Big Models, Few-Shot Learning, and Model Evaluation

Zhe Gan

Principal Researcher

6/19/2022

What has been covered so far

- The general landscape of image-text pre-training
  - OD-based methods: LXMERT, ViLBERT, UNITER, VL-BERT, OSCAR, VILLA, VinVL etc.
  - The current prevailing E2E methods: PixelBERT, SOHO, ViLT, ALBEF, CLIP-ViL, METER, BLIP etc.

- Unified image-text modeling
  - Output format: VL-T5, GPV, MDETR, UniT, UNICORN, OFA etc.
  - Architecture: UFO, VLMo, VL-BEiT etc.

- A typical *academic* setting for all these models
  - Model size: base (~110M) or large (~340M)
  - Pre-training data: 4M images in total (COCO+VG+SBU+CC3M)

So, what do we offer in this talk?

- Part I: Big multimodal foundation models
  - Beyond base/large sizes, and beyond 4M images
  - Examples include SimVLM, CoCa, and GIT
- Part II: Multimodal few-shot learning
  - How can we enable in-context few-shot learning?
  - Examples include Frozen, PICa, and Flamingo
- Part III: Model evaluation
  - What's next for VL model evaluation?
  - Diagnostic tests, challenge sets, probing analysis

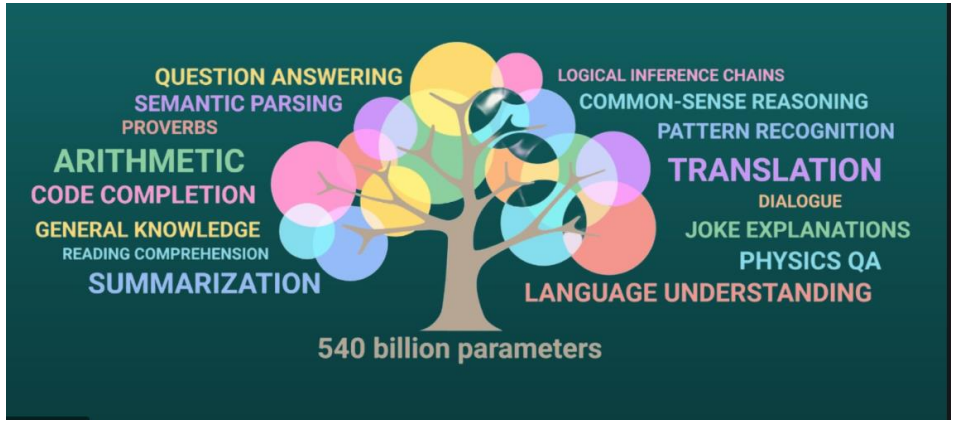
Part I: Big Models

What are big foundation models?

- Let's take a look at big language foundation models

BERT (base 110M, large 340M)

GPT-3 (175B)



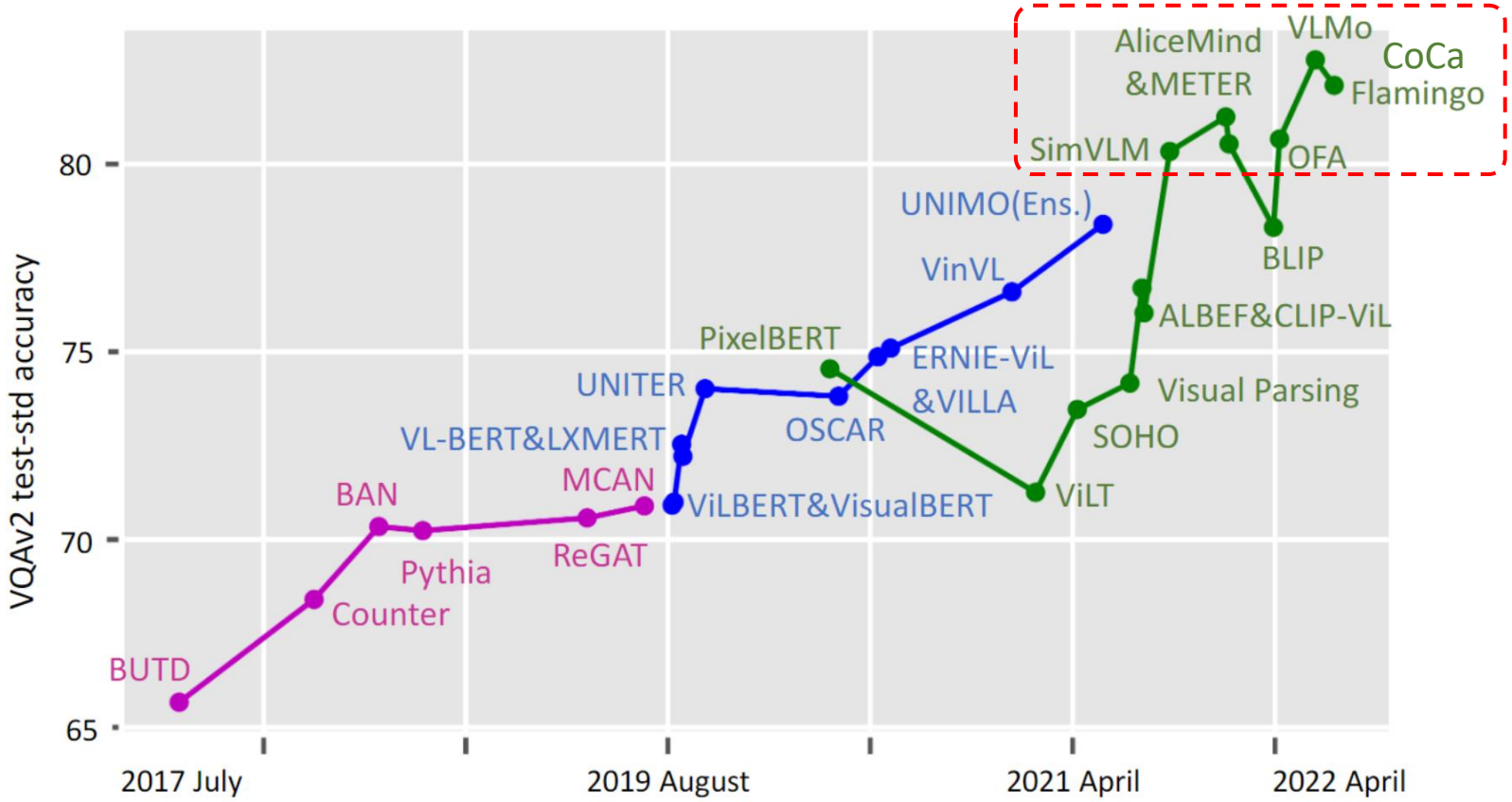
PaLM (540B)

How about big multimodal models?

- Models that have either billion-level parameters or use billion-level pre-training data are considered as *"big"* in this context
- First, note that foundation models are not necessarily needed to be big
- CLIP-like dual encoders and text-to-image big models (DALLE-2, Imagen) are not considered here (will be covered in the afternoon session)
- Take VQA as an example
  - OD-based models
  - E2E models

Large model sizes and pre-training data have been the driving force for SOTA performance.

We will also briefly talk about what's beyond SOTA chasing in later slides.



A summary of big multimodal models

Model	Model Size				#Pre-training image-text data	Pre-training tasks
	Img Enc	Txt Enc	Fusion	Total		
CLIP ViT-L/14	302M	123M	0	425M	400M	ITC
ALIGN	480M	340M	0	820M	1.8B	ITC
Florence	637M	256M	0	893M	900M	ITC
SimVLM-huge	300M	39M	600M	939M	1.8B	PrefixLM
METER-huge	637M	125M	220M	982M	20M*	MLM+ITM
LEMON	147M	39M	636M	822M	200M	MLM
Flamingo	200M	70B	10B	80.2B	2.1B+27M video-text	LM
GIT	637M	40M	70M	747M	800M	LM
VLMo++	--	--	--	565M	1B	MLM+ITM+ITC
CoCa	1B	477M	623M	2.1B	4.8B (before filtering)	ITC+LM

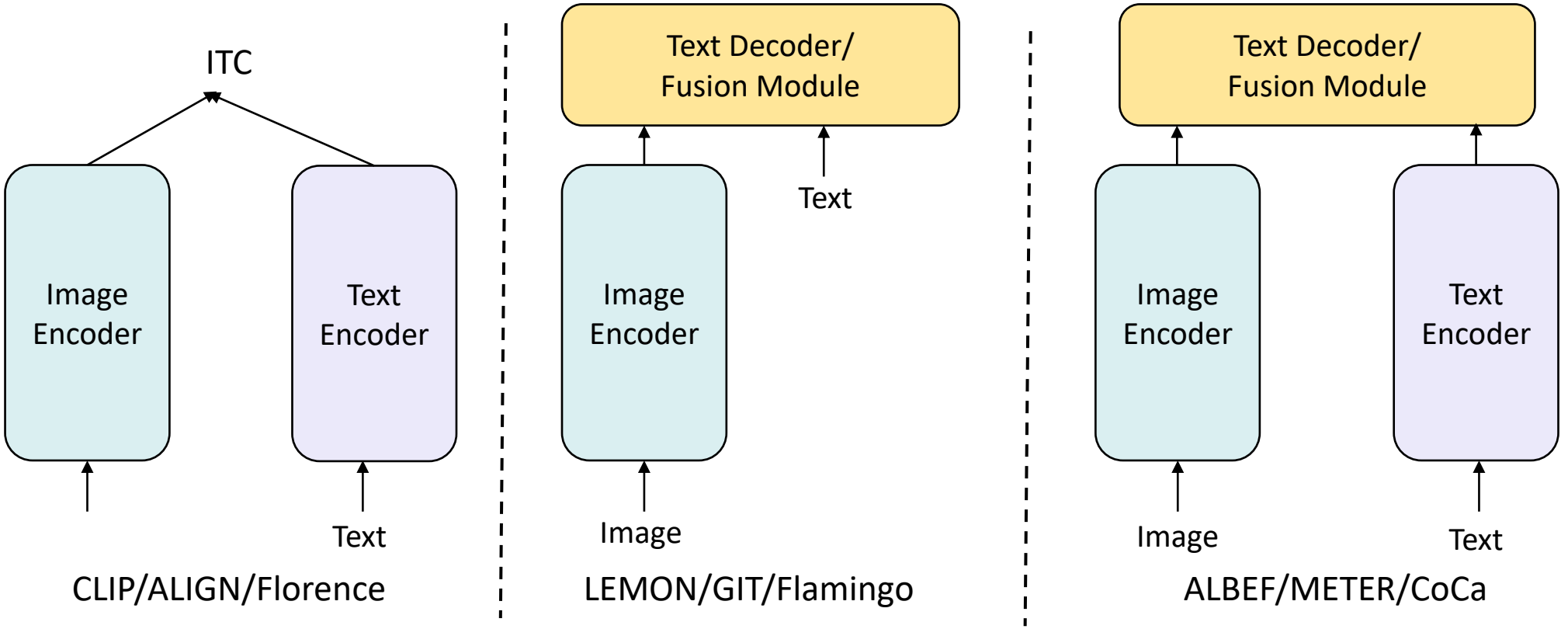
Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision
Learning Transferable Visual Models From Natural Language Supervision
Florence: A New Foundation Model for Computer Vision
SimVLM: Simple Visual Language Model Pretraining with Weak Supervision
An Empirical Study of Training End-to-End Vision-and-Language Transformers
Scaling Up Vision-Language Pre-training for Image Captioning
Flamingo: a Visual Language Model for Few-Shot Learning

Note: Some of the numbers here are based on our best estimate
**: excluding the data used to pre-train the Florence image encoder*

GIT: A Generative Image-to-text Transformer for Vision and Language
VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts
CoCa: Contrastive Captioners are Image-Text Foundation Models

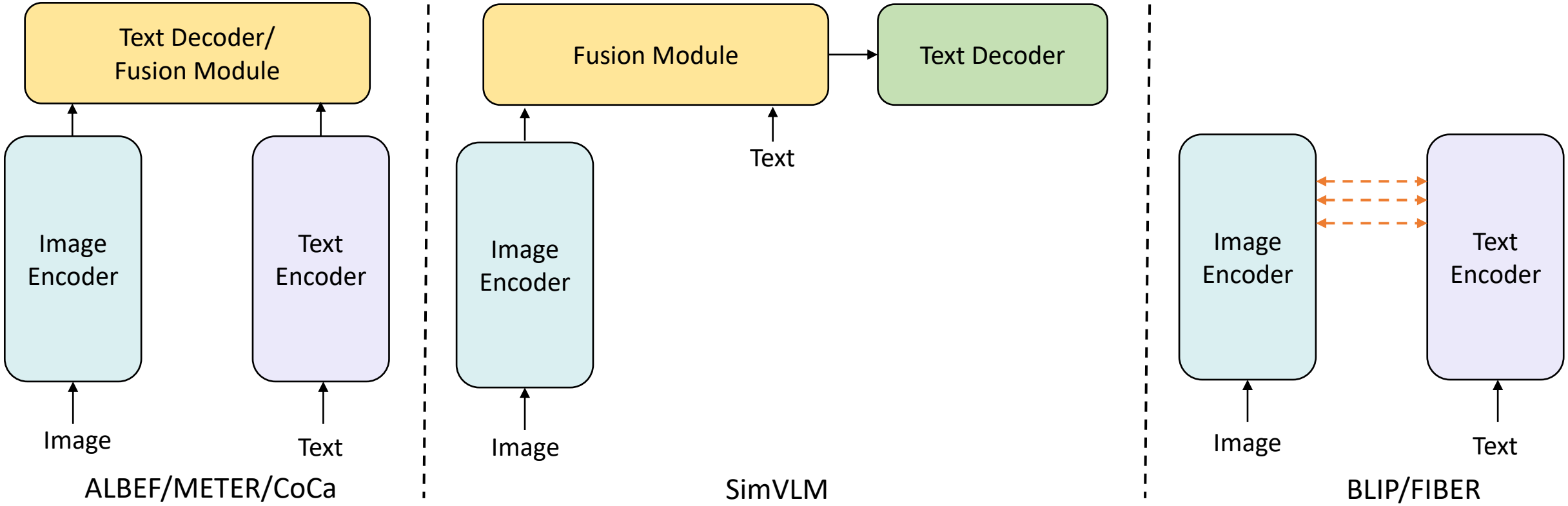
How do these models look like?

- CLIP/ALIGN/Florence models are dual-encoders that only use ITC loss for pre-training
- LEMON/GIT/Flamingo models are fusion encoders
  - LEMON uses a strong OD module to first extract image features offline
  - GIT uses a big Swin-like image encoder, but a small text decoder
  - On the other hand, Flamingo uses a relatively small image encoder, but a big text decoder
  - Since only MLM or LM losses are used for pre-training, it is not friendly to *fast* retrieval



How do these models look like?

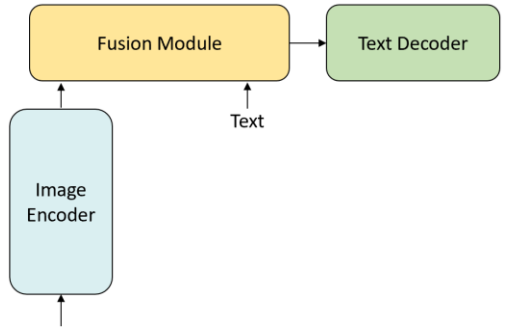
- ALBEF/METER/CoCa use an encoder-only design, but the fusion module can be used/called as a text decoder as well
  - In METER, ITC loss is not used; while in ALBEF and CoCa, ITC loss is used, which enables fast retrieval
  - There are also models like BLIP and FIBER that performs *fusion in the backbone*
- SimVLM uses an encoder-decoder design, and pre-trained with PrefixLM only



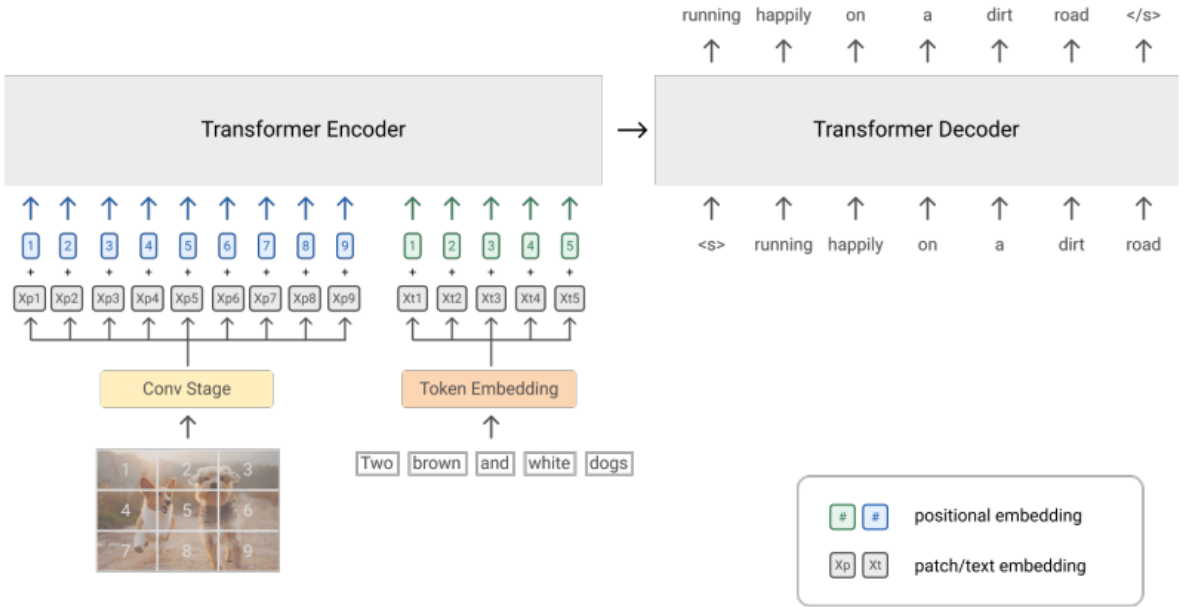
BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation
Coarse-to-Fine Vision-Language Pre-training with Fusion in the Backbone

A closer look at SimVLM

- The model was pre-trained with PrefixLM on 1.8B image-text pairs as used in ALIGN (*private data*)
- PrefixLM: partition the text input randomly into two parts
- Additional text corpus can be naturally used for pre-training
- Strong performance after finetuning



SimVLM

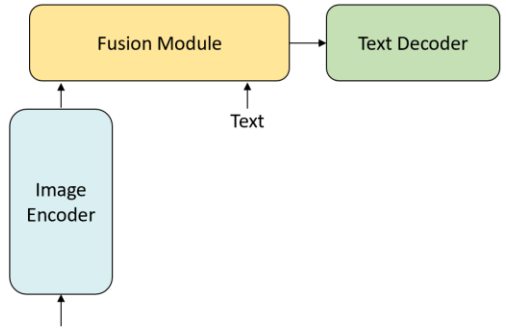


	VQA	
	test-dev	test-std
LXMERT	72.42	72.54
VL-T5	-	70.30
SOHO	73.25	73.47
SimVLM$_{base}$	77.87	78.14
UNITER	73.82	74.02
OSCAR	73.61	73.82
Villa	74.69	74.87
UNIMO	75.06	75.27
VinVL	76.56	76.60
SimVLM$_{large}$	79.32	79.56
SimVLM$_{huge}$	**80.03**	**80.34**

SimVLM: Simple Visual Language Model Pretraining with Weak Supervision

A closer look at SimVLM

- Besides SOTA after finetuning, it also shows strong zero-shot generalization capability
  - Few-shot: using 1% training data
  - Using a prefix prompt “A picture of” improves the quality of decoded captions
- The generative approach also enables open-ended VQA naturally
- Also strong results on ImageNet linear probe



SimVLM

	Setup	CoCo Caption B@4	M	C	S
BUTD[a†]	supervised	36.3	27.7	120.1	21.4
AoANet[b†]		39.5	29.3	129.3	23.2
M2 Transformer[c†]		39.1	29.2	131.2	22.6
$\text{SimVLM}_{\text{base}}$	zero-shot	9.5	11.5	24.0	7.5
$\text{SimVLM}_{\text{large}}$		10.5	12.0	24.9	8.3
$\text{SimVLM}_{\text{huge}}$		11.2	14.7	32.2	8.5
$\text{SimVLM}_{\text{base}}$	few-shot	34.7	29.2	118.7	21.9
$\text{SimVLM}_{\text{large}}$		35.4	30.2	124.1	22.7
$\text{SimVLM}_{\text{huge}}$		36.8	31.5	131.3	24.0
OSCAR[†]	pretrain-finetune	**41.7**	30.6	140.0	24.5
VinVL[†]		41.0	31.1	140.9	25.2
$\text{SimVLM}_{\text{huge}}$		40.6	**33.7**	**143.3**	**25.4**

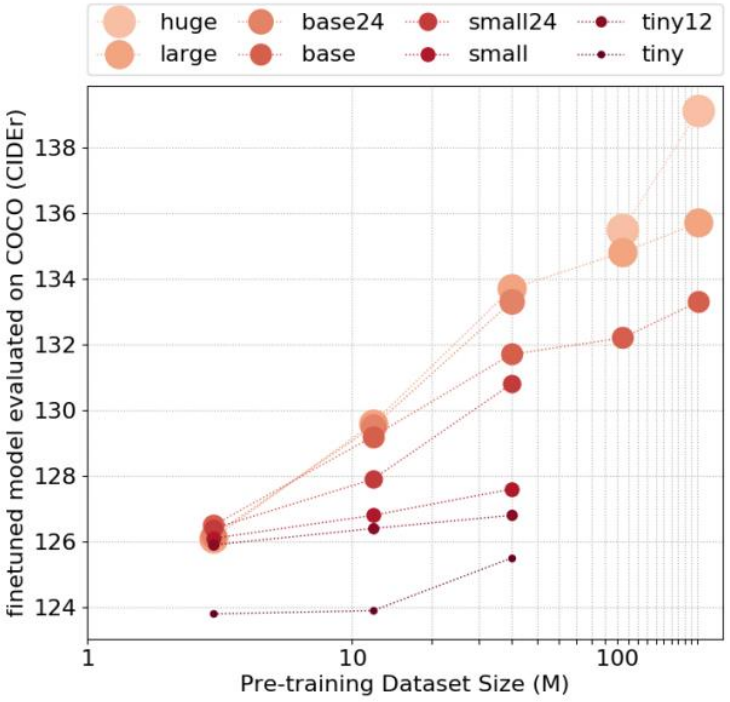
	Dev	Karpathy-test In-domain	Out-domain	Overall
				Discriminative
UNITER	-	74.4	10.0	70.5
VL-T5	-	70.2	7.1	66.4
VL-BART	-	69.4	7.0	65.7
$\text{SimVLM}_{\text{base}}$	73.8	79.0	16.7	75.3
$\text{SimVLM}_{\text{large}}$	76.0	80.4	17.3	76.7
$\text{SimVLM}_{\text{huge}}$	**76.5**	**81.0**	17.5	**77.2**
				Generative
VL-T5	-	71.4	13.1	67.9
VL-BART	-	72.1	13.2	68.6
$\text{SimVLM}_{\text{base}}$	73.2	78.3	25.8	75.2
$\text{SimVLM}_{\text{large}}$	75.2	79.5	29.6	76.5
$\text{SimVLM}_{\text{huge}}$	75.5	79.9	**30.3**	77.0

Method	Acc@1
SimCLRv2 (Chen et al., 2020a)	79.8
DINO (Caron et al., 2021)	80.1
CLIP (Radford et al., 2021)	85.4
ALIGN (Jia et al., 2021)	**85.5**
$\text{SimVLM}_{\text{base}}$	80.6
$\text{SimVLM}_{\text{large}}$	82.3
$\text{SimVLM}_{\text{huge}}$	83.6

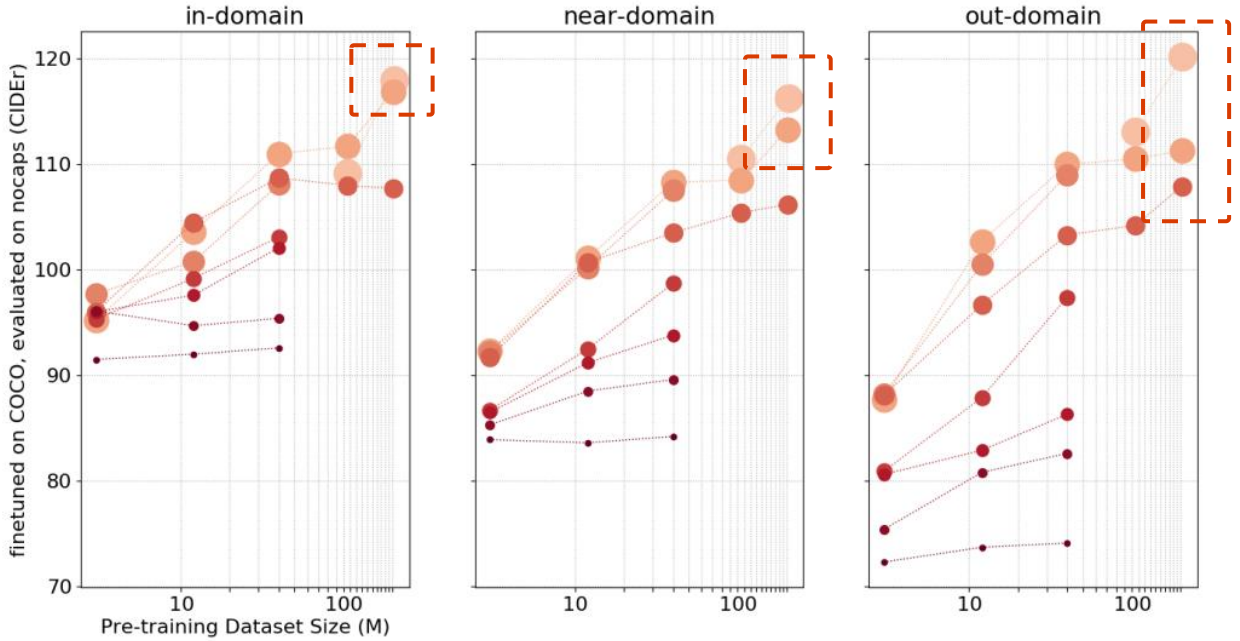
Table 5: Linear evaluation on ImageNet classification, compared to state-of-the-art representation learning methods.

A closer look at LEMON

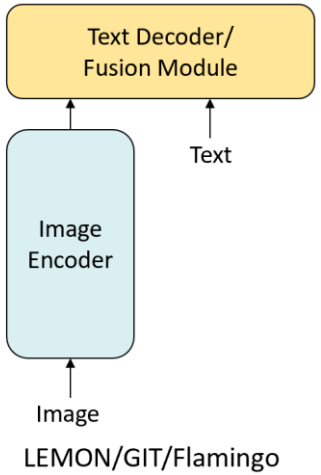
- The authors study the scaling law of VLP models for image captioning
- Scaling up model size (from 13M tiny to 675M huge) and data size (from 3M to 200M)
  - Strong performance boost for out-domain captioning with large-scale pre-training
  - Larger models benefit more from the large-scale data



(a) finetuned and evaluated on COCO



(b) finetuned on COCO, evaluated on nocaps



Scaling Up Vision-Language Pre-training for Image Captioning

A closer look at LEMON

- The authors also collect ALT200M
  - No human annotation required
  - Cover rich visual concepts, while some texts are not well-formed sentences, some do not exactly reflect image content

Dataset	#images (M)	#cap./image	Unigram	
			#unique	#unique in 0.1% tail
COCO Caption [5]	0.1	5	19,264	1,184
CC3M [40]	3.1	1	49,638	22,677
CC12M [4]	12.2	1	1,319,284	193,368
ALT200M (Ours)	203.4	1	2,067,401	1,167,304

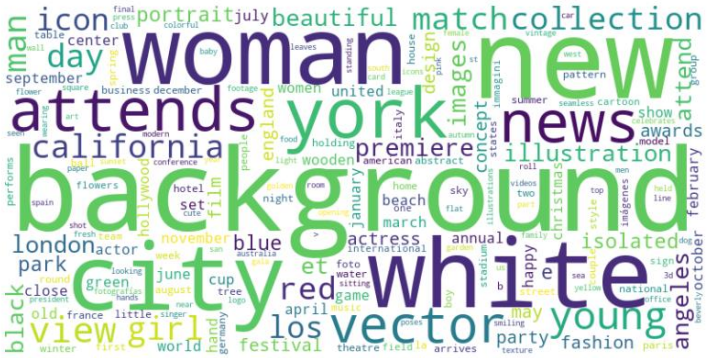
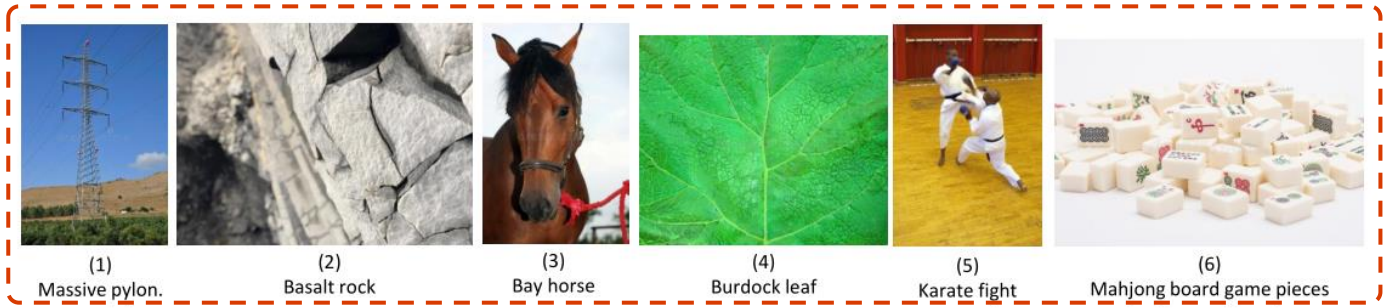


Figure 3. **Word cloud of the top 200 words** in our pre-training dataset ALT200M, excluding the stop words, *e.g.*, a, the, of, *etc*.



(1) Massive pylon.

(2) Basalt rock

(3) Bay horse

(4) Burdock leaf

(5) Karate fight

(6) Mahjong board game pieces

(7) Great landscape with vicia cracca flower in nature

(8) Portrait of a great blue heron standing on a tree branch with one leg tucked up inside. A great blue heron stands on a tree bough on one leg against a background

(9) Cacao fruit, Fresh cocoa pod in hands. With a cocoa plant in background agriculture asia bahia bean beans botanical botany branch brown chocolate climate

(10) Cartoon fat man riding recumbent exercise bikes

(11) Watercolor Potion Halloween Wreath Garland Leaves Wreath Bats

(12) I got you

(13) Happy father's day

(14) After the first COVID-19 positive case was detected, the residents of a number of locations in <LOC> town sealed their areas. After the first COVID-19 Novel

(15) First lady <PERSON> speaks about the government and community efforts to fight childhood obesity while visiting the YMCA January 28 2010

(16) NASCAR - Fans in <LOC> Pre Race

A closer look at LEMON

- SOTA on public benchmarks (COCO, nocaps)
- Zero-shot image captioning, capable of recognizing diverse visual contents
  - B: no pretrain
  - F: pre-trainig + finetuning
  - Z: pre-training only

#	Model	Pre-training data	in-domain CIDEr	in-domain SPICE	near-domain CIDEr	near-domain SPICE	out-of-domain CIDEr	out-of-domain SPICE	overall CIDEr	overall SPICE
15	Human		80.6	15.0	84.6	14.7	91.6	14.2	85.3	14.6
16	SimVLM_{base}	1.8B	-	-	-	-	-	-	94.8	13.1
17	SimVLM_{large}	1.8B	-	-	-	-	-	-	108.5	14.2
18	$\text{SimVLM}_{huge}{}^{\ddagger}$	1.8B	109.0	14.6	110.8	14.6	109.5	**13.9**	110.3	14.5
19	LEMON_{large}	ALT200M	111.2	**15.6**	112.3	**15.2**	105.0	13.6	110.9	**15.0**
20	LEMON_{huge}	ALT200M	**112.8**	15.2	**115.5**	15.1	**110.1**	13.7	**114.3**	14.9

B: a woman holding a pink umbrella over her head.
F: a woman in a kimono holding a purple umbrella.
Z: a picture of a geisha

B: a close up of a tree branch
F: a close up of a dinosaur skull with a black background
Z: a picture of a dinosaur skeleton

B: a picture of a cat with a red tail.
F: a black and white image of a killer whale.
Z: a picture of a killer whale

B: a man wearing a hat and a straw hat standing in front of a large metal instrument.
F: a man in a green hat playing a tuba.
Z: a picture of a bavarian musician playing a tuba

B: a collection of wooden tools on a white background.
F: a collection of swords on display in a museum.
Z: a picture of ancient swords

Scaling Up Vision-Language Pre-training for Image Captioning

A closer look at GIT

- GIT is a simple generative image-to-text transformer, which is pre-trained on 800M image-text pairs (*public+private*) via a simple LM loss
  - Instead of using an OD module, the authors use a big Swin-like model as image encoder
- The image encoder is first pre-trained via image-text contrastive loss
  - The same strategy is also used in Flamingo, while in CoCa, contrastive and captioning losses are used together

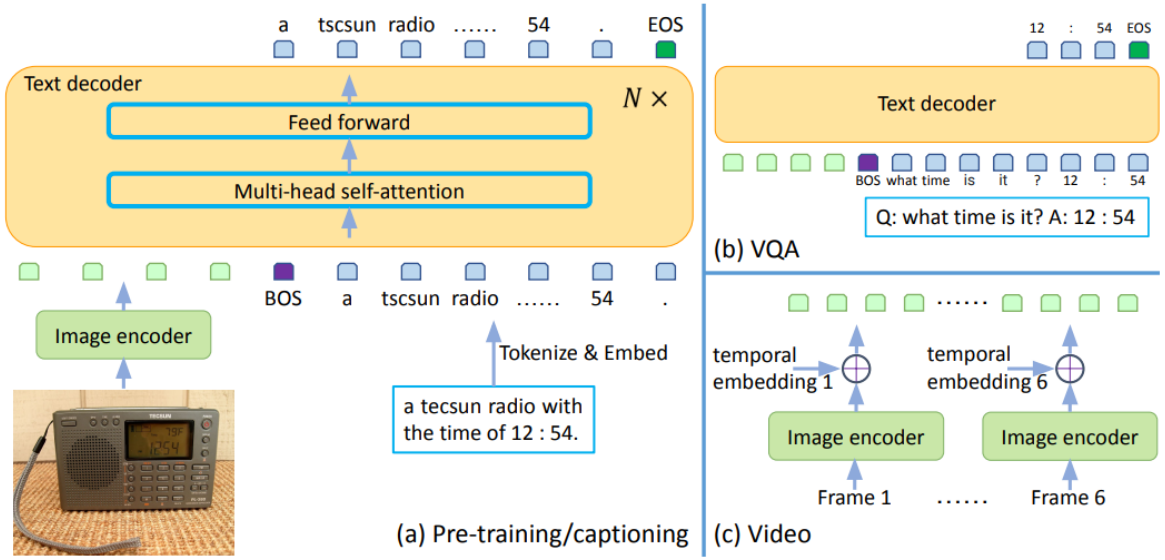
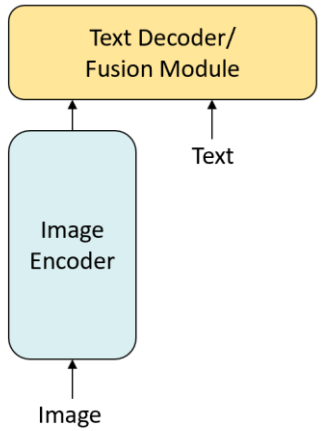


Figure 2: Network architecture of our GIT, composed of one image encoder and one text decoder. (a): The training task in both pre-training and captioning is the language modeling task to predict the associated description. (b): In VQA, the question is placed as the text prefix. (c): For video, multiple frames are sampled and encoded independently. The features are added with an extra learnable temporal embedding (initialized as 0) before concatenation.



LEMON/GIT/Flamingo

- VQA tasks are tackled as a text generation tasks
- Video tasks are tackled via concatenating image frame features

A closer look at GIT

- GIT achieves SOTA over 12 image/video captioning and QA tasks, including the first human parity on TextCaps (no OCR engine is used)

Table 1: New state-of-the-art performance with our GIT across 12 image/video captioning and question answering (QA) tasks. *: evaluated on the public server. CIDEr scores are reported for Captioning tasks.

	Image captioning				Image QA			Video captioning			Video QA	
	COCO*	nocaps*	VizWiz*	TextCaps*	ST-VQA*	VizWiz*	OCR-VQA	MSVD	MSRVTT	VATEX*	MSVD-QA	TGIF-Frame
Prior SOTA	138.7 [114]	120.6 [109]	94.1 [22]	109.7 [107]	59.7 [107]	65.4 [2]	64.1 [28]	120.6 [60]	60 [80]	86.5 [88]	48.3 [91]	69.5 [112]
GIT (ours)	148.8	123.4	114.4	138.2	69.6	67.5	68.1	180.2	73.9	93.8	56.8	72.8
Δ	+10.1	+3.7	+20.3	+28.5	+9.9	+2.1	+4.0	+59.6	+13.9	+7.3	+8.5	+3.3

a

Model: a macdonald ' s sign that is on a brick wall

Human: A tile wall with a red circle on it reading Mornington Crescent

b

Model: a sign that has the time of 12 : 37 on it

Human: A kiosk of track 13 of Metra which states that the 5:43 train has moved tracks

- GIT provides a generative scheme for image classification and scene text recognition

Table 10: Results on ImageNet-1k classification task. Our approach takes the class name as the caption and predict the label in an auto-regressive way without pre-defining the vocabulary.

Vocabulary	Method	Top-1
Closed	ALIGN [39]	88.64
	Florence [110]	90.05
	CoCa [109]	**91.0**
Open	GIT	88.79

Generate class names token-by-token

Table 11: Results on scene text recognition. MJ and ST indicate the MJSynth (MJ) [36, 37] and SynthText (ST) [25] datasets used for training scene text recognition models.

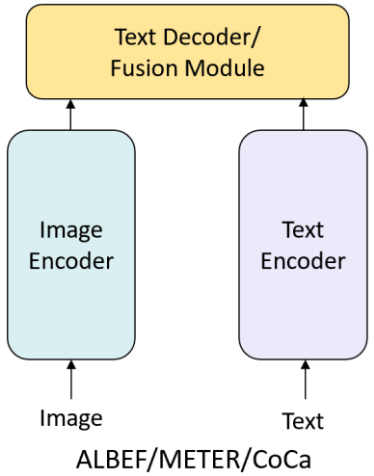
Method	Fine-tuning Data	Regular Text			Irregular Text			Average
		IC13[43]	SVT [95]	IIIT [68]	IC15[42]	SVTP[74]	CUTE [79]	
SAM [58]	MJ+ST	95.3	90.6	93.9	77.3	82.2	87.8	87.8
Ro.Scanner [111]	MJ+ST	94.8	88.1	95.3	77.1	79.5	90.3	87.5
SRN [108]	MJ+ST	95.5	91.5	94.8	82.7	85.1	87.8	89.6
ABINet [16]	MJ+ST	**97.4**	93.5	**96.2**	**86.0**	89.3	89.2	91.9
S-GTR [29]	MJ+ST	96.8	94.1	95.8	84.6	87.9	92.3	91.9
GIT	TextCaps	94.2	91.5	92.9	78.2	87.1	95.5	89.9
	MJ+ST	97.3	**95.2**	95.3	83.7	**89.9**	**96.2**	**92.9**

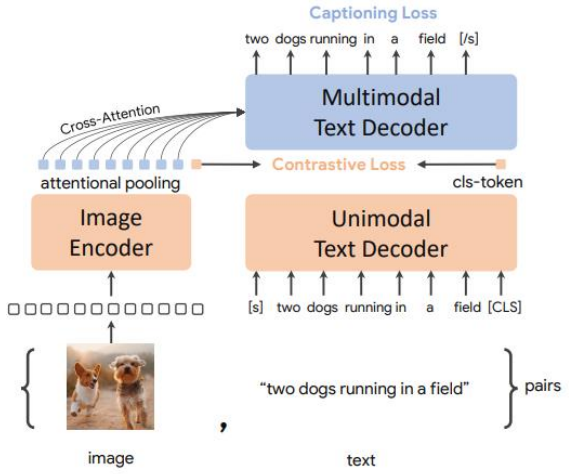
A closer look at GIT

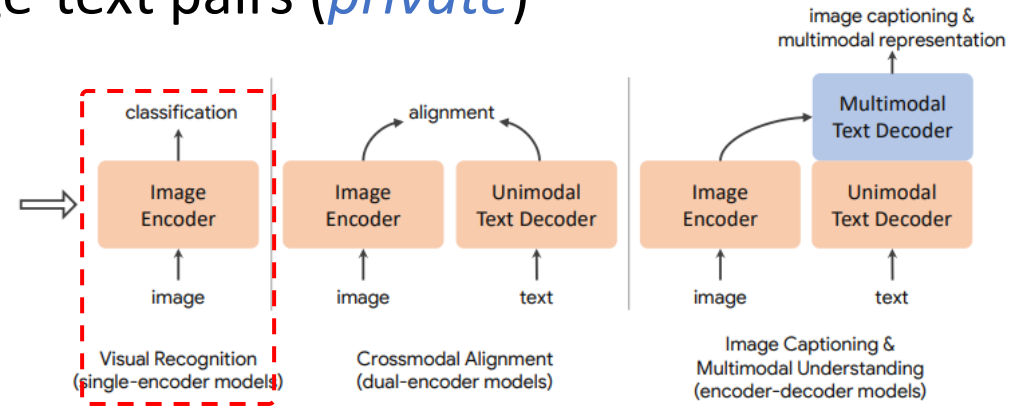
Figure 1: Example captions generated by GIT. The model demonstrates strong capability of recognizing scene text, tables/charts, food, banknote, logos, landmarks, characters, products, *etc.*

A closer look at CoCa

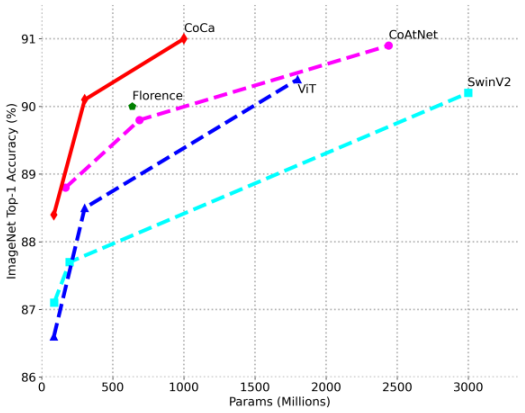
- All trained from scratch, including both image and text encoders
- Pre-trained via a combination of contrastive and generative losses
- Pre-training dataset: JFT-3B + 1.8B image-text pairs (*private*)



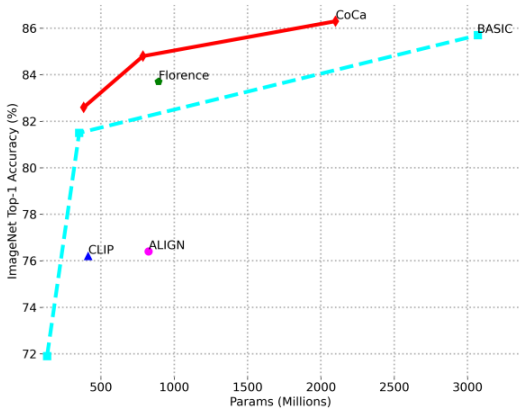




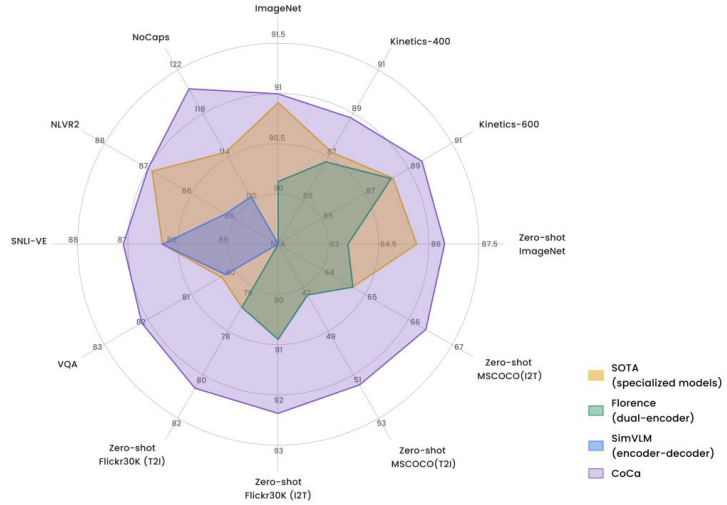
Zero-shot, frozen-feature or finetuning



(a) Finetuned ImageNet Top-1 Accuracy.



(b) Zero-Shot ImageNet Top-1 Accuracy.



What are the publicly available pre-training data?

	#Image-Text Pairs	Avg. text length
COCO [66]	0.9M	12.4
SBU Captions [77]	1.0M	12.1
Localized Narratives [82]	1.9M	13.8
Conceptual Captions [92]	3.1M	10.3
Visual Genome [57]	5.4M	5.1
Wikipedia Image Text [99]	4.8M	12.8
Conceptual Captions 12M [14]	11.0M	17.3
Red Caps [27]	11.6M	9.5
YFCC100M [103], filtered	30.3M	12.7
Total	70M	12.1

Table 2. Public Multimodal Datasets (PMD) corpus used in FLAVA multimodal pretraining, which consists of publicly available datasets with a total size of 70M image and text pairs.

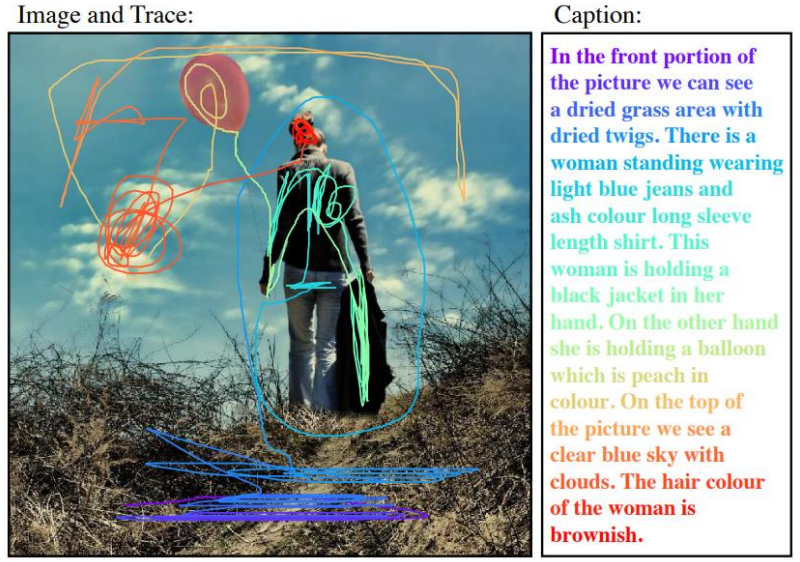
FLAVA: A Foundational Language And Vision Alignment Model

<PERSON> was the first US president to attend a tournament in sumo's hallowed Ryogoku Kokugikan arena. (AFP photo)

Hand holding a fresh mangosteen

#jellyfish #blue #ocean #pretty Sea Turtle Wallpaper, Aquarius Aesthetic, Blue Aesthetic Pastel, The Adventure Zone, Capricorn And <PERSON>, Life Aquatic, Ocean Life, Jellyfish, Marine Life

Figure 1: **CC12M** Even when the alt-texts do not precisely describe their corresponding Web images, they still provide rich sources for learning long-tail visual concepts such as sumo, mangosteen, and jellyfish. We scale up vision-and-language pre-training data to 12 million by relaxing overly strict filters in Conceptual Captions [70].



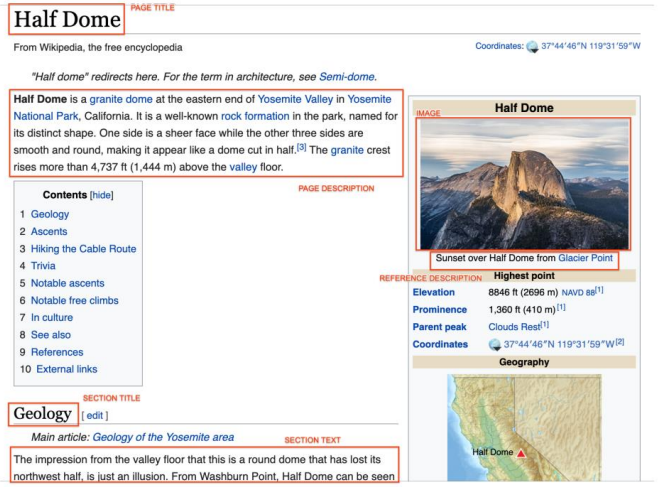


Figure 2: The Wikipedia page for Half Dome, Yosemite, California via Wikimedia Commons with examples of the different fields extracted and provided in WIT.

WIT features images and multilingual texts collected from the Wikipedia content pages.

LAION-400M/5B

However, it remains unclear how useful each dataset is

Figure 1: **RedCaps dataset** comprises 12M image-text pairs from 350 subreddits. RedCaps data is created *by the people, for the people* – it contains everyday things that users like to share on social media, for example hobbies (**r/crafts**) and pets (**r/shiba**). Captions often contain specific and fine-grained descriptions (northern cardinal, taj mahal). Subreddit names provide relevant image labels (**r/shiba**) even when captions may not (mlem!), and sometimes may group many visually unrelated images through a common semantic meaning (**r/perfectfit**).

WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning

RedCaps: web-curated image-text data created by the people, for the people

Part II: In-Context Few-Shot Learning

What's in-context few-shot learning?

- Getting SOTA via full finetuning is great. But can we train a model that can quickly adapt to different downstream tasks via only providing a few in-context examples?

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French:        ← task description
2 cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French:        ← task description
2 sea otter => loutre de mer          ← example
3 cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

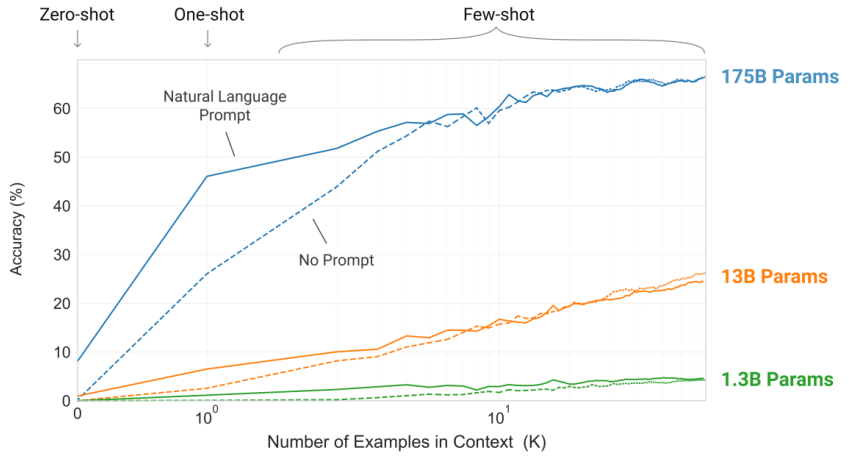
```
1 Translate English to French:        ← task description
2 sea otter => loutre de mer          ← examples
3 peppermint => menthe poivrée
4 plush girafe => girafe peluche
5 cheese =>                           ← prompt
```

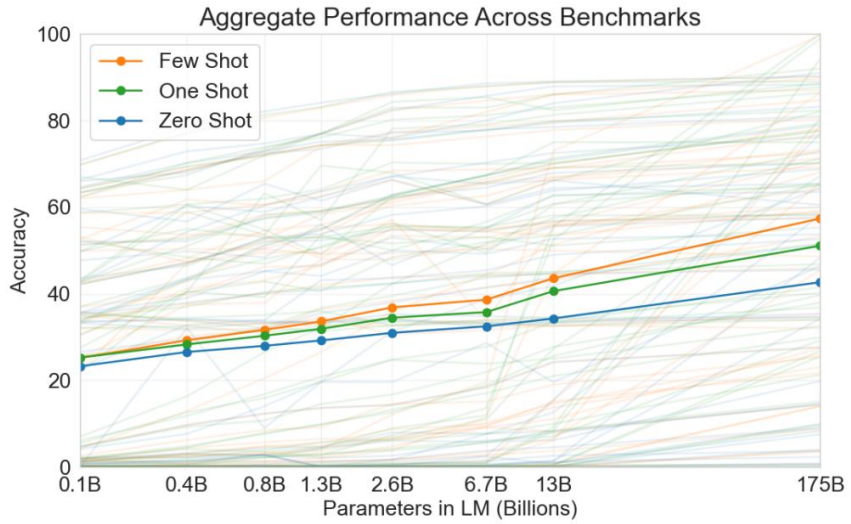
Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
1 sea otter => loutre de mer          ← example #1
        ↓
   gradient update
        ↓
1 peppermint => menthe poivrée        ← example #2
        ↓
   gradient update
        ↓
        ...
        ↓
1 plush giraffe => girafe peluche     ← example #N
        ↓
   gradient update
        ↓
1 cheese =>                           ← prompt
```





Multimodal Few-Shot Learning with Frozen LM

- We need a strong language model that is kept frozen, and align an image encoder towards this language embedding space
  - This is an opposite design choice compared with LiT, since the end-goal is different
  - Frozen: in-context learning capability based on LM
  - LiT: zero-shot transfer on image classification based on CLIP

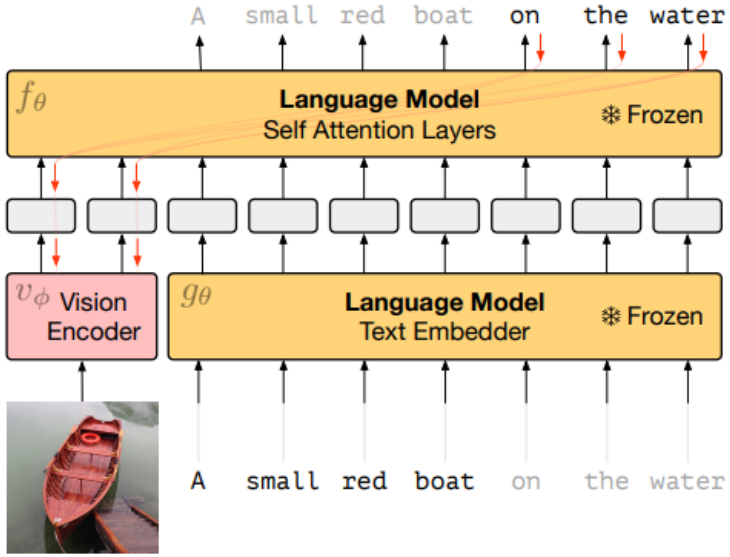


Figure 2: Gradients through a frozen language model's self attention layers are used to train the vision encoder.

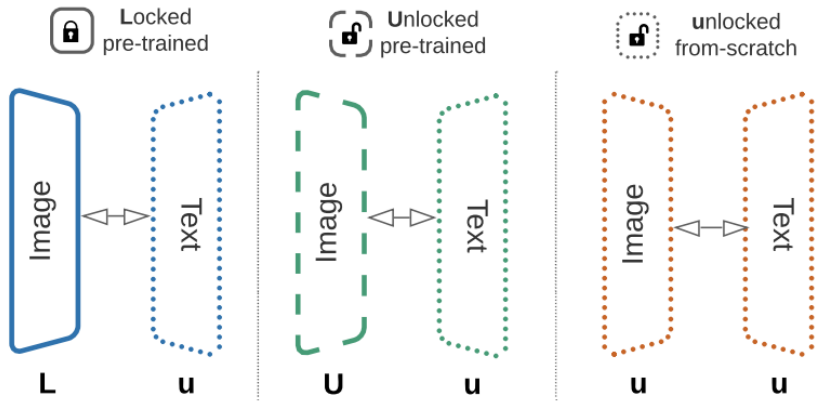


Figure 2. Design choices for contrastive-tuning on image-text data. Two letters are introduced to represent the image tower and text tower setups. `L` stands for locked variables and initialized from a pre-trained model, `U` stands for unlocked and initialized from a pre-trained model, `u` stands for unlocked and randomly initialized. `Lu` is named as "Locked-image Tuning" (LiT).

Multimodal Few-Shot Learning with Frozen LM

- Interesting in-context few-shot learning can be inherited from the frozen LM
  - Image encoder: NF-ResNet-50; however, an image is compressed into 2 global vectors
  - LM: 7B parameter transformer pre-trained on C4

Figure 1: Curated samples with about five seeds required to get past well-known language model failure modes of either repeating text for the prompt or emitting text that does not pertain to the image. These samples demonstrate the ability to generate open-ended outputs that adapt to both images and text, and to make use of facts that it has learned during language-only pre-training.

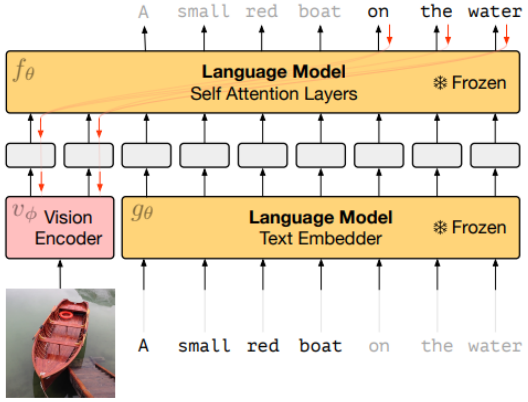
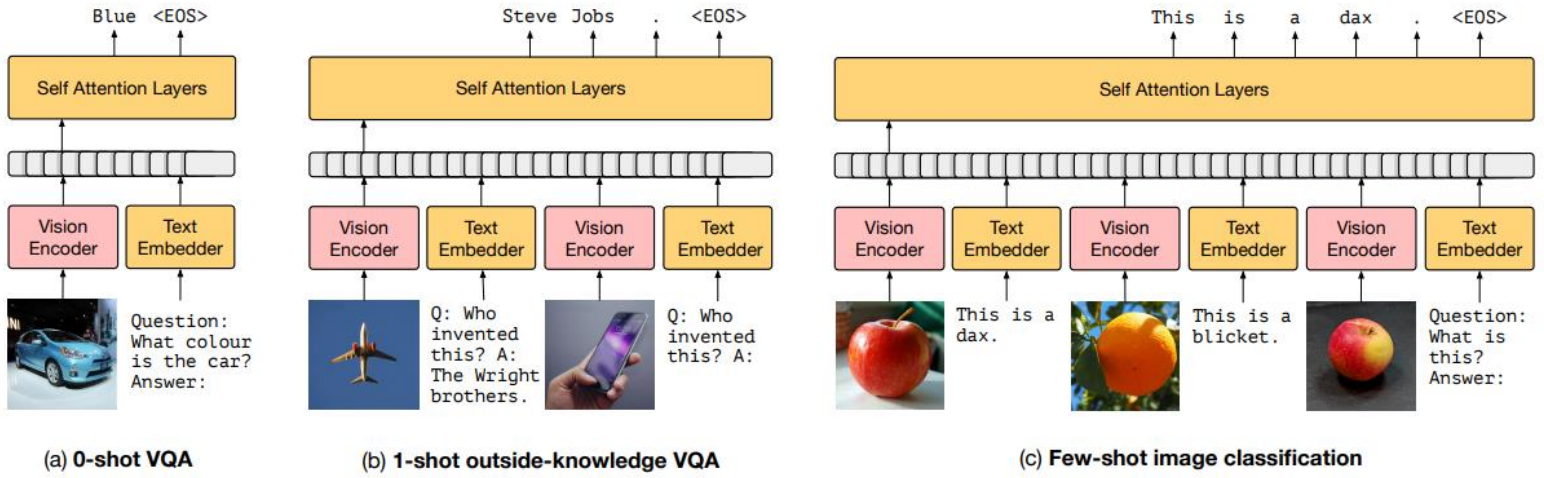


Figure 2: Gradients through a frozen language model's self attention layers are used to train the vision encoder.

Multimodal Few-Shot Learning with Frozen LM

- How to perform in-context few-shot learning



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

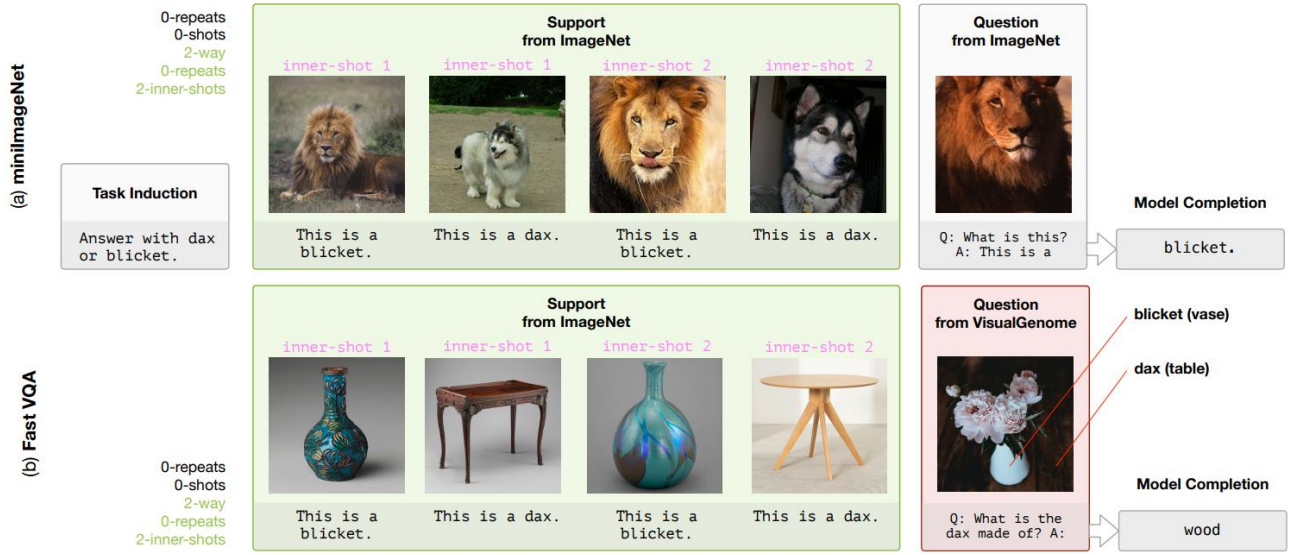


Figure 4: Examples of (a) the Open-Ended miniImageNet evaluation (b) the Fast VQA evaluation.

n-shot Acc.	n=0	n=1	n=4	τ
Frozen	29.5	35.7	38.2	✗
Frozen scratch	0.0	0.0	0.0	✗
Frozen finetuned	24.0	28.2	29.2	✗
Frozen train-blind	26.2	33.5	33.3	✗
Frozen VQA	48.4	–	–	✓
Frozen VQA-blind	39.1	–	–	✓
Oscar [23]	73.8	–	–	✓

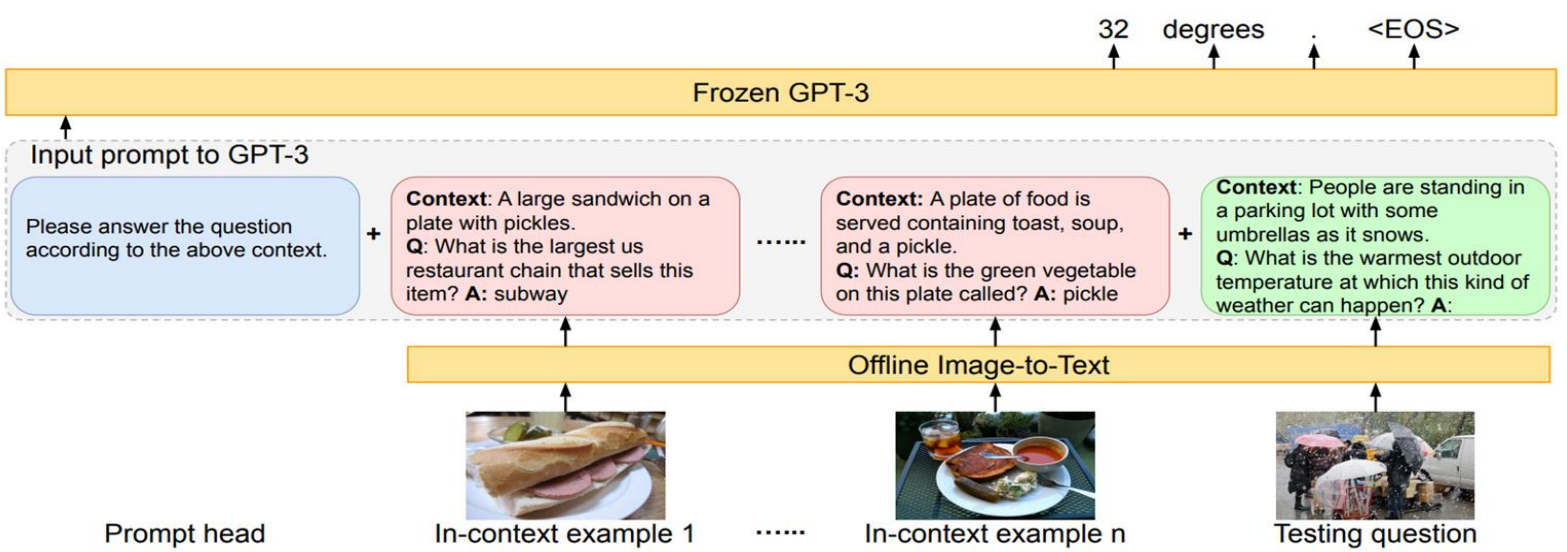
Table 1: Transfer from Conceptual Captions to VQAv2. The τ column indicates whether a model uses training data from the VQAv2 training set. The row denoted *Frozen* train-blind is the blind baseline described in subsection 4.1. *Frozen* VQA is a baseline which mixes in VQAv2 training data.

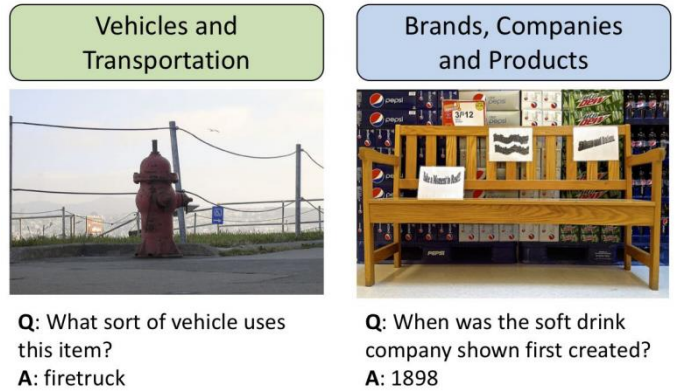
n-shot Acc.	n=0	n=1	n=4	τ
Frozen	5.9	9.7	12.6	✗
Frozen 400mLM	4.0	5.9	6.6	✗
Frozen finetuned	4.2	4.1	4.6	✗
Frozen train-blind	3.3	7.2	0.0	✗
Frozen VQA	19.6	–	–	✗
Frozen VQA-blind	12.5	–	–	✗
MAVEx [42]	39.4	–	–	✓

Table 2: Transfer from Conceptual Captions to OKVQA. The τ column indicates if a model uses training data from the OKVQA training set. *Frozen* does not train on VQAv2 except in the baseline row, and it never trains on OKVQA.

PICa: How about using GPT-3 directly?

- Model size: Frozen 7B vs. GPT-3 175B
  - However, getting gradients from GPT-3 would be non-trivial, which renders training an image encoder to align the language embedding space challenging
- PICa: Prompting GPT-3 via the use of Image Captions
  - Treating GPT-3 as an *implicit* and *unstructured* KB
  - Translate images into captions/tags so that GPT-3 can understand it
  - Focus on OK-VQA that requires external knowledge to correctly answer the question
  - 4 shots outperform supervised SOTA on OK-VQA; also reasonable results on VQA-v2





Why GPT-3 are so powerful for OK-VQA?

- It encodes encyclopedia and commonsense knowledge
- GPT-3 also generates answer rationales reasonably well

(a) When was this type of transportation invented?

Context: A blue and yellow train traveling down train tracks.
Answer: 1804
GT Answer: ['1804', '1804', '1804', '1804', '1800s']
Acc.: 1.0

(a) This is because: first steam-powered locomotive was invented in 1804.

(b) When was this type of two wheeled vehicle invented?

Context: A row of motorcycles parked next to each other.
Answer: 1885
GT Answer: ['1885', '1885', '30's', '1845', '1915']
Acc.: 1.0

(b) This is because: first motorcycle was invented in 1885

(c) Where can you get these?

Context: A shopping cart filled with bananas and other items.
Answer: grocery store
GT Answer: ['grocery', 'grocery', 'supermarket', 'store', 'grocery store']
Acc.: 0.6

(c) This is because: grocery store is most common place get food

(d) Where does this vehicle stop?

Context: A blue and white train traveling down train tracks.
Answer: train station
GT Answer: ['train station', 'train station', 'train station', 'station', 'station']
Acc.: 1.0

(d) This is because: train station is only place where train stops

(e) What animal do you have to watch out for when doing this sport?

Context: A man holding a surfboard on top of a beach.
Answer: shark
GT Answer: ['shark', 'shark', 'shark', 'wave', 'shark']
Acc.: 1.0

(e) This is because: sharks are dangerous animals

Limitations of this approach

- Converting images into captions could lose important visual information

Method	Image Repr.	Few-shot	Acc.
Oscar (Li et al. 2020)	Feature Emb.	✗	73.8
Frozen	Feature Emb.	✓	38.2
PICa-Base	Caption	✓	53.2
PICa-Base	Caption+Tags	✓	54.3
PICa-Full	Caption	✓	55.9
PICa-Full	Caption+Tags	✓	**56.1**
PICa-Full†	GT-Caption-5	✓	59.7

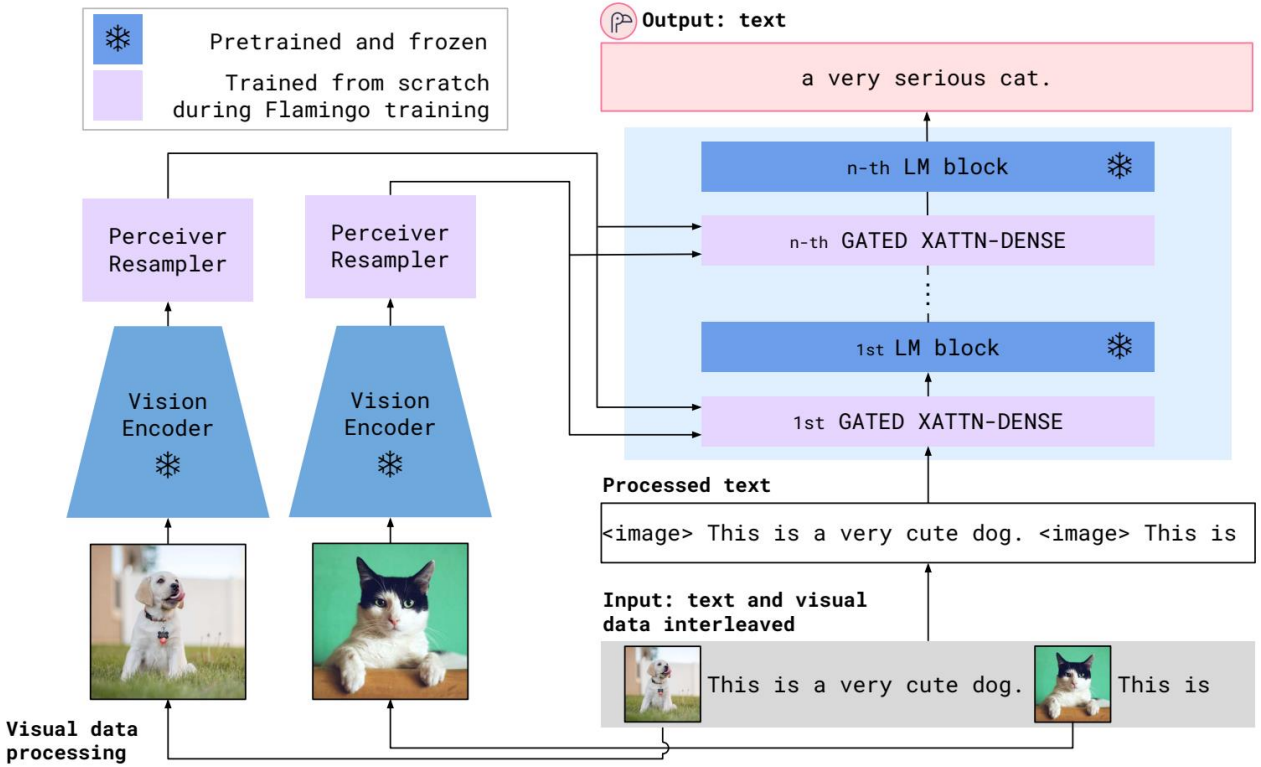
(f) How many giraffes are there?
Context: A herd of giraffe standing next to a wooden fence.
Answer: 3
GT Answer: ['6', '6', '8', '6', '8', '6', '6', '7', '8', '7']

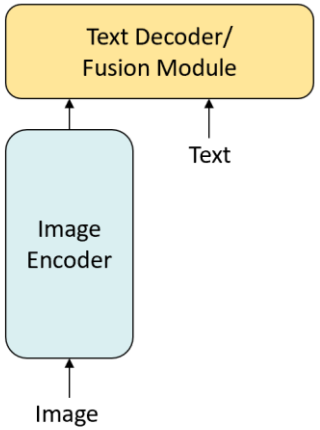
Acc.: 0.0

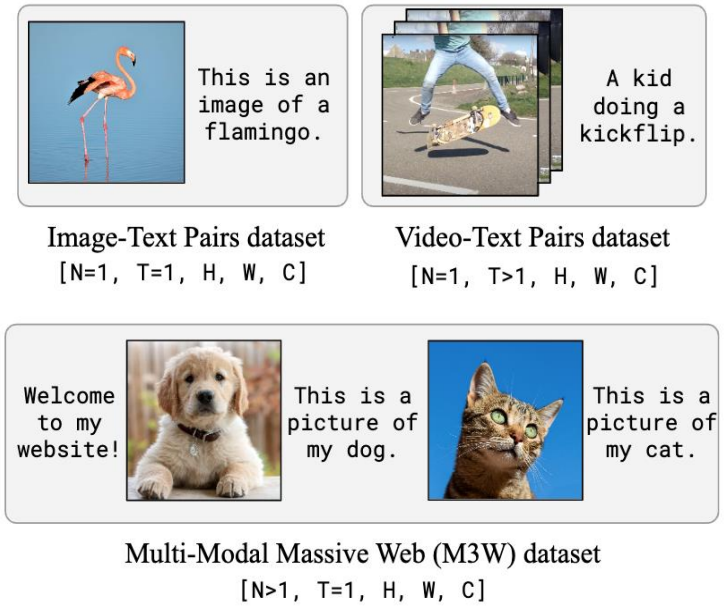
- How about if we have enough computing resources and go beyond this?

Flamingo for few-shot learning

- Both the visual encoder and the LLM are kept frozen
- Besides image/video-text pairs, the pre-training data also includes interleaved image-text data (M3W), which is crucial for in-context learning
- The visual encoder is pre-trained with CLIP-like loss beforehand





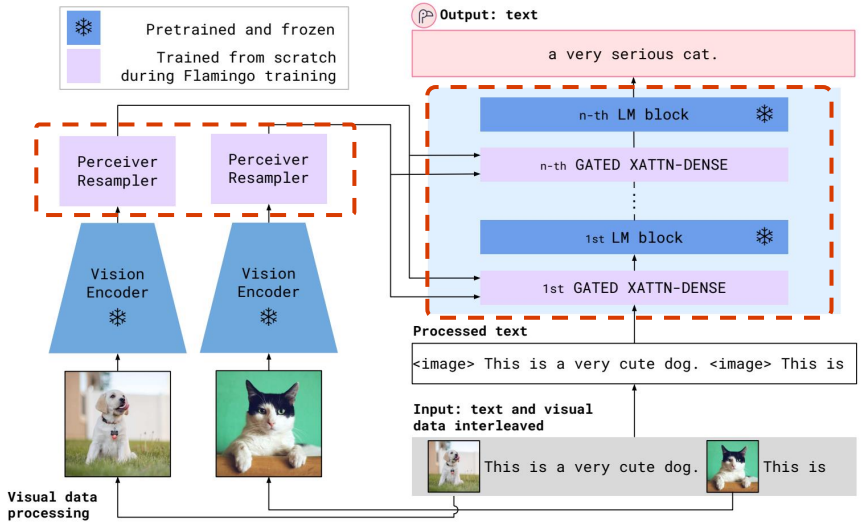


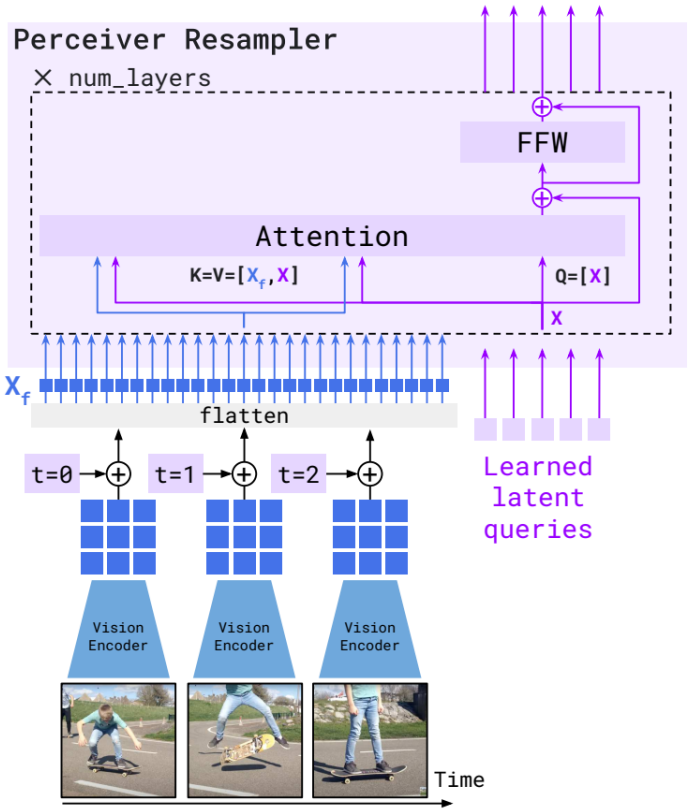
How the pre-training data looks like

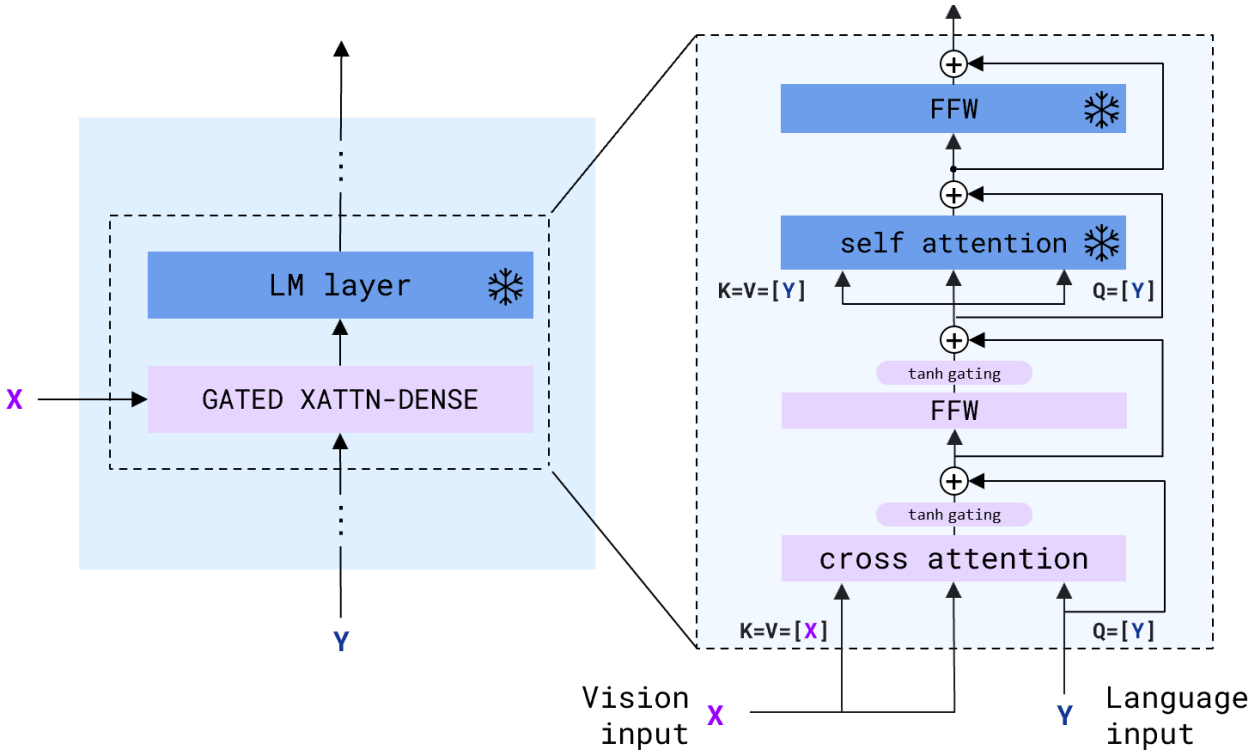
Flamingo: a Visual Language Model for Few-Shot Learning

Flamingo for few-shot learning

- Perceiver resampler: takes as input a variable number of features and outputs a fixed number of “visual tokens”
- Gated X-attn design: bridge vision and text modalities, the tanh gates make the training dynamics of LLM not change too much at the early stage of training

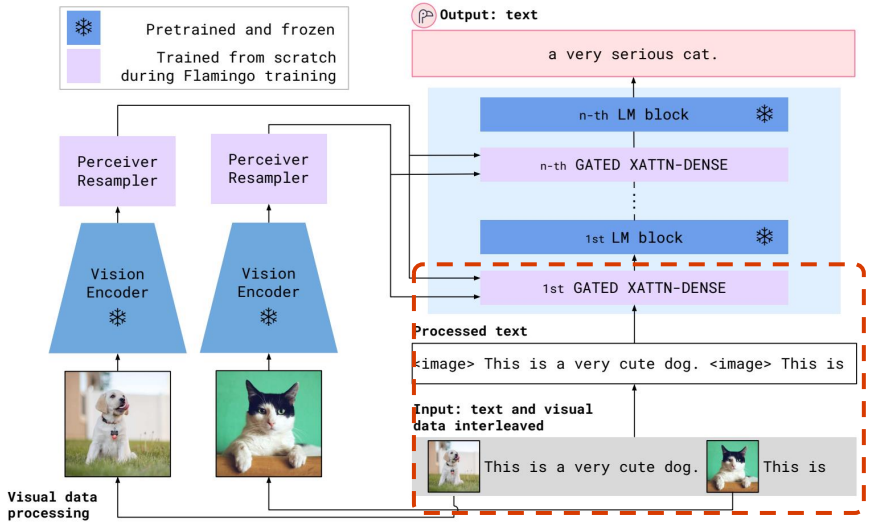


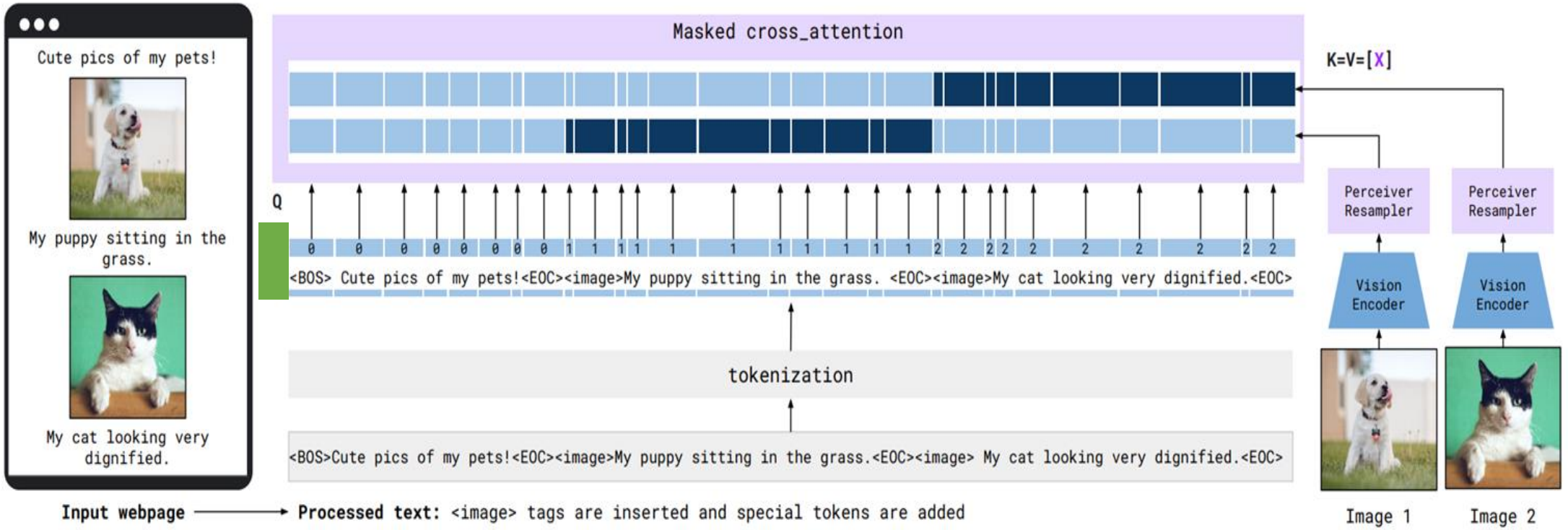




Flamingo for few-shot learning

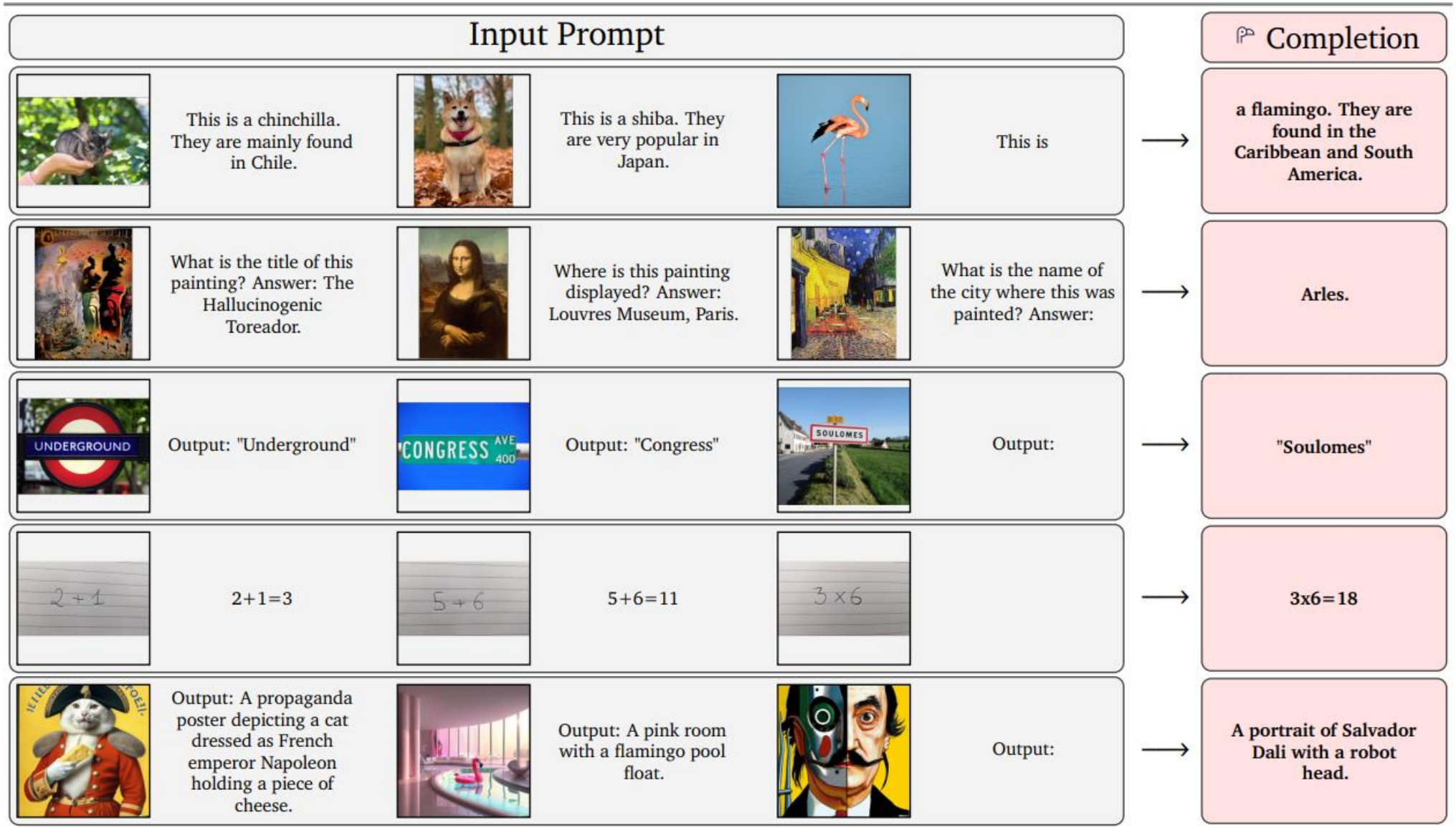
- Each text token cross-attends to the image that precedes it in the interleaved sequence
- The model is trained by autoregressive LM loss





Flamingo: a Visual Language Model for Few-Shot Learning

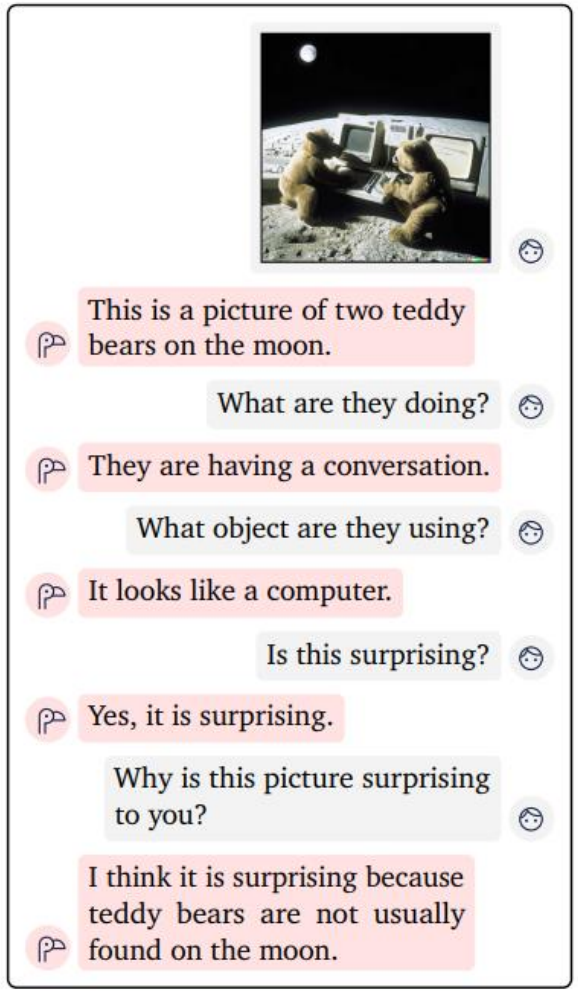
Flamingo for few-shot learning

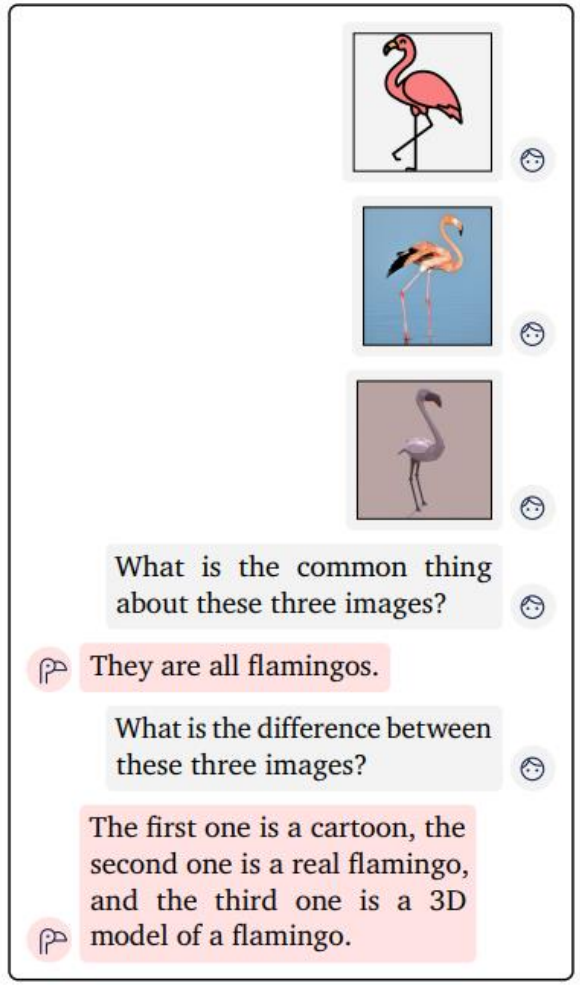


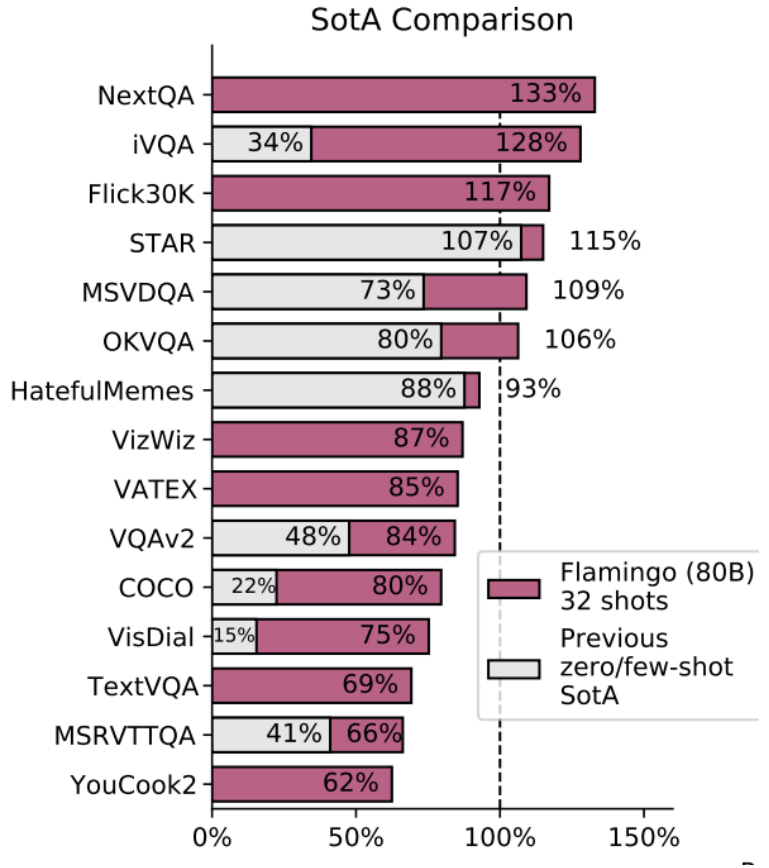
Flamingo: a Visual Language Model for Few-Shot Learning

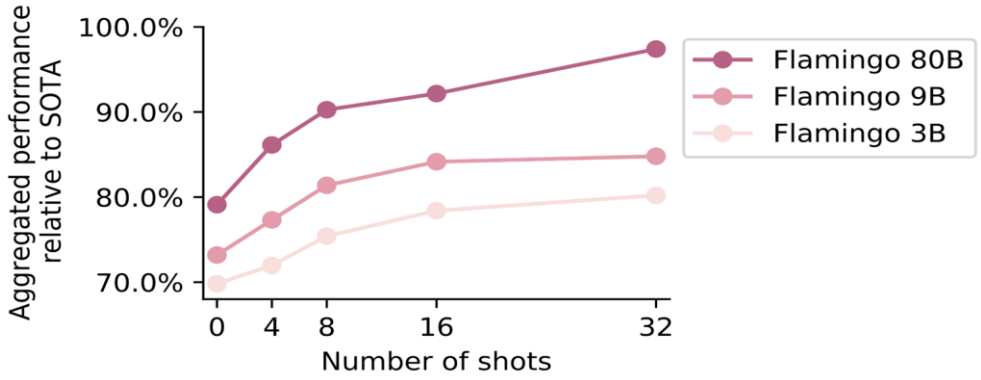
Flamingo results highlight

- Visual dialog capability powered by LLM





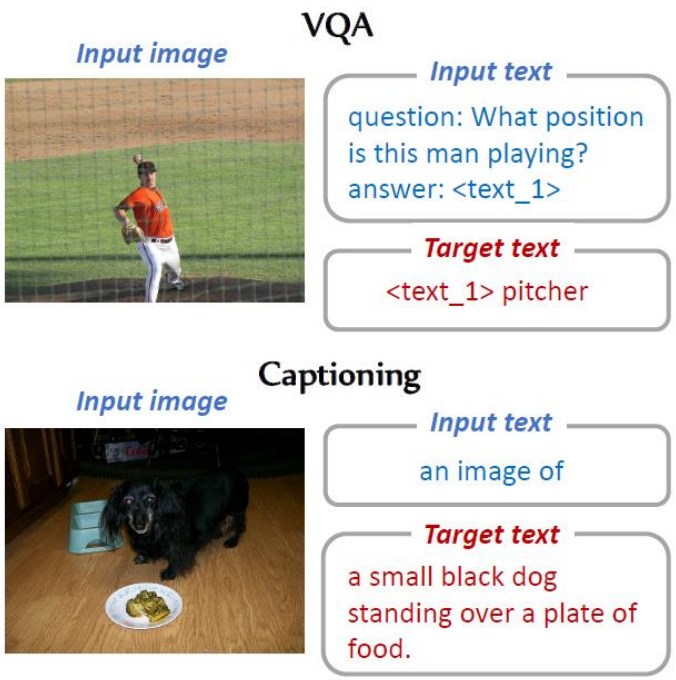


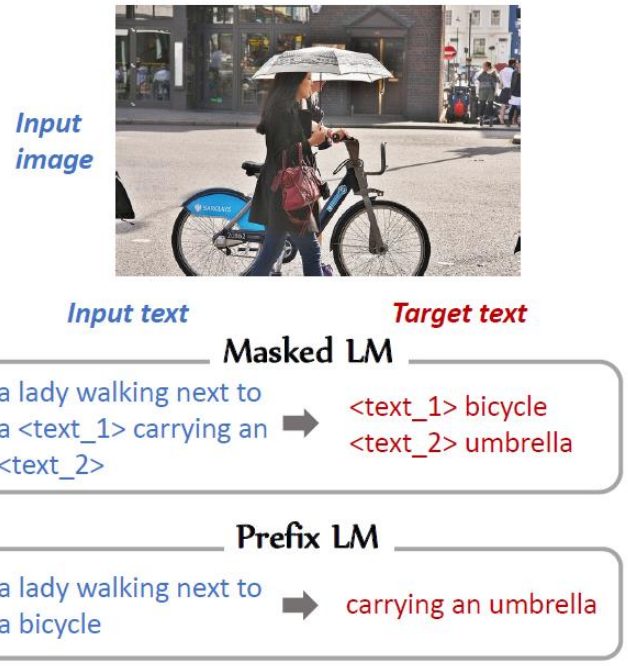


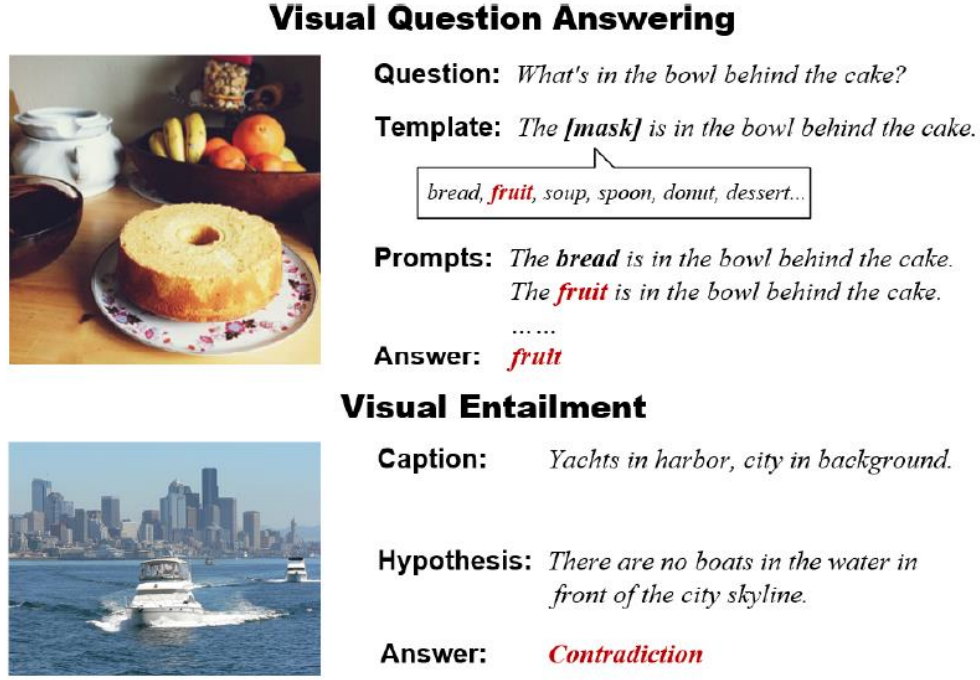
Flamingo: a Visual Language Model for Few-Shot Learning

Other work besides in-context learning

- FewVLM: train a VL-T5-like base model with PrefixLM and MLM, and found that PrefixLM is helpful for zero/few-shot captioning, while MLM is good for zero/few-shot VQA
- TAP-C: CLIP is few-shot learner for VQA and visual entailment.
  - For VQA, the authors propose to reformulate it as a retrieval task
  - For visual entailment, caption and hypothesis (text-text pairs) are used in training, while image and hypothesis (image-text pairs) are used at inference.







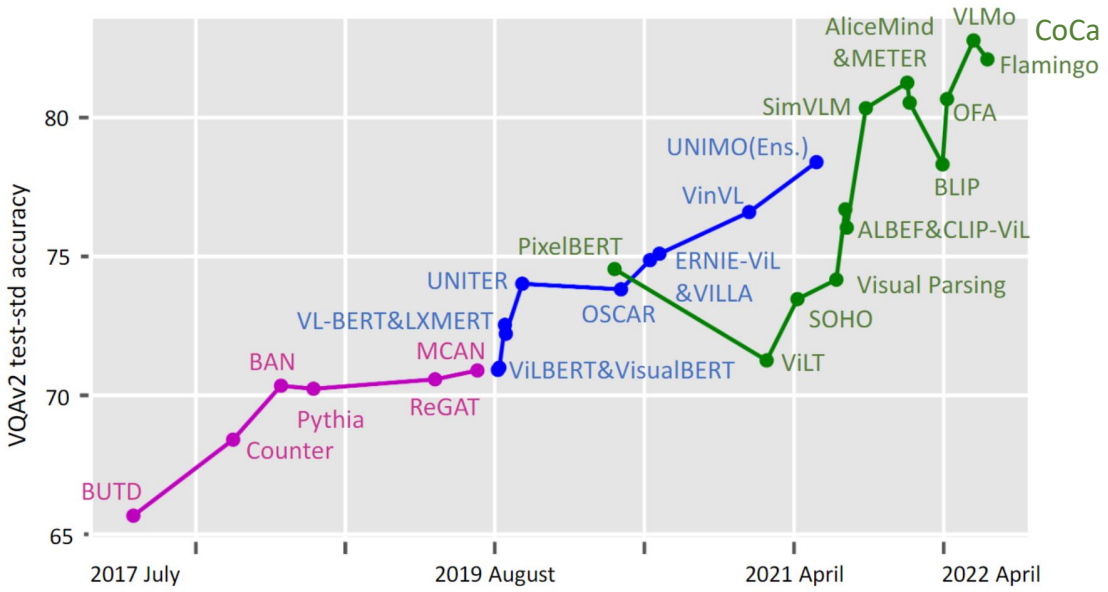
CLIP Models are Few-shot Learners: Empirical Studies on VQA and Visual Entailment
A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models

Part III: Model Evaluation

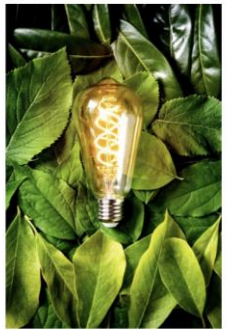
Model evaluation is difficult

- What we do at the current stage
  - We evaluate on tasks/datasets like VQAv2, NLVR2, SNLI-VE, VCR, image-text retrieval, image captioning, RefCOCO etc.
  - These benchmarks have driven tremendous progress in the field (VQA score from 70 to 82)
- They are good, but may not be good enough
  - Especially given the fact that big models are surpassing human performance on certain tasks
  - We should try to avoid both *over-claiming* and *under-claiming*
  - Not just focus on topping the leaderboard, but also testing the learned abilities
  - *An open question* for the community

- Some robustness analysis in the field
  - Diagnostic tests
  - Challenge sets (OOD)
  - Adversarial attacks
  - Probing



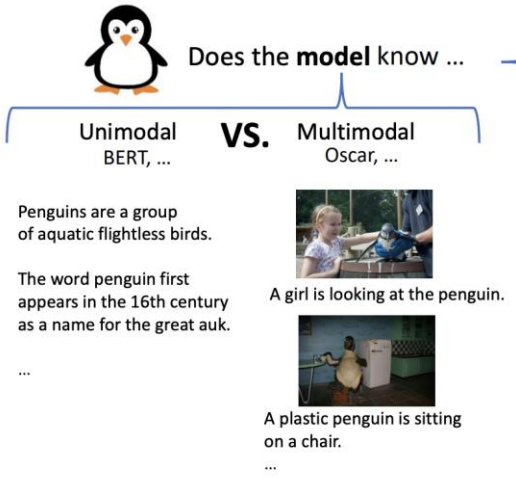
Example diagnostic tests



(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

Winoground: Compositionality

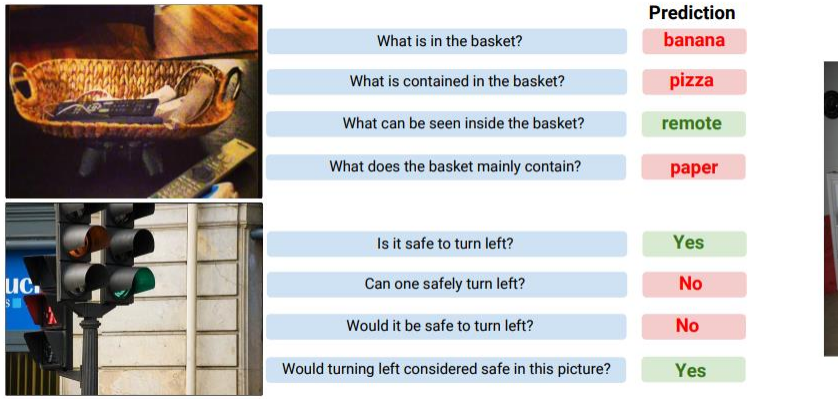


Commonsense

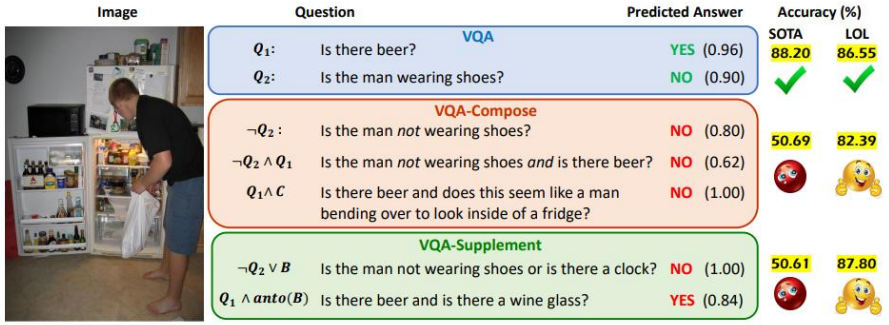
Q: How many magnets are on the bottom of the fridge?

A: 5
A: 2
A: 3
A: 4

Counting



Rephrasing



Logical reasoning

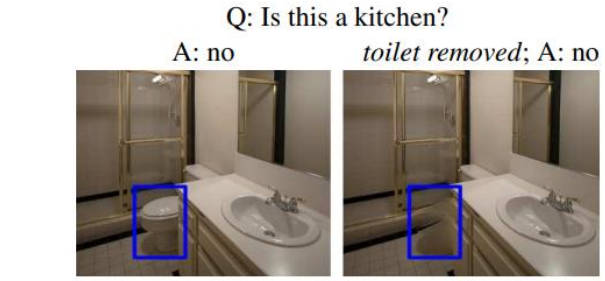
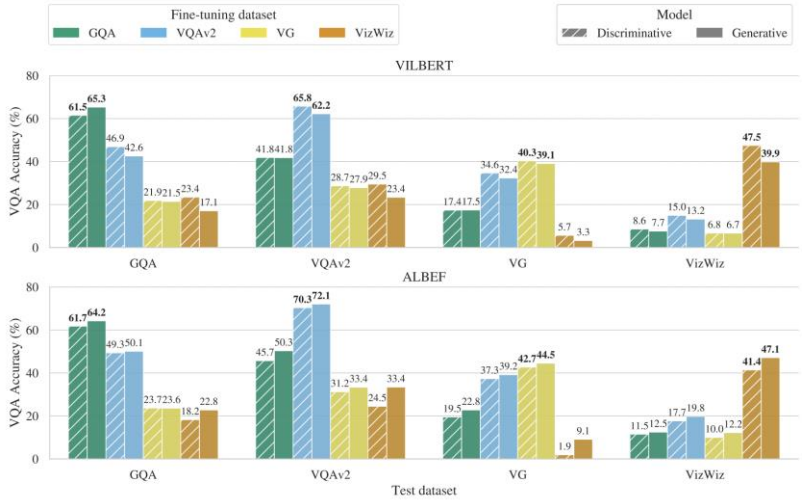
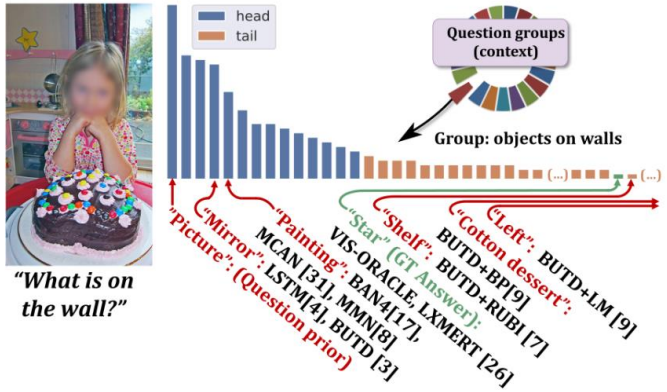


Image editing

Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality
Visual Commonsense in Pretrained Unimodal and Multimodal Models
Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks
Cycle-Consistency for Robust Visual Question Answering
VQA-LOL: Visual Question Answering under the Lens of Logic
Towards Causal VQA: Revealing and Reducing Spurious Correlations by Invariant and Covariant Semantic Editing

OOD generalization





GRIT

General Robust Image Task Benchmark

Generality

Evaluate performance to novel domains and novel concepts for each task

Robustness

Measure performance degradation with image perturbations

Calibration

Quantify misinformation and confidence calibration

Tasks

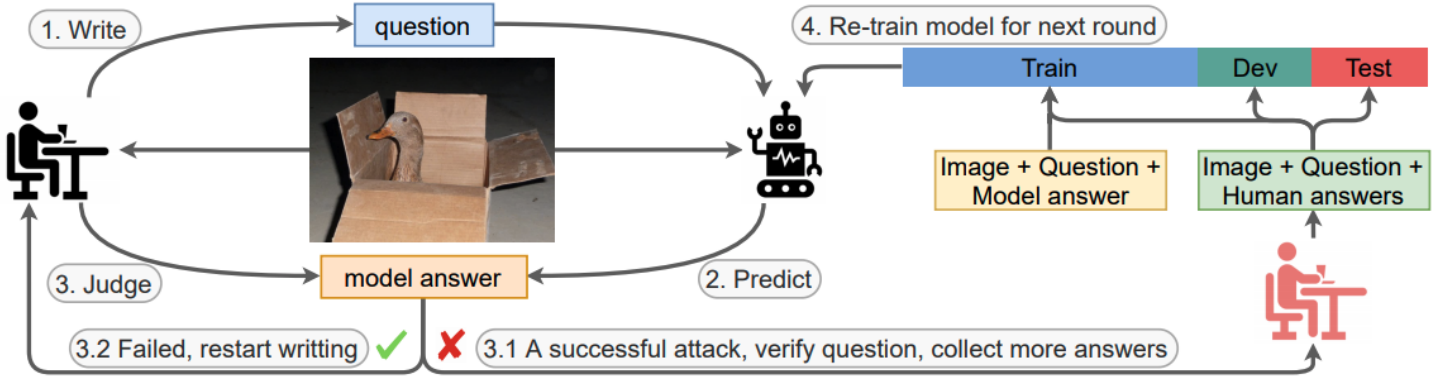
- Object Categorization
- Object Localization
- Segmentation
- Surface Normal Estimation
- Referring Expression Grounding
- Visual Question Answering
- Person Keypoint Detection

VLUE: A Multi-Task Benchmark for Evaluating VIsion-Language Pre-training

Task	Dataset	Image Domain	|Train|	|Dev|	|Test|	|OOD Test|	Metric
Image-Text Retrieval	MSCOCO	COCO	566,747	25,010	25,010	27,796	R@1
Image Captioning	MSCOCO	COCO	566,747	25,010	25,010	27,796	BLEU/CIDER
Visual Grounding	RefCOCO+	COCO	120,191	10,758	10,615	1,313	Accuracy
Visual Reasoning	NLVR2	Google Images[3]	86,373	6,982	6,967	5,662	Accuracy
Visual Question Answering	VQA 2.0	COCO	443,757	214,354	447,793	11,942	Accuracy

Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering
Roses Are Red, Violets Are Blue... but Should Vqa Expect Them To?
Rethinking Evaluation Practices in Visual Question Answering: A Case Study on Out-of-Distribution Generalization
GRIT: General Robust Image Task Benchmark
VLUE: A Multi-Task Benchmark for Evaluating Vision-Language Models

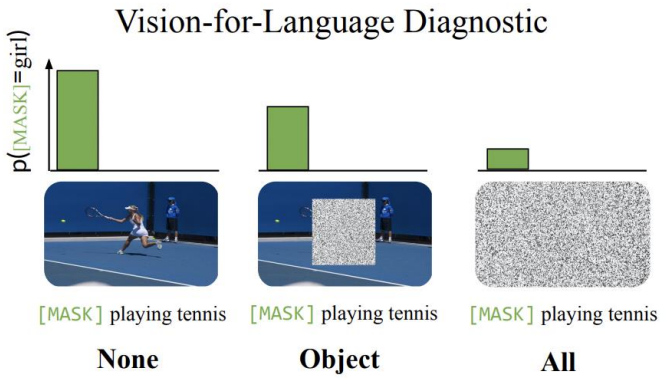
Adversarial VQA

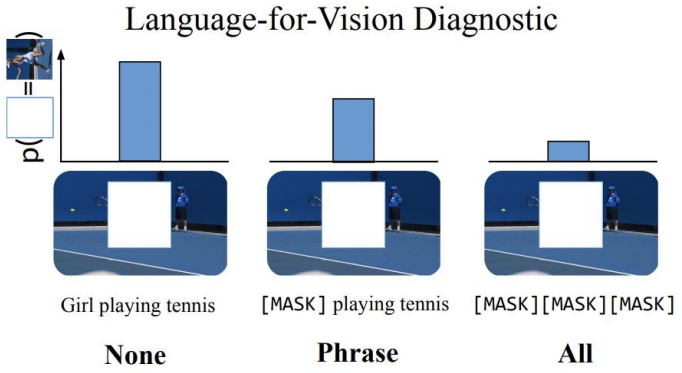


Round	Count	OCR	Reasoning				Visual Concept Recognition				
			Position	Relation	Common-sense	Other	Low-level	Action	Small Object	Occlusion	Abstract
R1	23.3%	10.7%	14.7%	8.3%	17.3%	0.7%	9.7%	4.3%	13.3%	14.7%	6.3%
R2	30.0%	22.7%	12.0%	27.7%	20.0%	4.3%	12.7%	9.3%	22.7%	10.0%	15.3%
R3	35.3%	13.0%	13.0%	28.3%	25.0%	6.3%	11.7%	4.3%	20.0%	20.0%	6.0%
Ave.	29.6%	15.4%	13.2%	21.4%	20.8%	3.8%	11.3%	6.0%	18.7%	14.9%	9.2%

Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models
Human-Adversarial Visual Question Answering

Probing (among many other works)





Cross-modal input ablation test

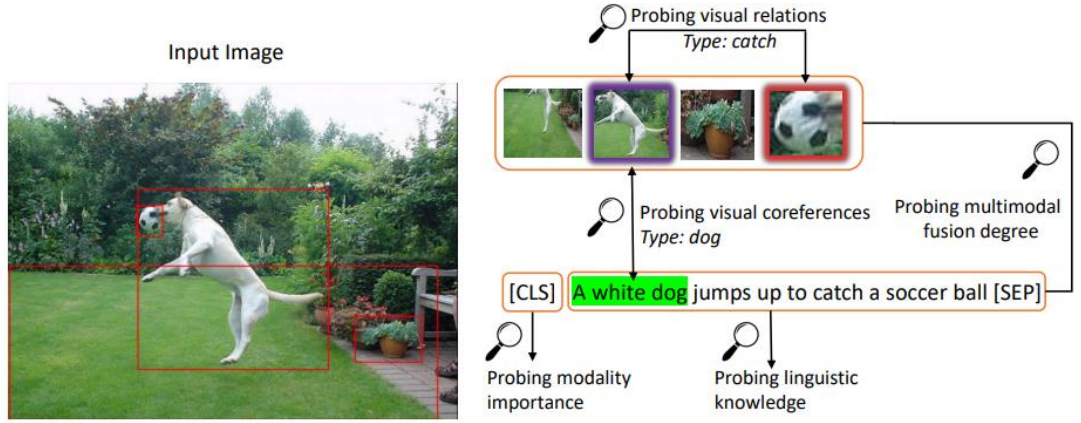
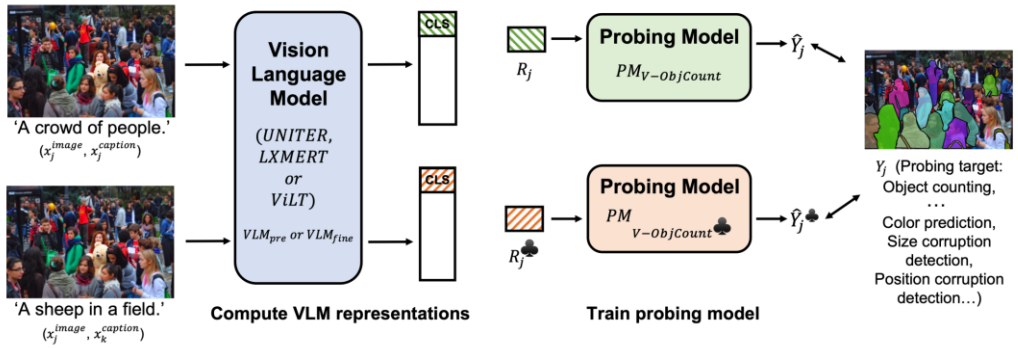


Fig. 1: Illustration of the proposed VALUE framework for investigating pre-trained vision-and-language models. VALUE consists of a set of well-designed probing tasks that unveil the inner mechanisms of V+L pre-trained models across: (i) Multimodal Fusion Degree; (ii) Modality Importance; (iii) Cross-modal Interaction via probing visual coreferences; (iv) Image-to-image Interaction via probing visual relations; and (v) Text-to-text Interaction via probing learned linguistic knowledge.



Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models

Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers

Are Vision-Language Transformers Learning Multimodal Representations? A Probing Perspective

Take-away messages

- What has been covered
  - Big foundation models: compared with big LMs, the development of big VL models is still in its infant stage
  - Few-shot in-context learning: you need a big LM at first
  - Model evaluation: we discussed diagnostic tests, OOD, adversarial examples, and probing analysis

- Future challenges
  - What's next beyond simple model scaling? Are larger models always better?
  - Beyond simple text output, can we train a MM model that quickly adapt to tasks that require bounding box output as well in a few-shot manner?
  - What do we mean when we say we improve VQA score by another +0.5 points? As MM foundation models become stronger and stronger, what's the next-generation benchmarking system?

Thank you!
Any Questions?