

Towards Language Models That Can See: Computer Vision Through the LENS of Natural Language

**William Berrios[†] Gautam Mittal^{†§} Tristan Thrush^{†§}
Douwe Kiela^{†§} Amanpreet Singh[†]**
[†]Contextual AI; [§]Stanford University

arXiv:2306.16410



帮我把图片处理成1920*1080分辨率



chicago.jpg

```
1 # Open the image file
----> 2 img = Image.open('/mnt/data/chicago.jpg')
4 # Resize the image
5 img_resized = img.resize((1920, 1080))
```

NameError: name 'Image' is not defined

看起来我忘记了导入必要的Python库。我需要首先导入Python的图像处理库PIL。让我修正这个错误并重新运行代码。

Finished working

Hide work ^

python

Copy code

```
from PIL import Image

# Open the image file
img = Image.open('/mnt/data/chicago.jpg')

# Resize the image
img_resized = img.resize((1920, 1080))

img_resized
```

RESULT

```
<PIL.Image.Image image mode=RGB size=1920x1080 at 0x7ECD244956D0>
```



我已经将图片的分辨率改变为了1920x1080。接下来，我将保存这个图片，然后将文件链接提供给您。

Finished working

Show work v

LZ

帮我把图片裁剪成兔子的形状



chicago.jpg



非常抱歉，我当前的能力并不包括裁剪出特定形状（例如兔子形状）。需要使用图像识别和处理技术，以便能够识别并裁剪出特定形状。形状的模板，我可以帮你按照这个模板裁剪图片。这需要模板图像黑色，而你想要删除的部分是白色。你是否有这样的模板图像？

LZ

帮我把这张图片转换为梵高的《星月夜》的风格



chicago.jpg

T

我明白你的要求，你希望将这张图片进行风格转换，使其风格接近梵高的《星月夜》。然而，我很抱歉地告诉你，虽然有一些深度学习的技术，如神经风格迁移，可以实现这种风格转换，但在当前的环境中，我无法执行这样的任务。因为这需要训练神经网络模型，并且需要使用到GPU等资源，以及互联网访问权来下载预训练的模型。这些超出了我目前的能力。



臣妾做不到啊

LZ

这张图片的内容是什么



chicago.jpg

T

我无法直接查看您上传的图片，但我可以使用计算机视觉技术来解析图片的内容。然而，我必须指出，虽然人工智能在图像识别方面已经取得了很大的进步，但它仍然无法与人类的视觉识别能力相比，特别是在理解图片的具体内容和上下文方面。

但是，由于当前的环境限制，我无法访问互联网，因此无法使用像Google Vision或Amazon Rekognition这样的云端图像识别服务来识别图片内容。这些服务能提供详细的图片内容分析，包括对象检测、场景识别、面部识别和文本识别等。

Content

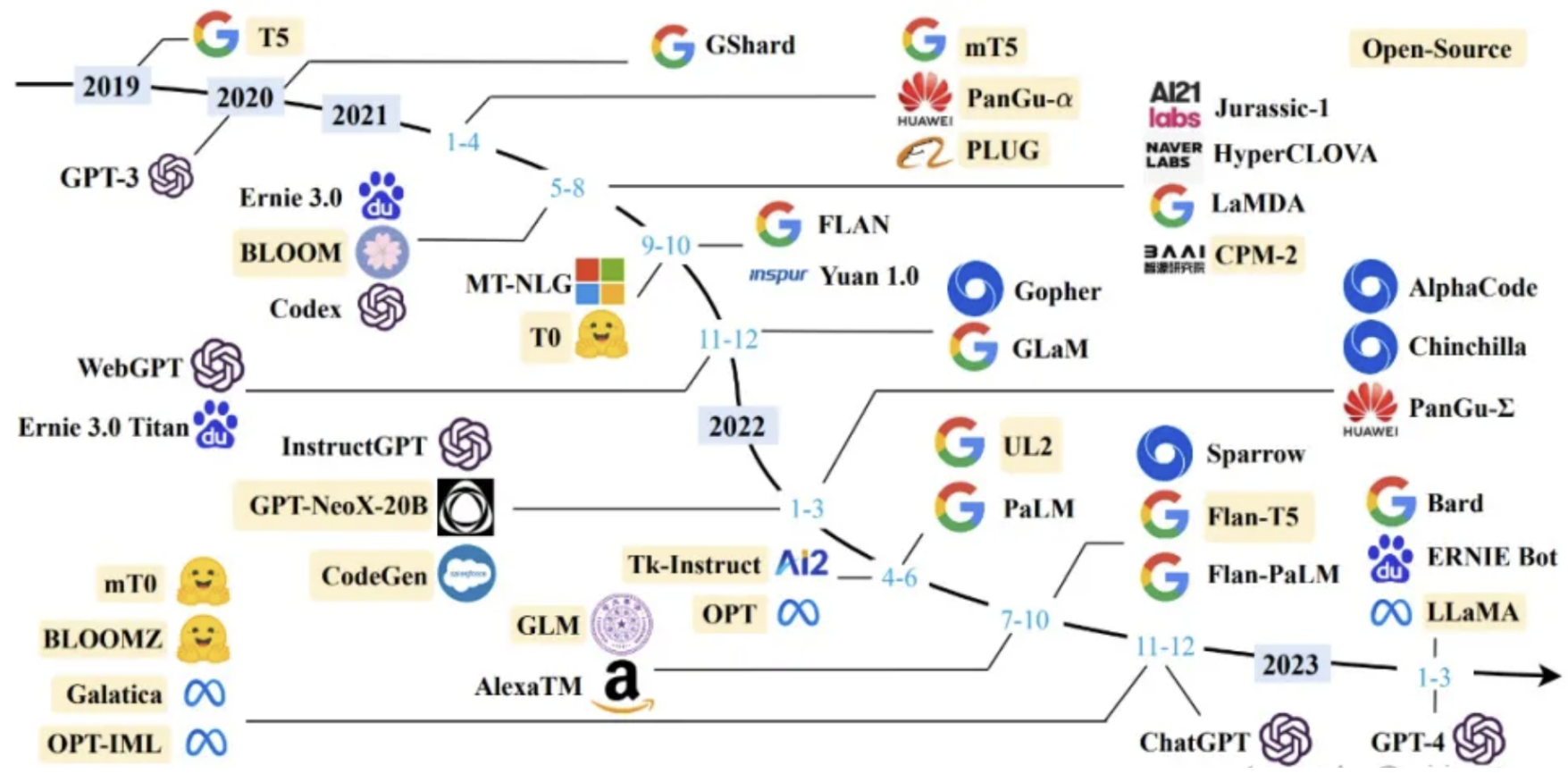
- Background
- Method
- Experiments

Content

- **Background**
- Method
- Experiments

Background

Large Language Models (LLMs)



Background

Large Language Models (LLMs)

- Large amount of text data
- Billions of parameters
- Extensive text related tasks
 - Text summary
 - Translation
 - Emotional analysis
- Zero-shot and few-shot

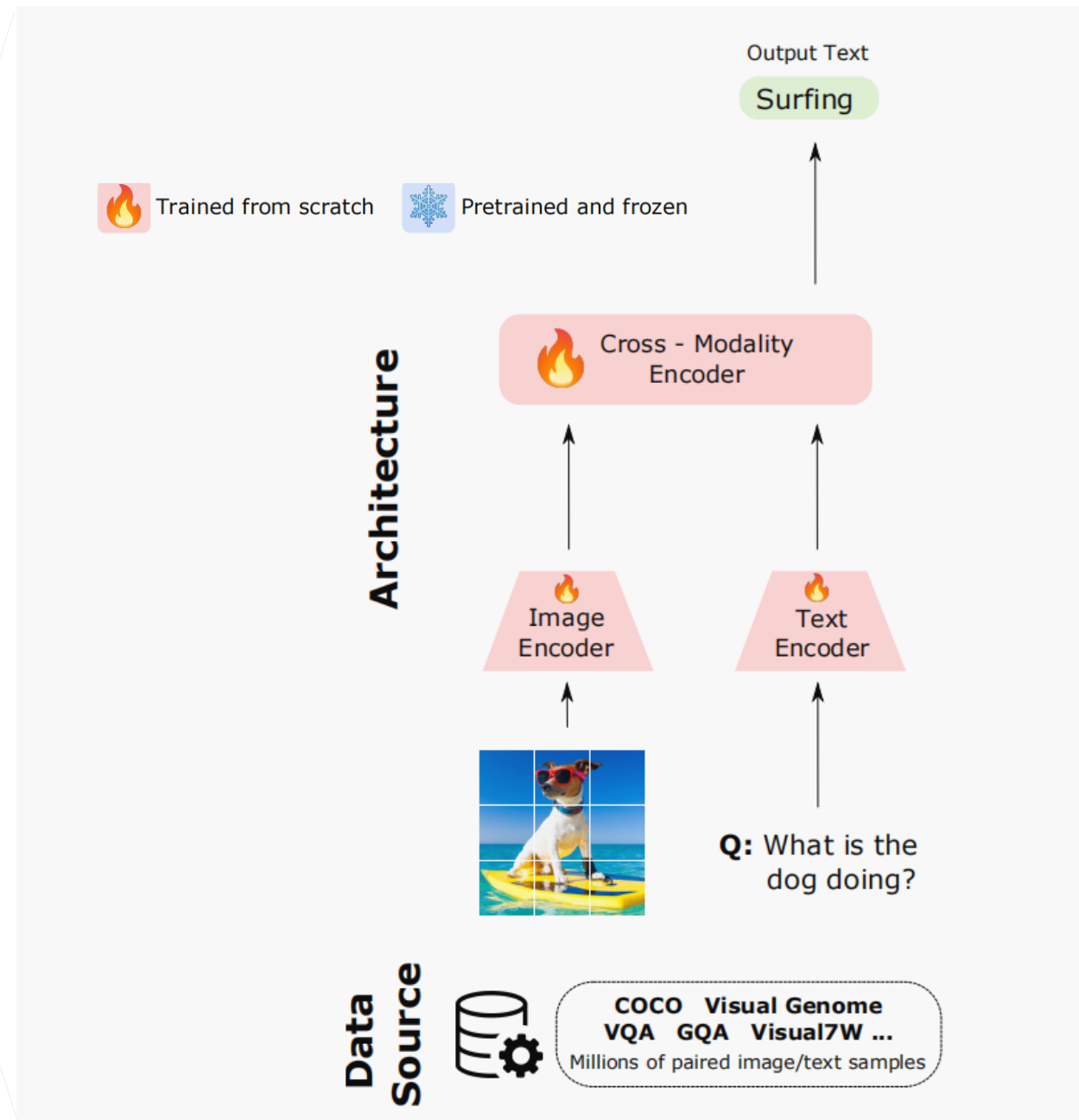
Background

Employing LLMs on vision-related tasks?



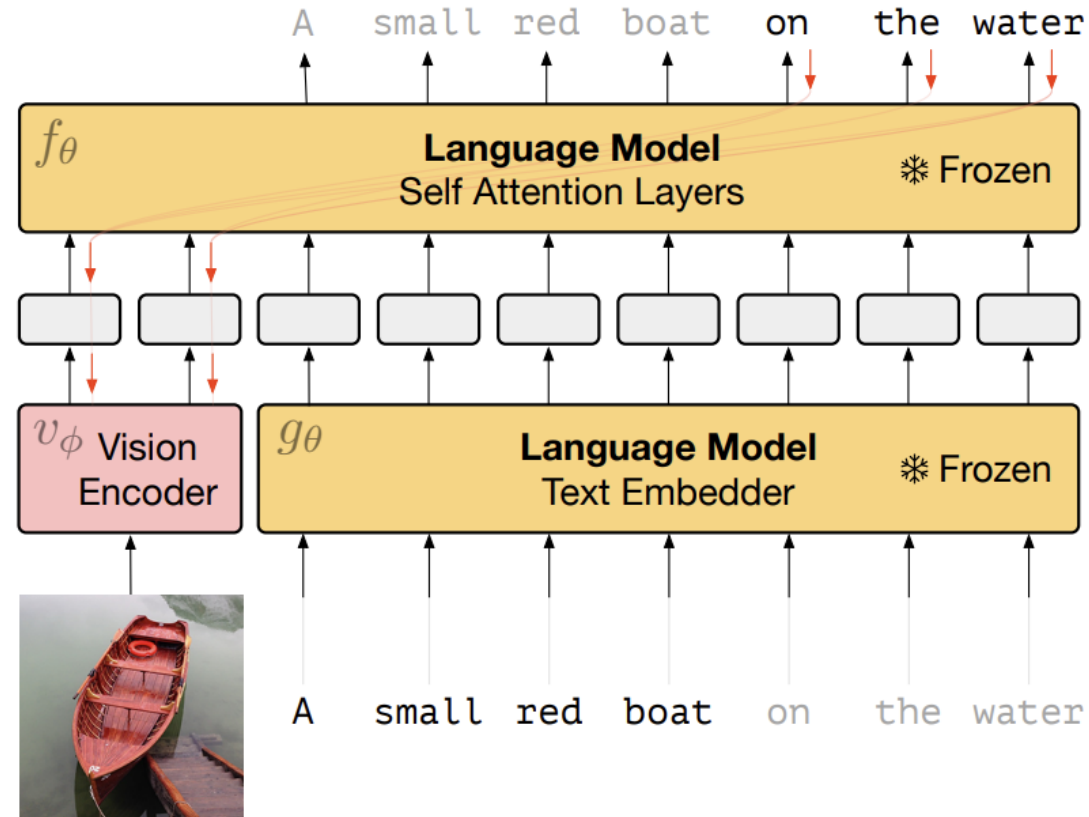
Background

Old-Style Pretraining



Background

Frozen



Background

Frozen

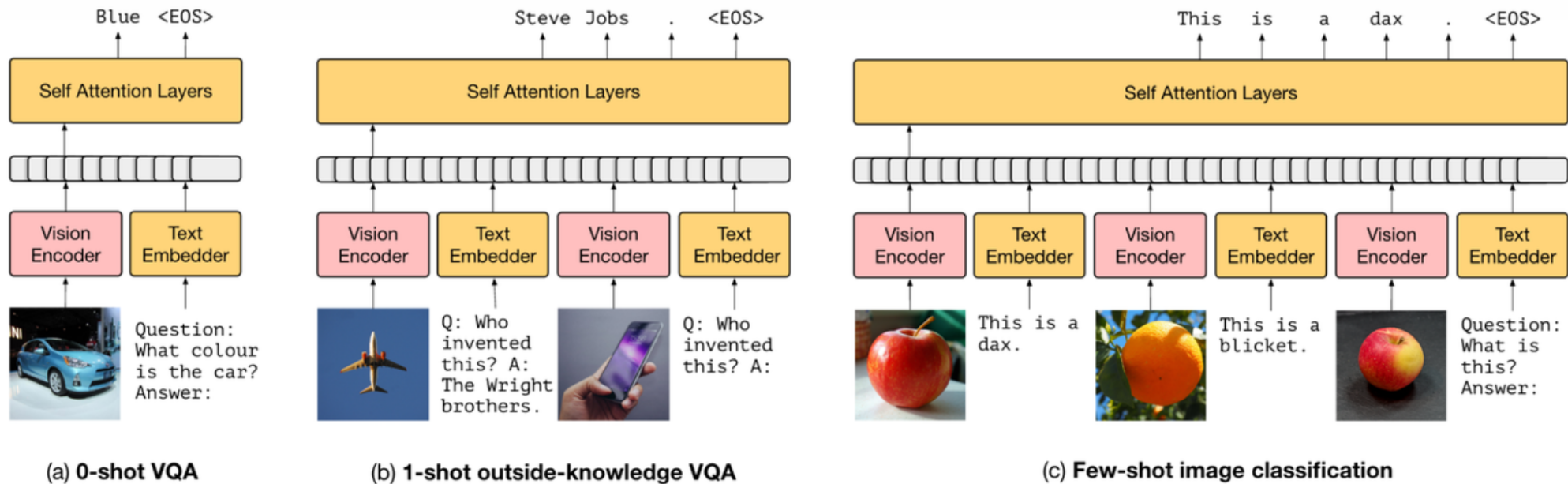


Figure 3: Inference-Time interface for *Frozen*. The figure demonstrates how we can support (a) visual question answering, (b) outside-knowledge question answering and (c) few-shot image classification via in-context learning.

Background

Frozen



Background

Frozen

Context



This was invented by
Zacharias Janssen.



This was invented by
Henry Ford.



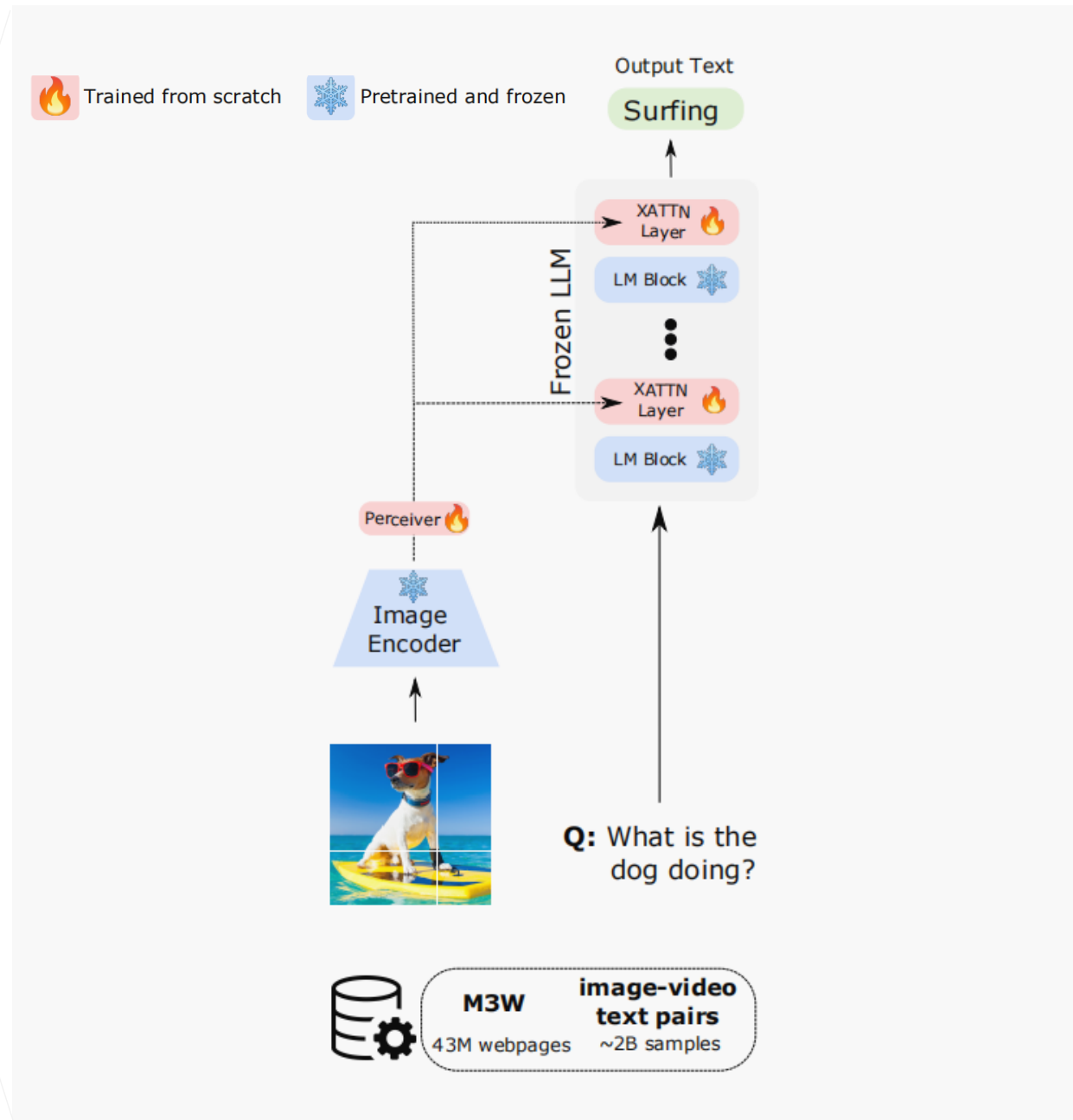
This was invented by

Model Completion

the Wright
brothers.

Background

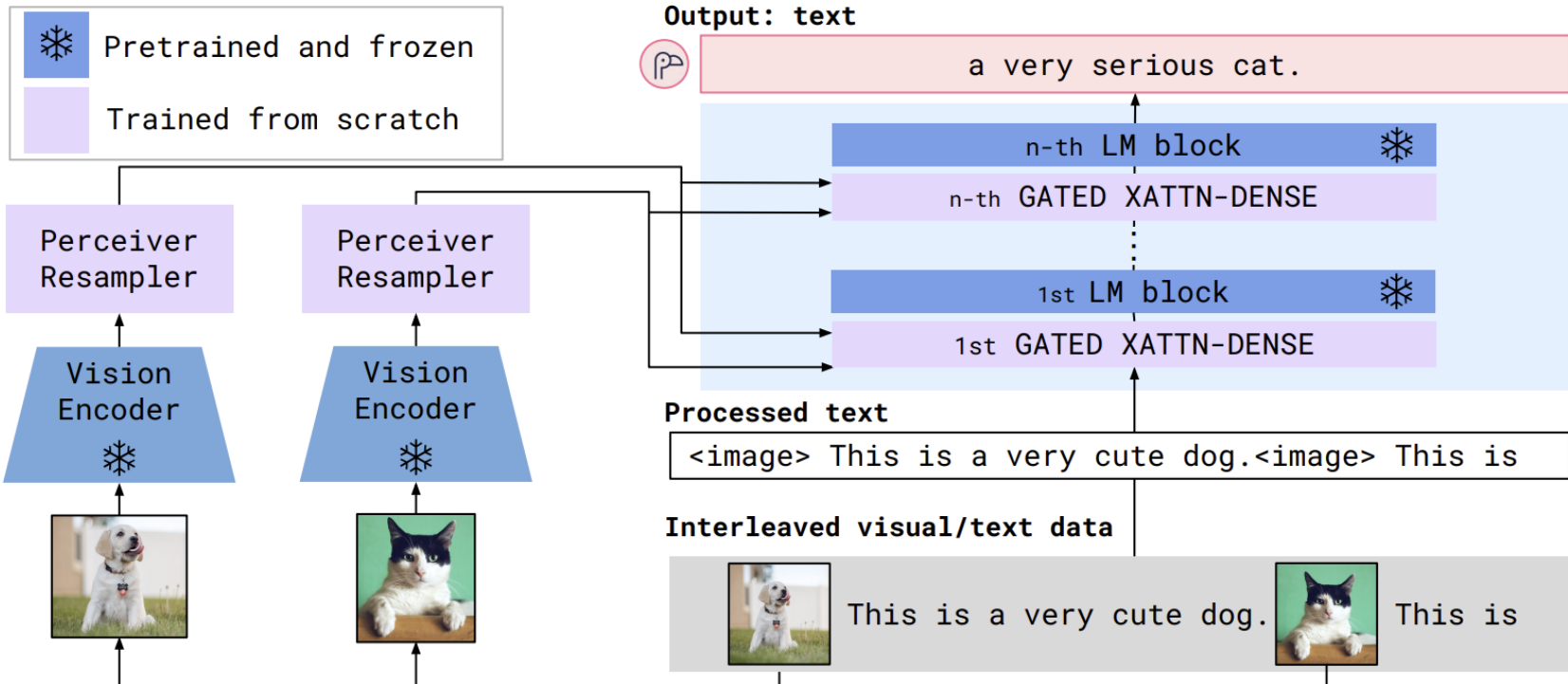
Frozen vision encoder & New layers



Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), Proceedings of Advances in Neural Information Processing Systems, 2022.

Background

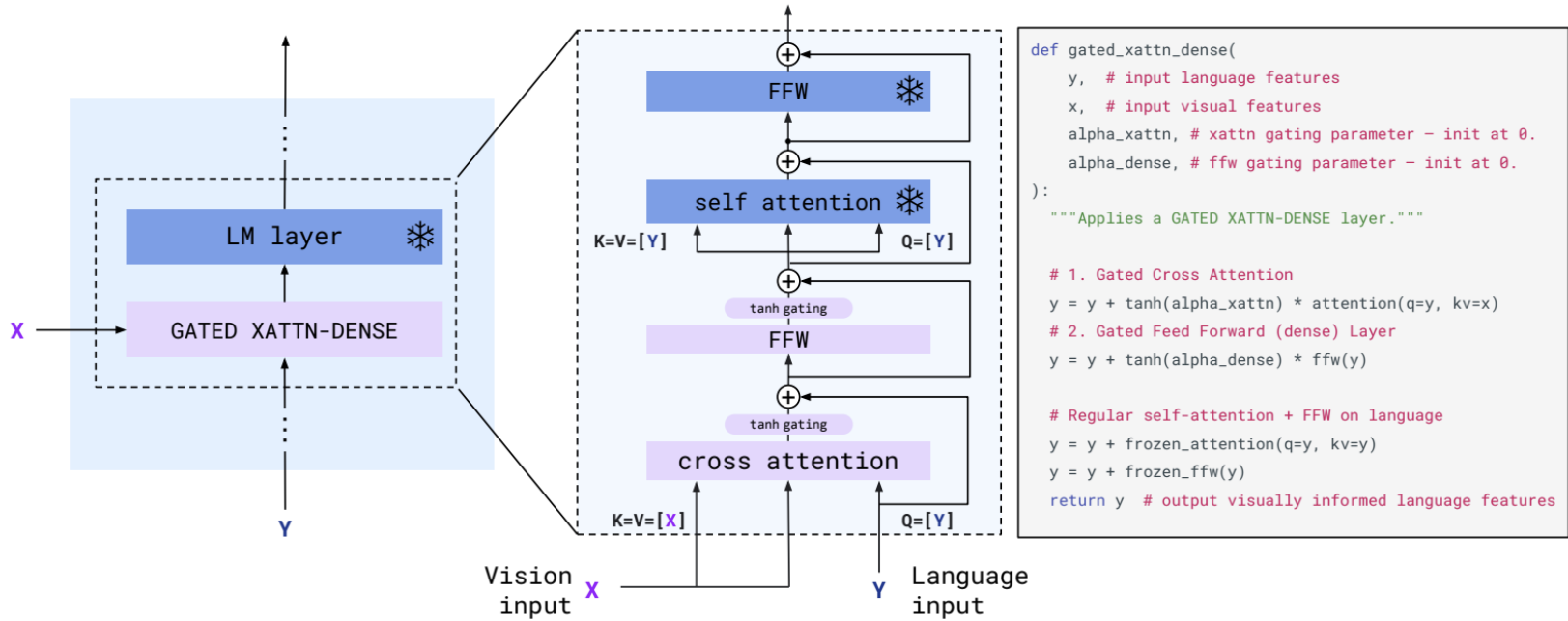
Flamingo



Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), Proceedings of Advances in Neural Information Processing Systems, 2022.

Background



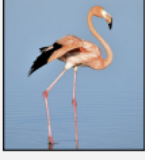
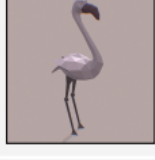



Flamingo



Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), Proceedings of Advances in Neural Information Processing Systems, 2022.

Background

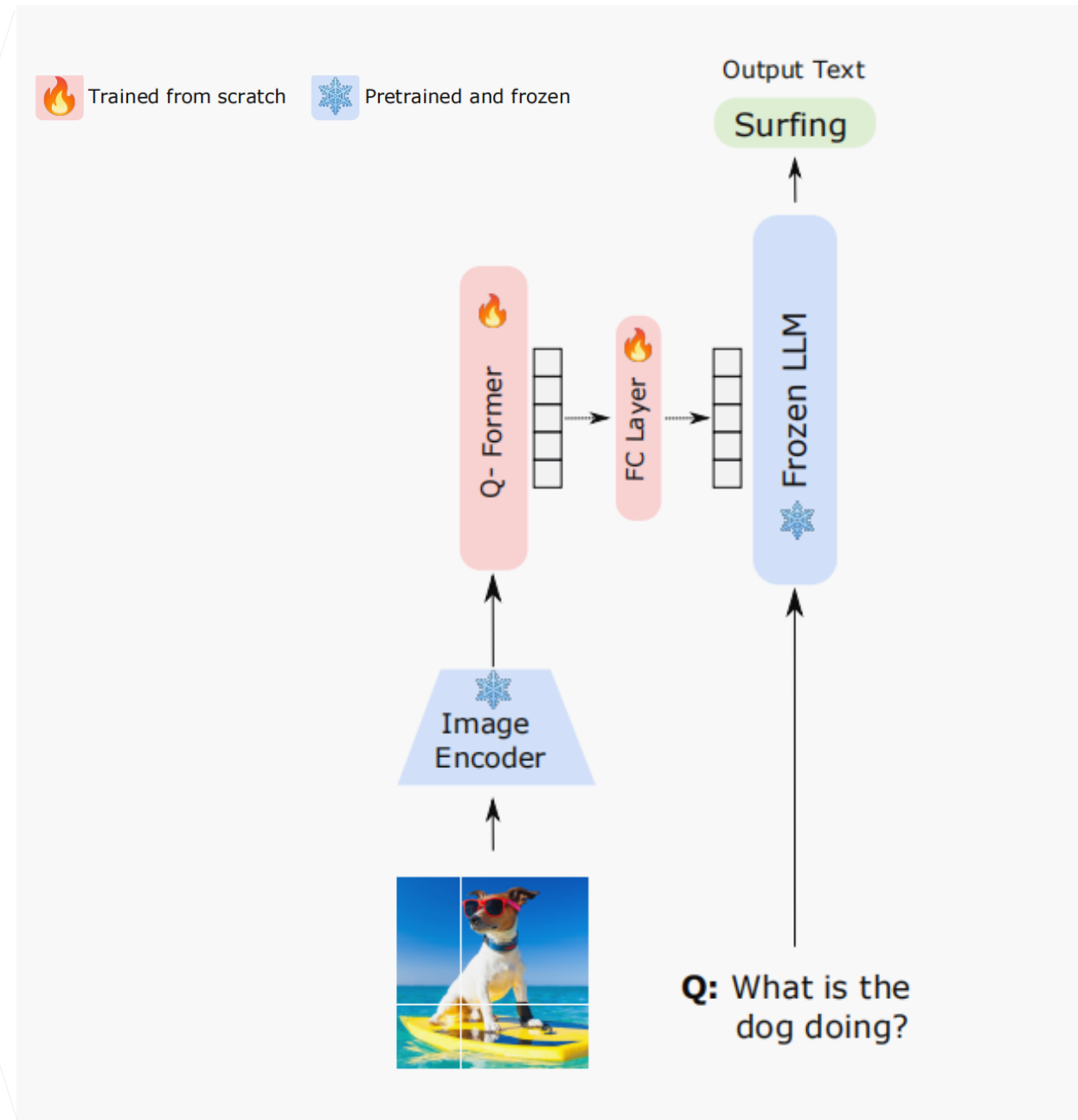
Flamingo

 <p>This is a picture of two teddy bears on the moon.</p> <p>What are they doing?</p> <p>They are having a conversation.</p> <p>What object are they using?</p> <p>It looks like a computer.</p> <p>Is this surprising?</p> <p>Yes, it is surprising.</p> <p>Why is this picture surprising to you?</p> <p>I think it is surprising because teddy bears are not usually found on the moon.</p>	   <p>What is the common thing about these three images?</p> <p>They are all flamingos.</p> <p>What is the difference between these three images?</p> <p>The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.</p>	 <p>This is an apple with a sticker on it.</p> <p>What does the sticker say?</p> <p>The sticker says "iPod".</p> <p>Where is the photo taken?</p> <p>It looks like it's taken in a backyard.</p> <p>Do you think it is printed or handwritten?</p> <p>It looks like it's handwritten.</p> <p>What color is the sticker?</p> <p>It's white.</p>	 <p>This is a cityscape. It looks like Chicago.</p> <p>What makes you think this is Chicago?</p> <p>I think it's Chicago because of the Shedd Aquarium in the background.</p>  <p>What about this one? Which city is this and what famous landmark helped you recognise the city?</p> <p>This is Tokyo. I think it's Tokyo because of the Tokyo Tower.</p>
---	---	---	---

Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), Proceedings of Advances in Neural Information Processing Systems, 2022.

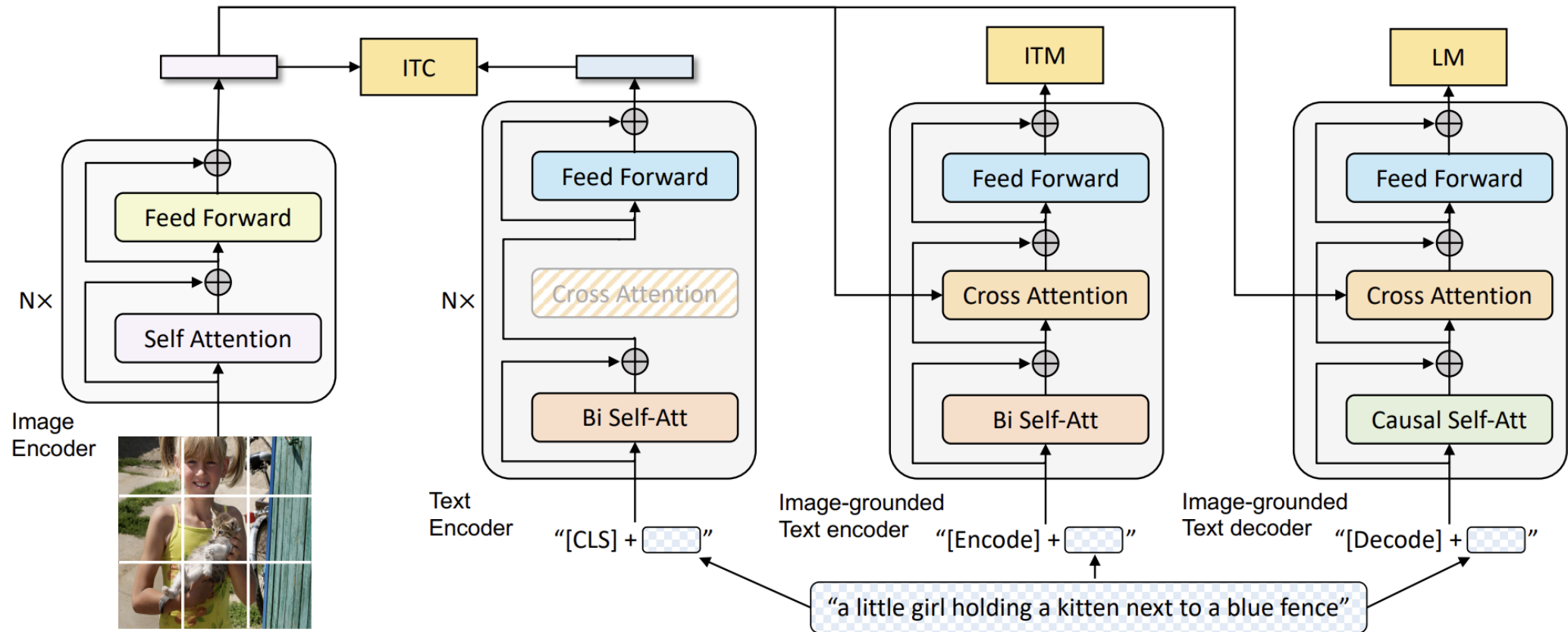
Background

Lightweight transformer



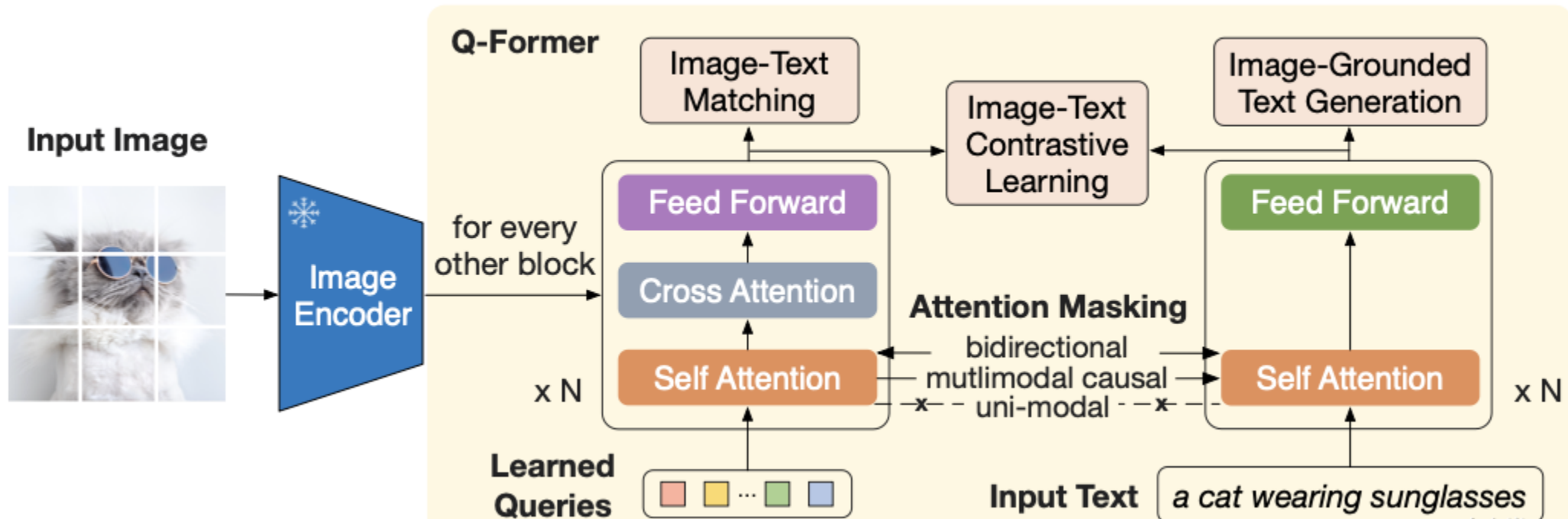
Background

BLIP



Background

BLIP-2



Content

- Background
- **Method**
- Experiments

Motivation

Massive additional vision-and-language alignment or multimodal data, Long term training

Motivation

Massive additional vision-and-language alignment or multimodal data, Long term training

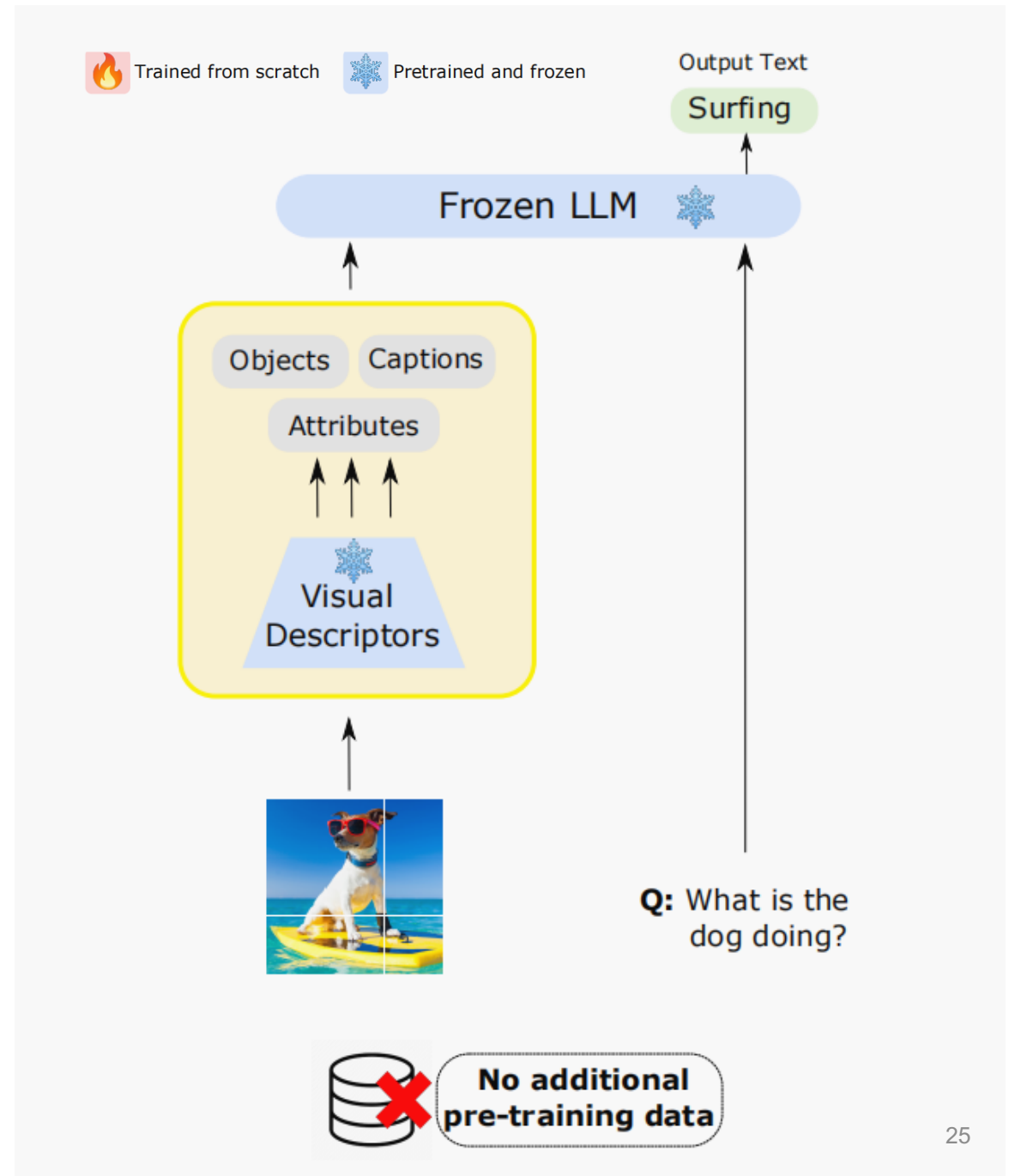


LENS: Without any need for additional vision-and-language alignment or multimodal data.

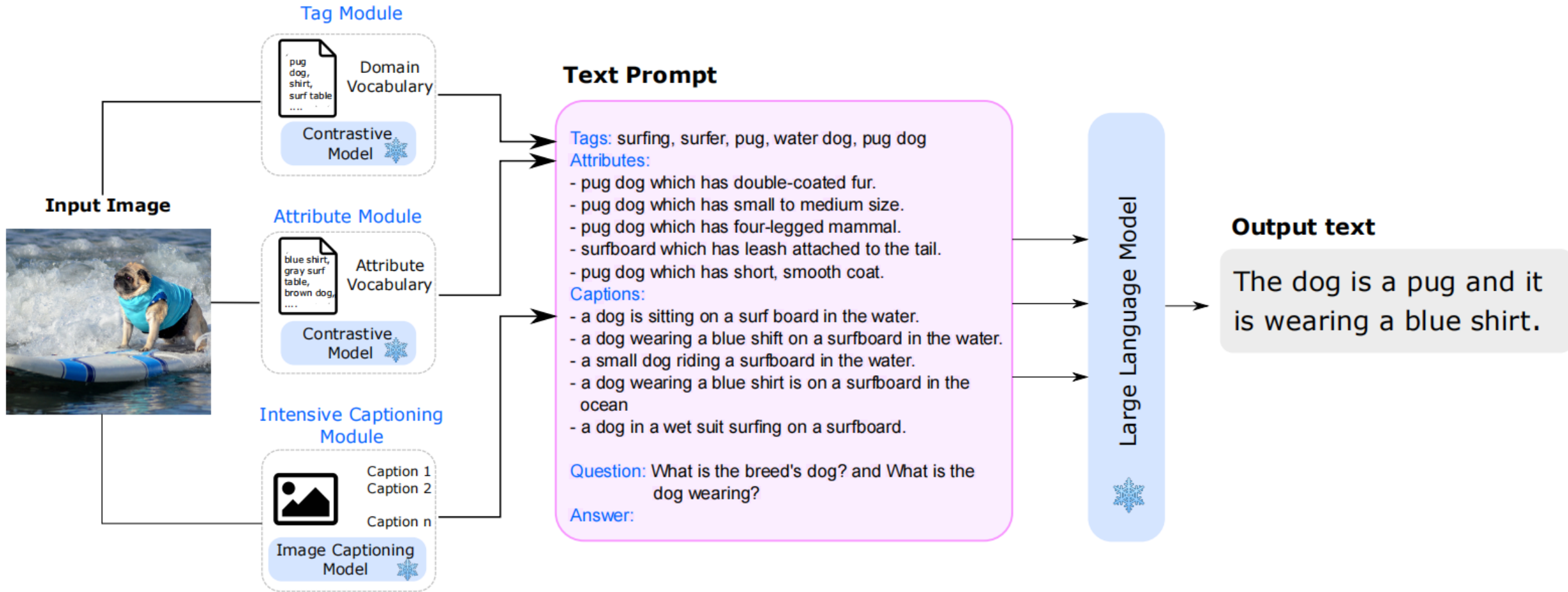
Method

Leverages a LLM as the reasoning module

Operates over independent vision modules



Framework



Content

- Background
- Method
- **Experiments**

Experiments

1. Object recognition task
 - 9 benchmark datasets
 - zero-shot, 1-shot, and 3-shot
2. Vision and language reasoning
 - zero-shot benchmarks

Prompt Design

Examples:

Components	Prompt
Tag:	Top-1 CLIP Tag
Attributes:	Top-K Attributes
Question:	Task specific prompt
Short Answer:	{answer}

Table 8: Object recognition prompt used in LENS

Components	Prompt
Captions:	Top-N captions
Question:	e.g Who is doing "x" action?
Short Answer:	{answer}

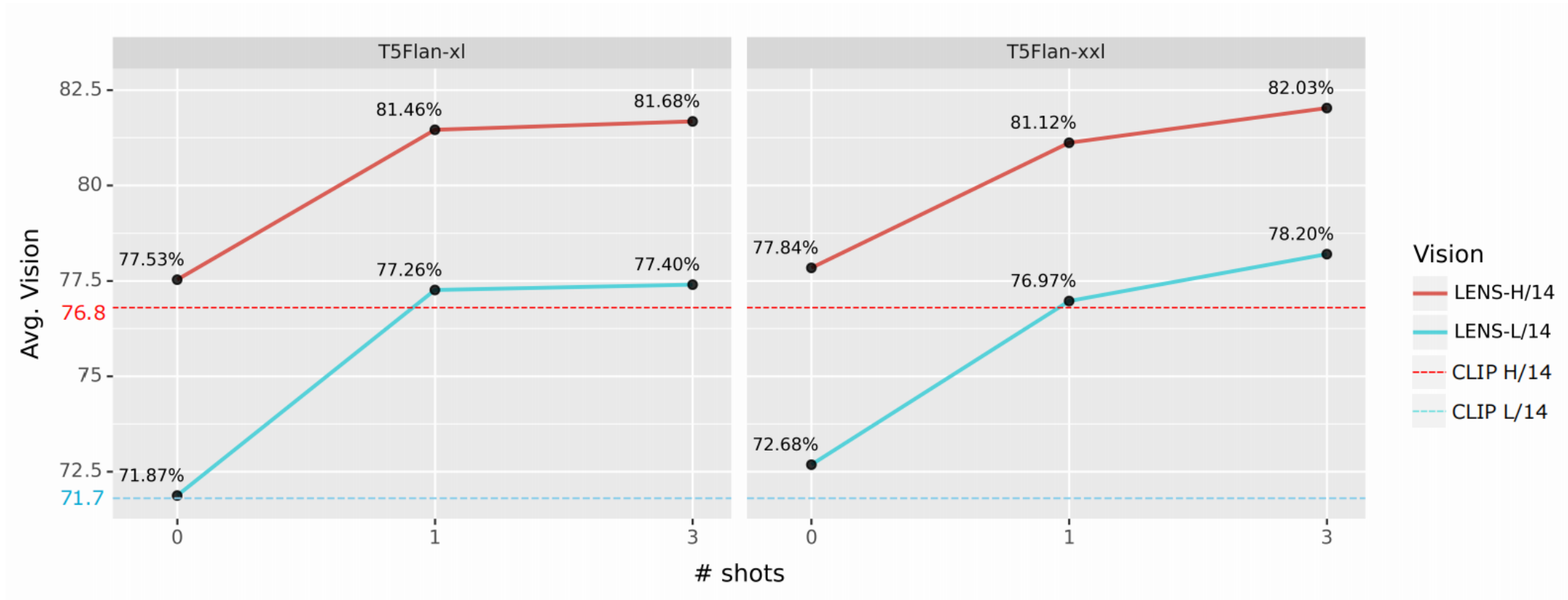
Table 9: VQA prompt used in LENS

Experiments

- Object Recognition

Datasets	LENS				CLIP	
	L ₁₄ - FlanT5 _{XL}	L ₁₄ - FlanT5 _{XXL}	H ₁₄ - FlanT5 _{XL}	H ₁₄ - FlanT5 _{XXL}	L ₁₄	H ₁₄
Pets [46]	90.1	92.0	92.6	92.4	87.8	90.1
DTD [8]	47.6	49.0	57.8	58.5	50.7	53.7
Aircraft [38]	31.1	30.1	38.5	38.5	29.5	38.0
Caltech101 [33]	71.3	71.9	75.4	75.5	70.4	75.6
Flowers102 [41]	73.0	76.4	76.6	76.7	75.5	74.9
Food101 [4]	90.9	90.9	90.8	92.1	89.8	92.6
Cars [28]	75.9	76.3	92.9	93.6	75.9	93.4
Cifar10 [30]	95.0	94.9	95.7	95.5	95.0	95.6
ImageNet-1k [9]	69.6	69.2	73.0	73.1	70.7	75.6
Vision Avg.	71.6 (-0.1)	72.3 (+0.6)	77.0 (+0.4)	77.3 (+0.7)	71.7	76.6

Experiments



Experiments

Models	# Trainable Params	VQAv2 test-dev	OK-VQA test	Rendered - SST2 test	Hateful Memes dev	test-seen
Kosmos-1	1.6B	51.0	-	67.1	63.9	-
Flamingo _{3B}	1.4B	49.2	41.2	-	-	53.7
Flamingo _{9B}	1.8B	51.8	44.7	-	-	57.0
Flamingo _{80B}	10.2B	56.3	50.6	-	-	46.4
BLIP-2 _{ViT-L FlanT5_{XL}}	103M	62.3	39.4	-	-	-
BLIP-2 _{ViT-g FlanT5_{XXL}}	108M	65.0	45.9	-	-	-
LENS Flan-T5 _{XL}	0	57.9	32.8	83.3	58.0	59.3
LENS Flan-T5 _{XXL}	0	62.6	43.3	82.0	59.4	62.5

Experiments



Q: How many white objects in this picture ?

A: 9



Q: What color is the chair in front of the wall on the left side of the stacked chairs ?

A: blue



Q: What is the largest white object on the left side of the picture ?

A: printer

Experiments



Experiments



Tell me something about the history of this place.

The Great Wall of China is a fortification built by the ancient Chinese to keep out invaders.

something else?

The Great Wall of China is a UNESCO World Heritage Site.



Write down the facts that you know about this flower.

Phalaenopsis sanderiana is a species of orchid. It is also known as the moth orchid, butterfly orchid, or moth plant. It is a perennial plant with a waxy or velvety texture. It may have a single large bloom or several smaller blooms.

Where does it come from?

Phalaenopsis sanderiana is native to the Philippines.



Is this photo unusual?

Yes

Why?

The house is upside down.

How could you get out of the house?

The house has a tunnel coming out of it.



What is this?

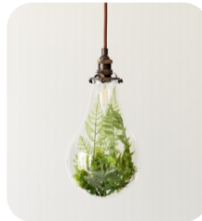
An incandescent light bulb surrounded by green leaves

Are the leaves inside of the lightbulb?

No, the leaves are outside the lightbulb

Is it glowing?

Yes, the lightbulb is glowing



What is this?

A light fixture that has a plant inside of it.

Are the plants inside of the lightbulb?

Yes.

Is it glowing?

No.

Red

What is the color of this text?

The text color is purple.

What does it say?

The text says "red".

Green

What about this one?

The text color is red.

What does it say?

The text says "green"

Experiments

Prompt Template	Acc. (Avg.)
Objects	76.6
Attributes	74.7
Objects + Attributes	77.0

Prompt Template	VQA-ACC
Question	37.2
Intensive Captioning (1) + Question	52.5
Intensive Captioning (5) + Question	56.6
Intensive Captioning (20) + Question	59.1
Intensive Captioning (50) + Question	60.4

Thanks For Listening