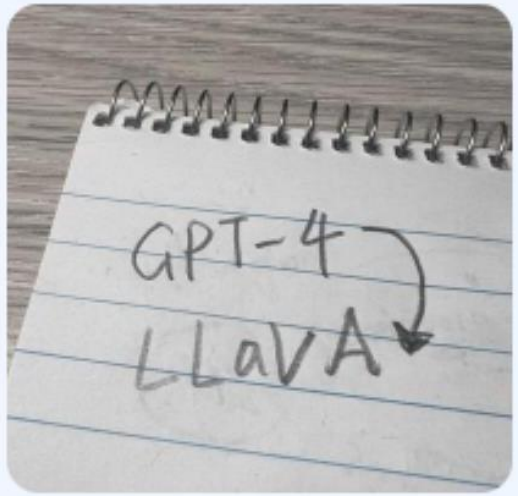
Example 1: CVPR & Vancouver

Example 2: CVinW workshop

what is image about?

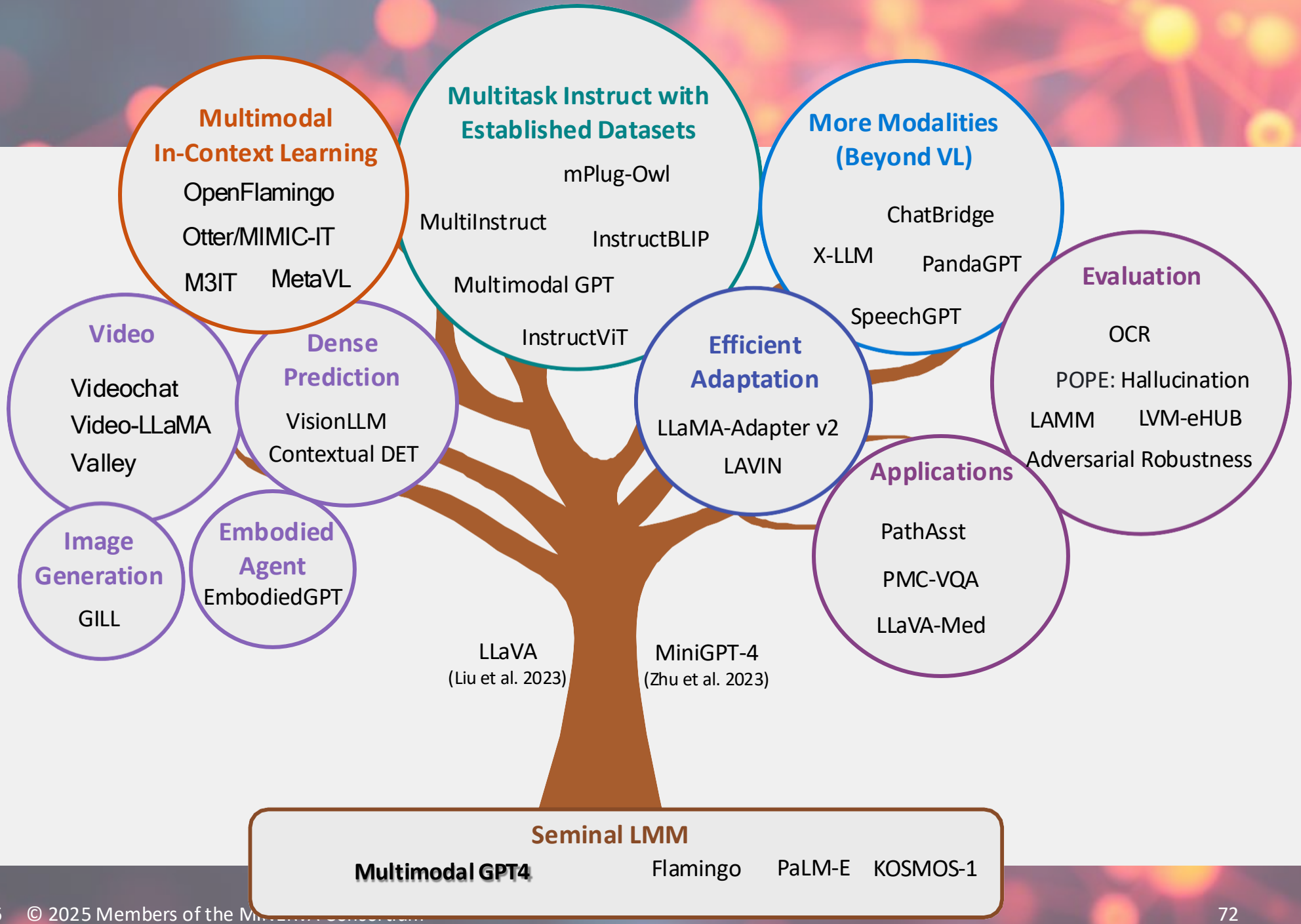
The image features a logo for "Computer Vision in the Wild," which is likely an educational or research program focused on computer vision and its applications in a wild setting. The logo consists of a large letter "C" with a smaller letter "v" on top of it, standing for "Computer Vision." The rest of the logo is not disclosed, as the picture is only showing a portion of the design. This image might represent a program or project related to the field of computer vision and how it can be used in natural environments.

what is written in the photo?



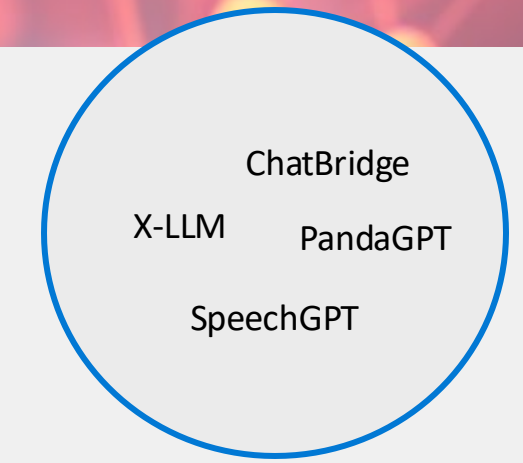
In the photo, there is written "Llava" which is presumably an abbreviation for a location or other identifier.

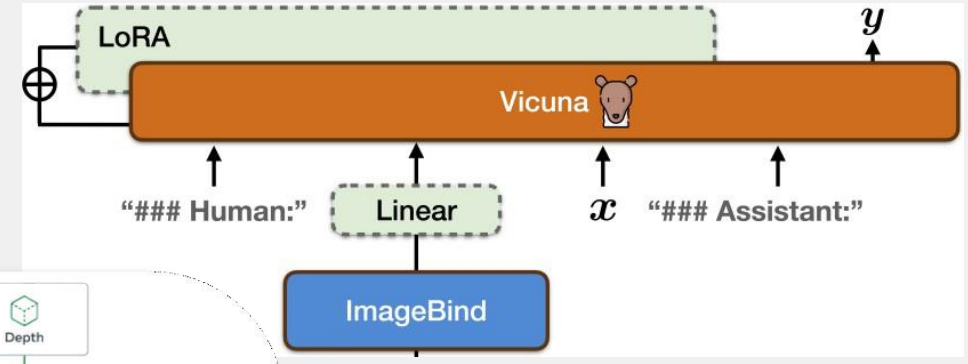
Example 3: LLaVA

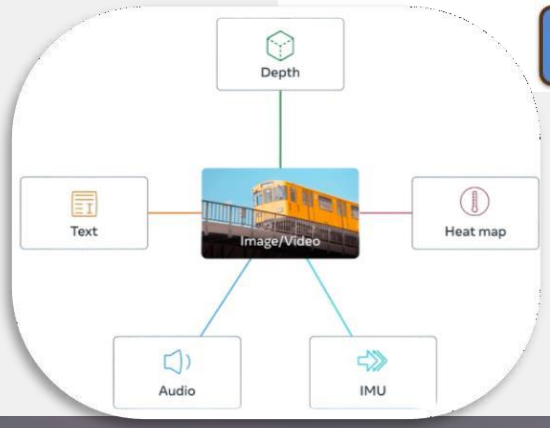


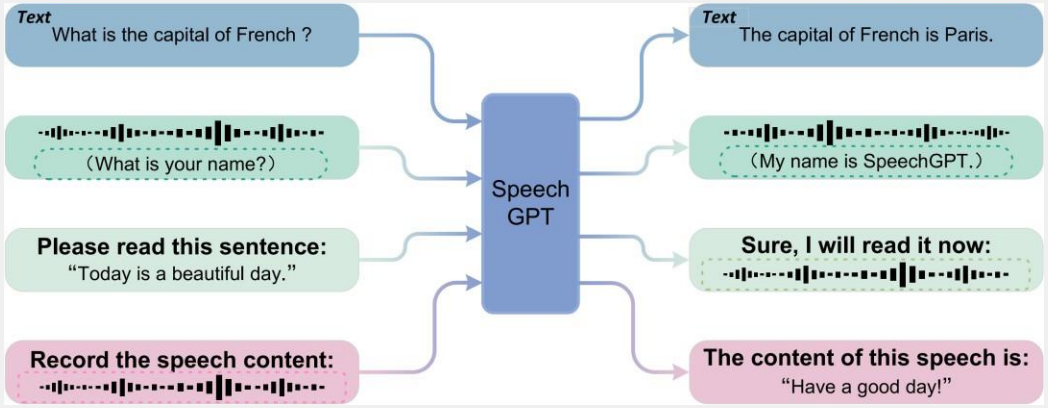
More Modalities (Beyond VL)

- 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









Applications

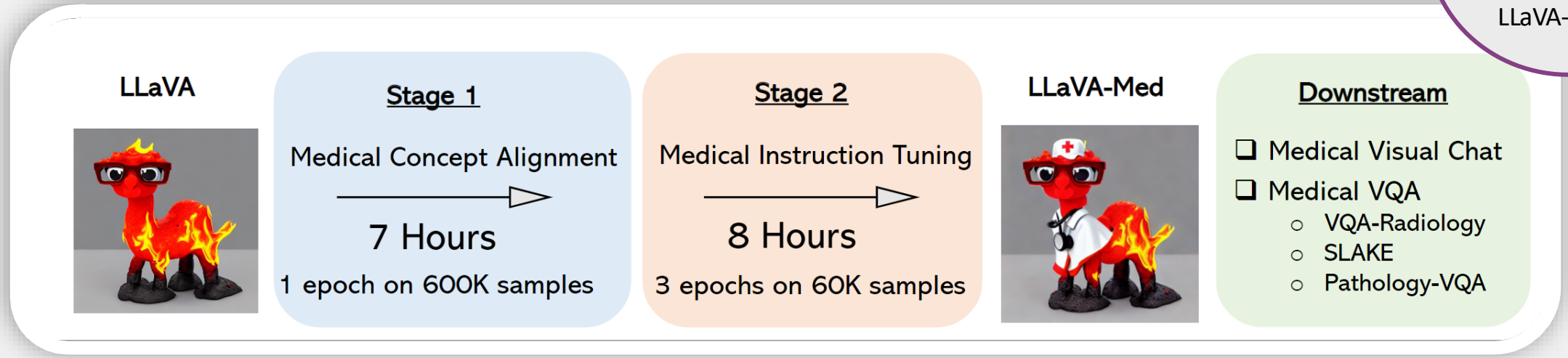
- PathAsst: Redefining Pathology through Generative Foundation AI Assistant for Pathology
- PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering
- LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

Applications

PathAsst

PMC-VQA

LLaVA-Med



User	What is this image about?
LLaVA-Med	The image is a coronal view of a computed tomography (CT) scan of the abdomen and pelvis. It provides a detailed view of the internal organs and structures within the abdominal and pelvic cavities.
User	What is abnormal about this image?
LLaVA-Med	The abnormality in this image is the presence of a large left-sided pelvic mass. This mass could be due to various reasons, such as a tumor, cyst, or other growth. Further evaluation and tests would be needed to determine the exact cause and appropriate treatment for this abnormality.

User	What is this image about?
LLaVA	The image is a medical illustration showing a person's internal anatomy, especially focusing on their large and small intestines, stomach, liver and kidneys. The organs are highlighted in different colors, making it easy to identify each organ
User	What is abnormal about this image?
LLaVA	As a visual AI assistant, I cannot determine the specific abnormality on the chest X-ray based on the provided image description. The image description mainly focuses on the organs and their colors...

Reach out to us @ *info@minerva4ai.eu*

Thank you

Co-funded by the European Union

EuroHPC Joint Undertaking

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101182737. The JU receives support from the Digital Europe Programme.