

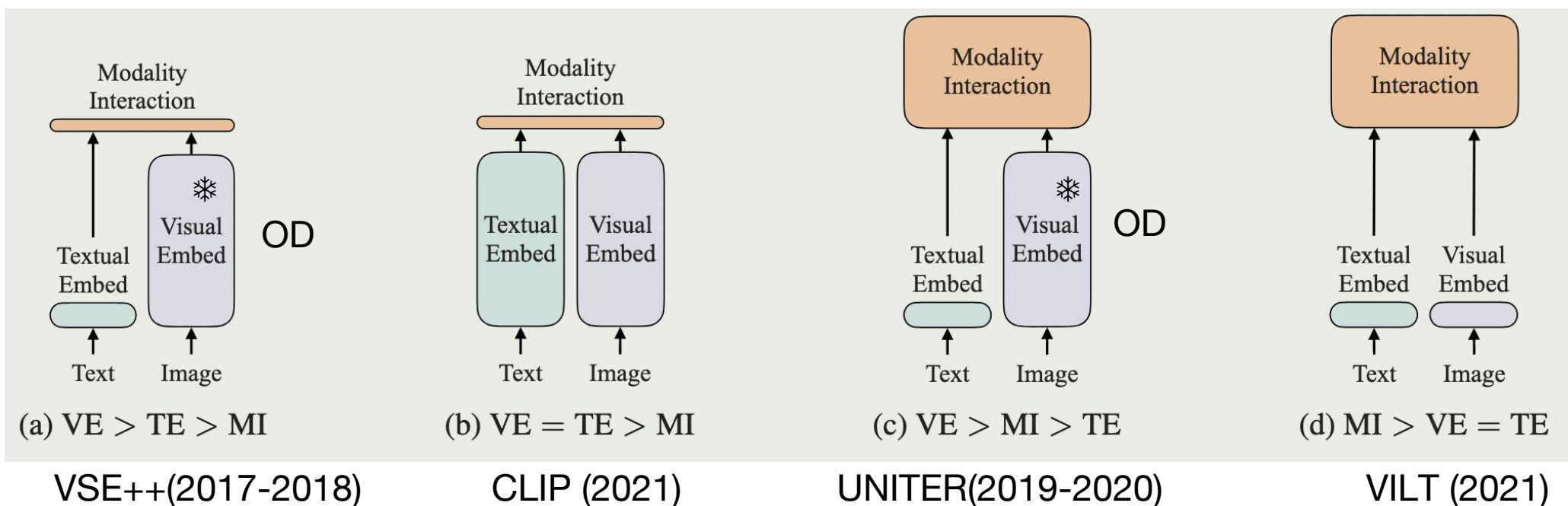
# 01 BLIPv2

## BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Junnan Li Dongxu Li Silvio Savarese Steven Hoi  
Salesforce Research

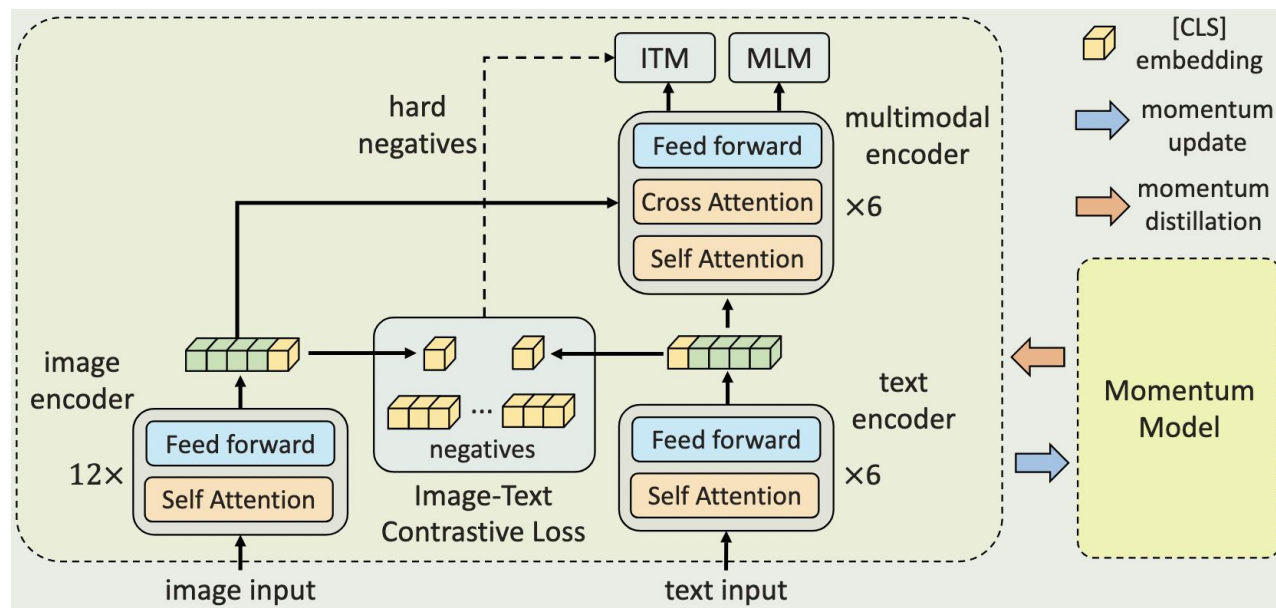
<https://github.com/salesforce/LAVIS/tree/main/projects/blip2>

BLIP trilogy



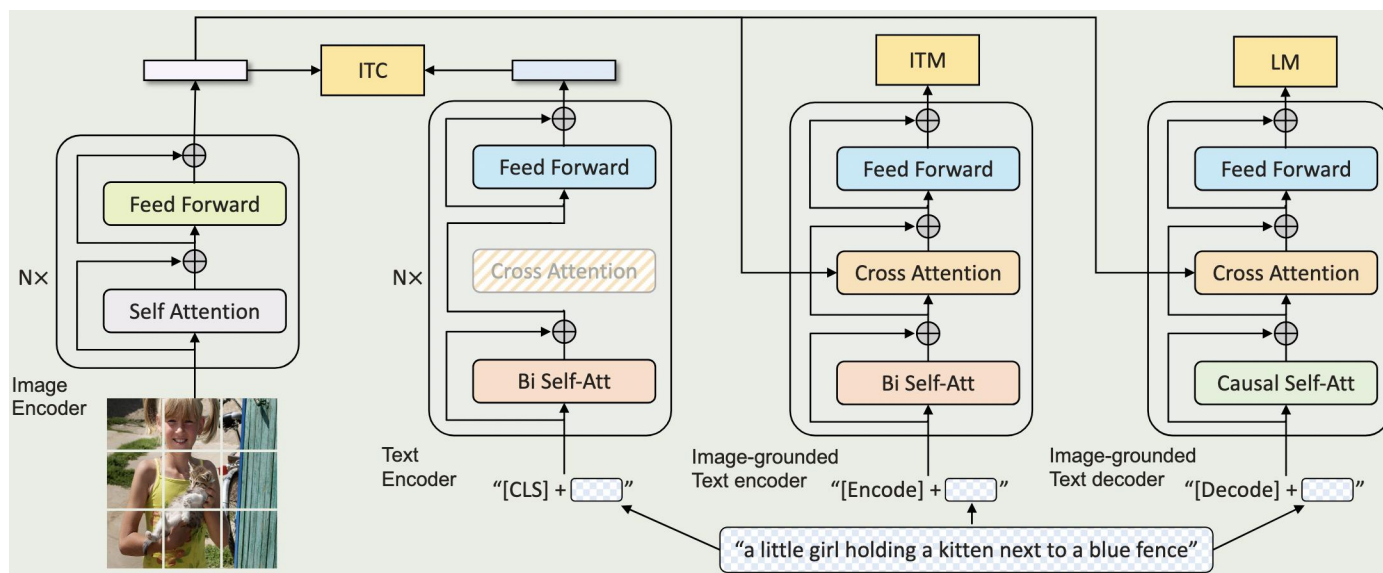
1. OD: cost/limitation
2.  $VE > TE$
3. MI

# 01 ALBEF->BLIP



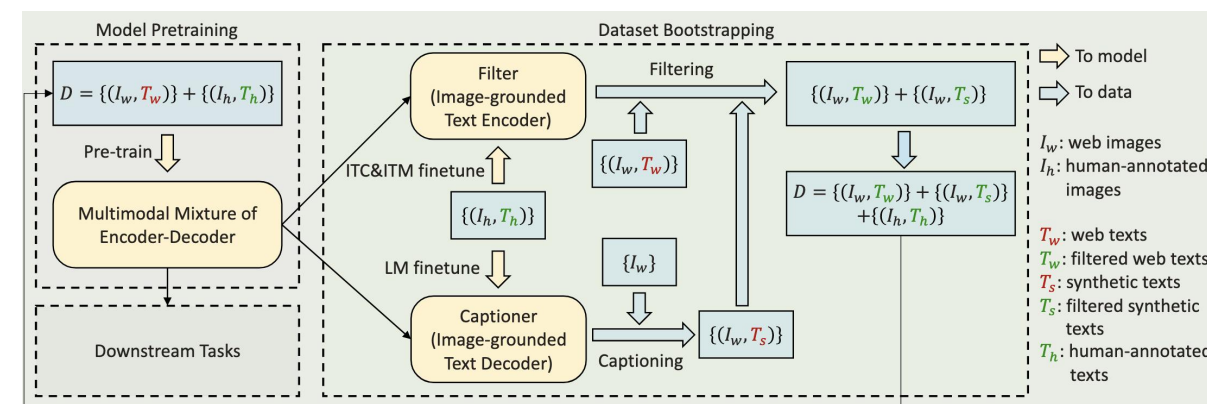
ALign BEfore Fuse(ALBEF, 2021)

1. ViT -> end-to-end, NO OD
2. contrastive loss -> efficient/general downstream
3. VE>TE(=MI)

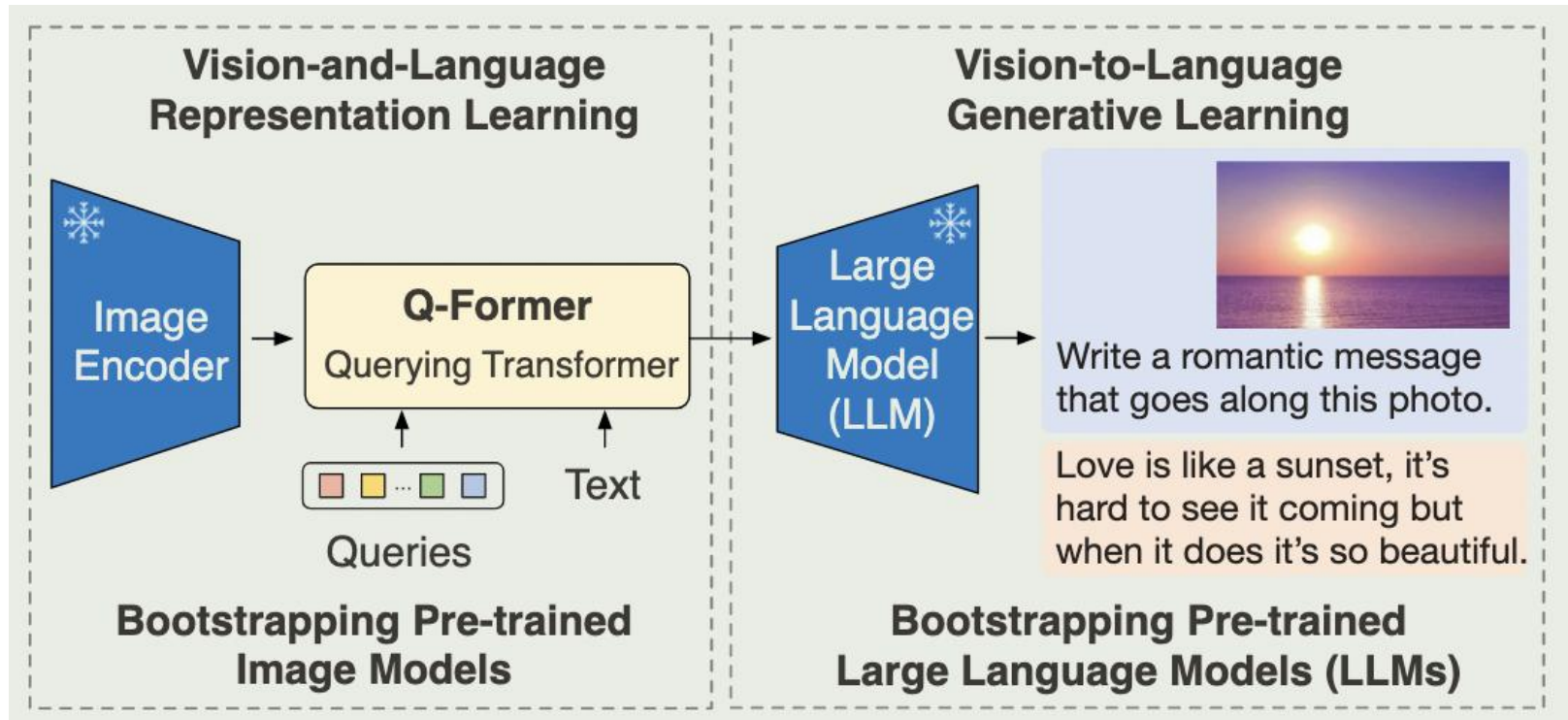


Bootstrapping Language-Image Pre-training (BLIP, 2022)

- Decoder: Generation/Retrieval Task
- Web Data: massive but noisy



## Two-stage



Lightweight Querying Transformer

Increasing Pretraining Cost -> unaffordable  
off-the-shelf frozen encoders -> efficient

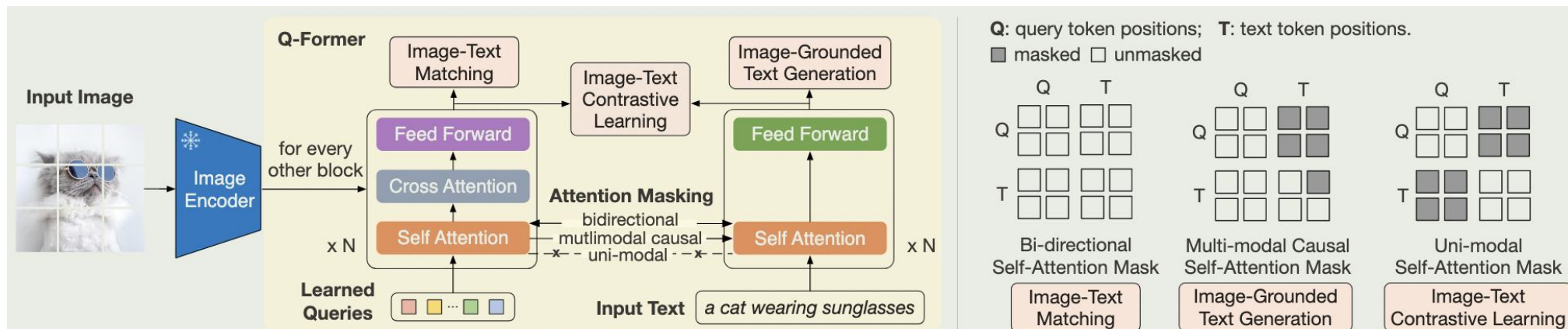
than Flamingo80B: 8.7%  $\uparrow$  VQAva2,  
54x fewer parameter

- reduce computation cost
- counteract catastrophic forgetting



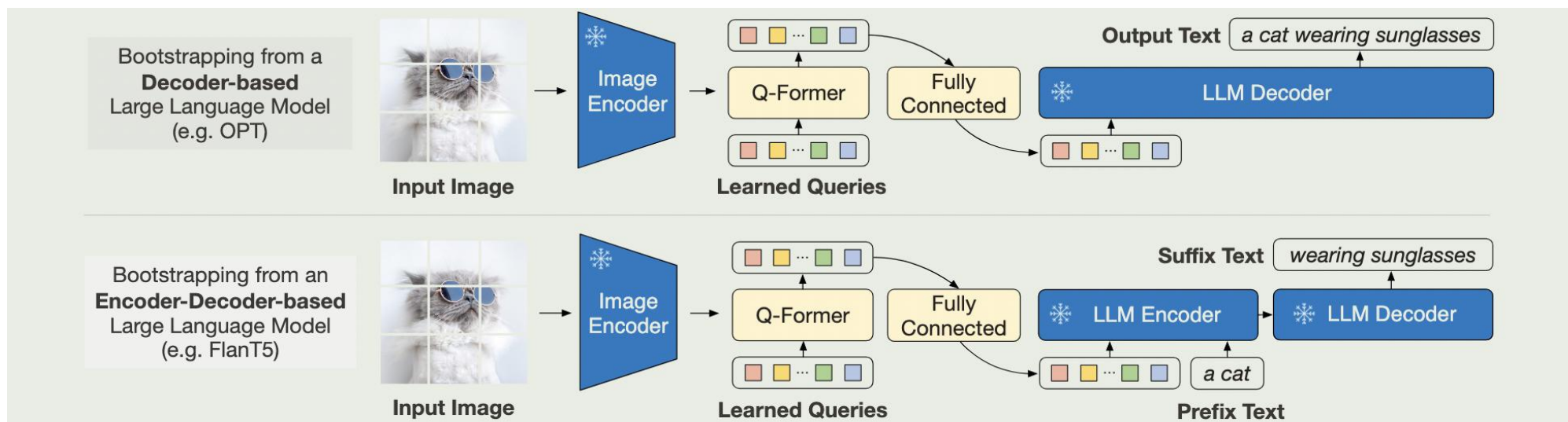
## 02 BLIPv2

stage1:



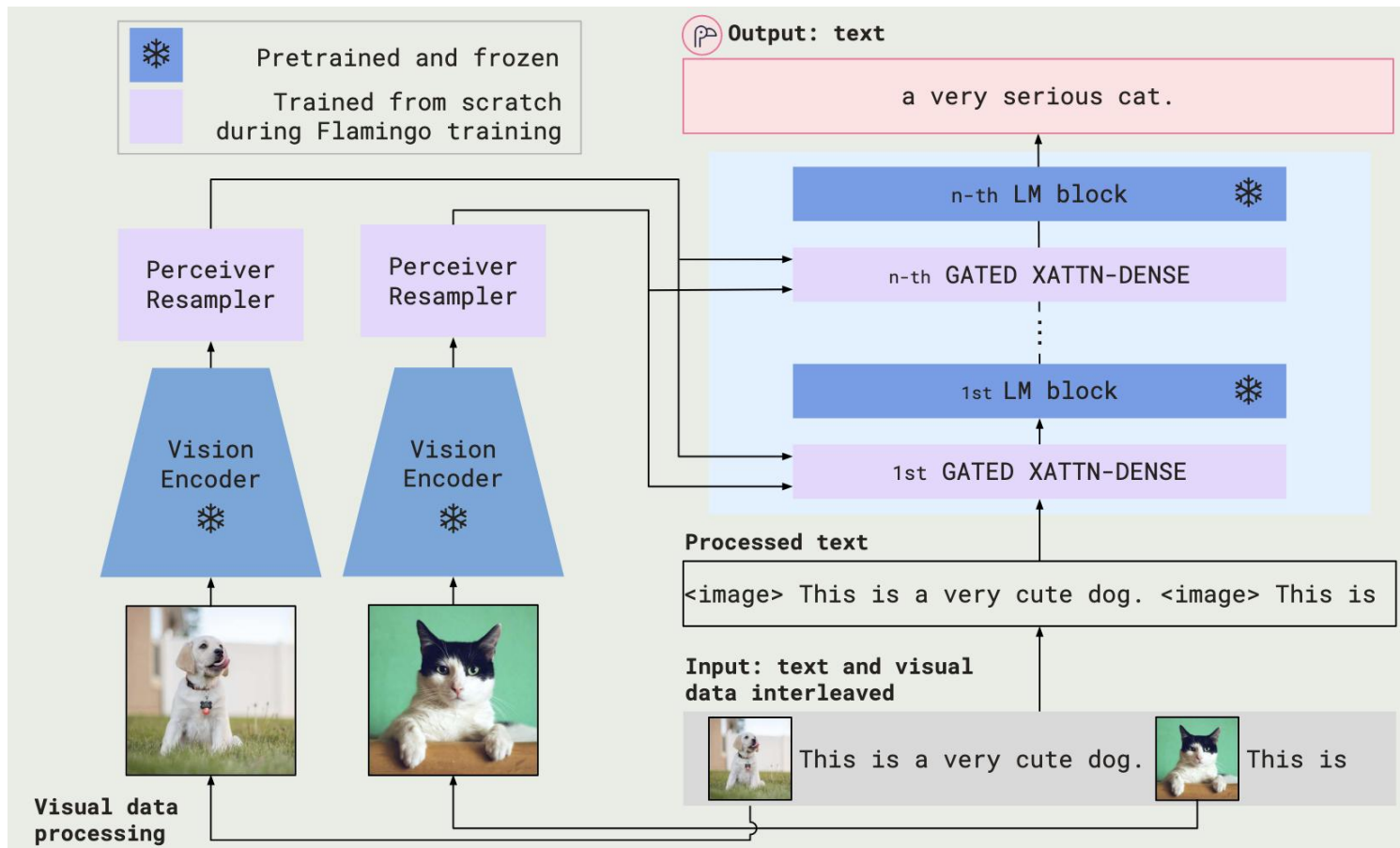
**Figure 2. (Left)** Model architecture of Q-Former and BLIP-2’s first-stage vision-language representation learning objectives. We jointly optimize three objectives which enforce the queries (a set of learnable embeddings) to extract visual representation most relevant to the text. **(Right)** The self-attention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

## information bottleneck

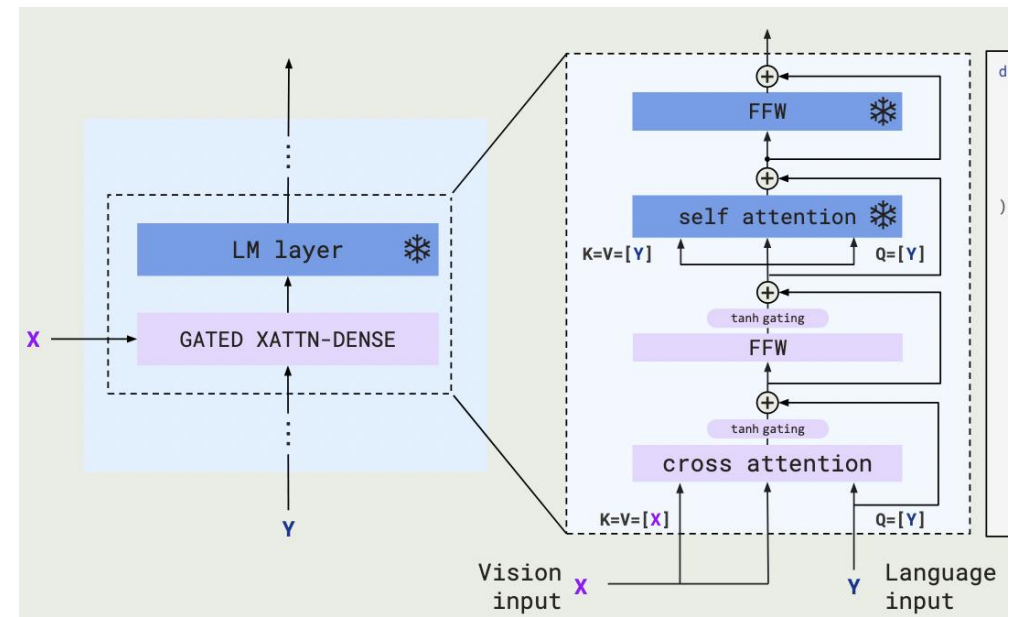
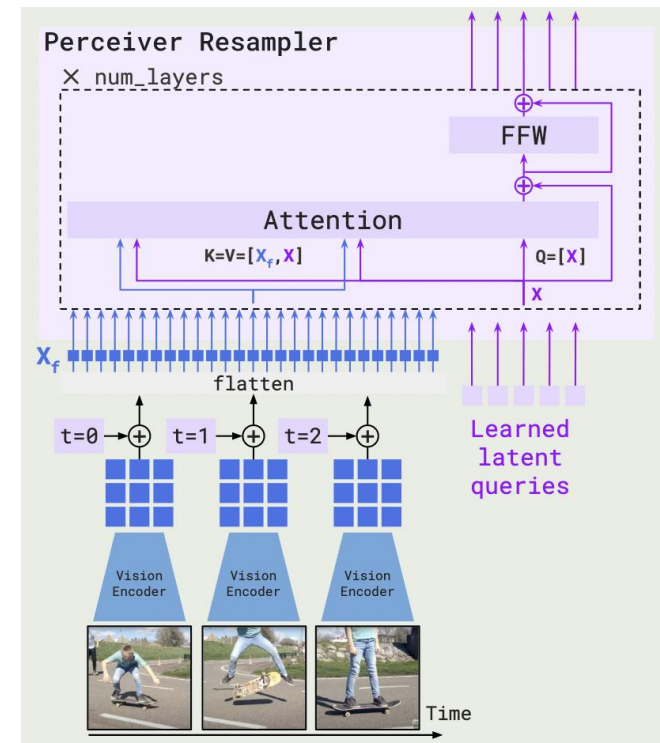


**Figure 3.** BLIP-2’s second-stage vision-to-language generative pre-training, which bootstraps from frozen large language models (LLMs). **(Top)** Bootstrapping a decoder-based LLM (e.g. OPT). **(Bottom)** Bootstrapping an encoder-decoder-based LLM (e.g. FlanT5). The fully-connected layer adapts from the output dimension of the Q-Former to the input dimension of the chosen LLM.

## 02 Flamingo DeepMind 2022



**Figure 3 | Overview of the Flamingo model.** The Flamingo models are a family of visual language model (VLM) that can take as input visual data interleaved with text and can produce free-form text as output. Key to its performance are novel architectural components and pretraining strategies described in Section 3.





## VQA

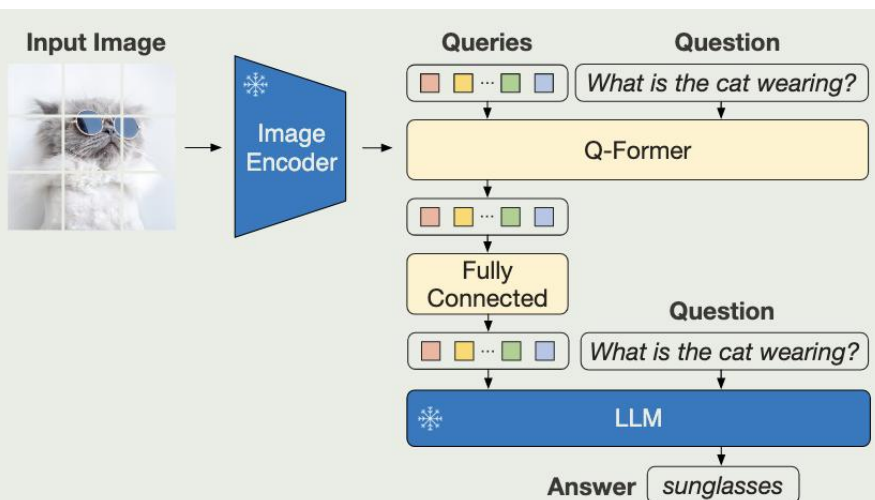


Figure 7. Model architecture for VQA finetuning, where the LLM receives Q-Former’s output and the question as input, then predicts answers. We also provide the question as a condition to Q-Former, such that the extracted image features are more relevant to the question.

Models	#Trainable Params	Open-sourced?	Visual Question Answering	Image Captioning		Image-Text Retrieval	
			VQAv2 (test-dev) VQA acc.	NoCaps (val) CIDEr	SPICE	Flickr (test) TR@1	IR@1
BLIP (Li et al., 2022)	583M	✓	-	113.2	14.8	96.7	86.7
SimVLM (Wang et al., 2021b)	1.4B	✗	-	112.2	-	-	-
BEIT-3 (Wang et al., 2022b)	1.9B	✗	-	-	-	94.9	81.5
Flamingo (Alayrac et al., 2022)	10.2B	✗	56.3	-	-	-	-
BLIP-2	188M	✓	<b>65.0</b>	<b>121.6</b>	<b>15.8</b>	<b>97.6</b>	<b>89.7</b>

Table 1. Overview of BLIP-2 results on various **zero-shot** vision-language tasks. Compared with previous state-of-the-art models. BLIP-2 achieves the highest zero-shot performance while requiring the least number of trainable parameters during vision-language pre-training.

Models	#Trainable Params	#Total Params	VQAv2		OK-VQA	GQA
			val	test-dev	test	test-dev
VL-T5 <sub>no-vqa</sub>	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	<b>50.6</b>	-
BLIP-2 ViT-L OPT <sub>2.7B</sub>	104M	3.1B	50.1	49.7	30.2	33.9
BLIP-2 ViT-G OPT <sub>2.7B</sub>	107M	3.8B	53.5	52.3	31.7	34.6
BLIP-2 ViT-G OPT <sub>6.7B</sub>	108M	7.8B	54.3	52.6	36.4	36.4
BLIP-2 ViT-L FlanT5 <sub>XL</sub>	103M	3.4B	62.6	62.3	39.4	<u>44.4</u>
BLIP-2 ViT-G FlanT5 <sub>XL</sub>	107M	4.1B	<u>63.1</u>	<u>63.0</u>	40.7	44.2
BLIP-2 ViT-G FlanT5 <sub>XXL</sub>	108M	12.1B	<b>65.2</b>	<b>65.0</b>	<u>45.9</u>	<b>44.7</b>

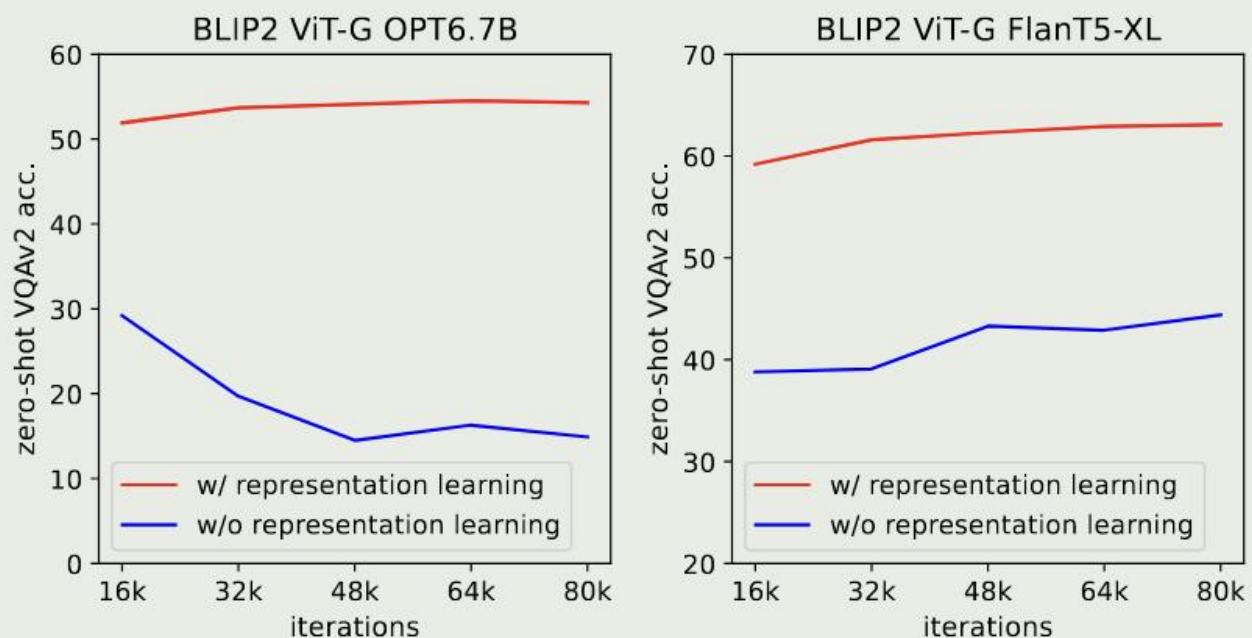
Table 2. Comparison with state-of-the-art methods on zero-shot visual question answering.

Retrieval:

Model	#Trainable Params	Flickr30K Zero-shot (1K test set)						COCO Fine-tuned (5K test set)					
		Image → Text			Text → Image			Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
<i>Dual-encoder models</i>													
CLIP (Radford et al., 2021)	428M	88.0	98.7	99.4	68.7	90.6	95.2	-	-	-	-	-	-
ALIGN (Jia et al., 2021)	820M	88.6	98.7	99.7	75.7	93.8	96.8	77.0	93.5	96.9	59.9	83.3	89.8
FILIP (Yao et al., 2022)	417M	89.8	99.2	99.8	75.0	93.4	96.3	78.9	94.4	97.4	61.2	84.3	90.6
Florence (Yuan et al., 2021)	893M	90.9	99.1	-	76.7	93.6	-	81.8	95.2	-	63.2	85.7	-
BEIT-3(Wang et al., 2022b)	1.9B	94.9	99.9	<b>100.0</b>	81.5	95.6	97.8	<u>84.8</u>	<u>96.5</u>	<u>98.3</u>	<u>67.2</u>	<b>87.7</b>	<b>92.8</b>
<i>Fusion-encoder models</i>													
UNITER (Chen et al., 2020)	303M	83.6	95.7	97.7	68.7	89.2	93.9	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
VinVL (Zhang et al., 2021)	345M	-	-	-	-	-	-	75.4	92.9	96.2	58.8	83.5	90.3
<i>Dual encoder + Fusion encoder reranking</i>													
ALBEF (Li et al., 2021)	233M	94.1	99.5	99.7	82.8	96.3	98.1	77.6	94.3	97.2	60.7	84.3	90.5
BLIP (Li et al., 2022)	446M	96.7	<b>100.0</b>	<b>100.0</b>	86.7	97.3	98.7	82.4	95.4	97.9	65.1	86.3	91.8
<b>BLIP-2</b> ViT-L	474M	<u>96.9</u>	<b>100.0</b>	<b>100.0</b>	<u>88.6</u>	<u>97.6</u>	<b>98.9</b>	83.5	96.0	98.0	66.3	86.5	91.8
<b>BLIP-2</b> ViT-G	1.2B	<b>97.6</b>	<b>100.0</b>	<b>100.0</b>	<b>89.7</b>	<b>98.1</b>	<b>98.9</b>	<b>85.4</b>	<b>97.0</b>	<b>98.5</b>	<b>68.3</b>	<b>87.7</b>	<u>92.6</u>

Table 5. Comparison with state-of-the-art image-text retrieval methods, finetuned on COCO and zero-shot transferred to Flickr30K.





*Figure 5.* Effect of vision-language representation learning on vision-to-language generative learning. Without representation learning, the Q-Former fails the bridge the modality gap, leading to significantly lower performance on zero-shot VQA.

COCO finetuning objectives	Image $\rightarrow$ Text		Text $\rightarrow$ Image	
	R@1	R@5	R@1	R@5
ITC + ITM	84.5	96.2	67.2	87.1
ITC + ITM + ITG	85.4	97.0	68.3	87.7

*Table 6.* The image-grounded text generation (ITG) loss improves image-text retrieval performance by enforcing the queries to extract language-relevant visual features.



## LLM to understand image















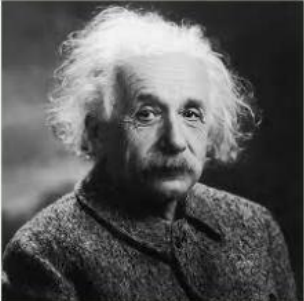
 <p>Explain the advantages of this product.</p> <p>The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.</p>	 <p>Tell me something about the history of this place.</p> <p>The great wall of china was built in 221 bc by emperor qin shihuang to protect the emperor's capital from invaders from the north.</p>	 <p>Write down the facts that you know about this flower.</p> <p>It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.</p>	 <p>Which city is this?</p> <p>Singapore.</p> <p>Why do you think so?</p> <p>The city has a statue of a merlion.</p>	 <p>What happened at the end of this movie?</p> <p>The titanic sank.</p> <p>Did Leonardo Dicaprio's character survive?</p> <p>No, he drowned.</p>	 <p>What is in the photo?</p> <p>A pizza that looks like a cat.</p> <p>What is the nose made of?</p> <p>A slice of pepperoni.</p>
 <p>Is this photo unusual?</p> <p>Yes, it's a house that looks like it's upside down.</p> <p>How could someone get out of the house?</p> <p>It has a slide on the side of the house.</p>	 <p>What are shown in the photo?</p> <p>A man and a chicken.</p> <p>What does the man feel and why?</p> <p>He is scared of the chicken because it is flying at him.</p>	 <p>What are the ingredients I need to make this?</p> <p>Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.</p> <p>What is the first step?</p> <p>Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.</p>	 <p>Write a romantic message that goes along this photo.</p> <p>Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.</p>	 <p>Tell us about the photo you took for Darren and Jade.</p> <p>Darren and Jade are a couple who have a passion for horses, so we decided to take a photo of them with a horse in the desert on their wedding day.</p>	 <p>Write a conversation between the two animals.</p> <p>cat: hey dog, can i ride on your back? dog: sure, why not? cat: i'm tired of walking in the snow.</p>

Figure 4. Selected examples of **instructed zero-shot image-to-text generation** using a BLIP-2 model w/ ViT-G and FlanT5<sub>XXL</sub>, where it shows a wide range of capabilities including visual conversation, visual knowledge reasoning, visual commonsense reasoning, storytelling, personalized image-to-text generation, etc.


zero-shot image-to-text capabilities

# unsatisfactory image-to-text generation results

inherits the risks of LLMs





Write a famous quote said by this person.




albert einstein - the world is a book, and those who do not travel read only one page.

Inaccurate knowledge  
(quote is from a different person)






Can I wear this for my trip to Canada in December?




yes, it's a nice shirt and shorts, but it's a little too casual for a trip to Canada.

Incorrect reasoning path  
(should have considered weather)



Please write the specifics of this product.



apple iphone 11 128gb space gray

Information not up-to-date  
(this is iphone 14)

offensive language, propagating social bias, or leaking private information