Early VLMs

Philipp Krähenbühl, UT Austin

Computer vision

"Traditional" Computer Vision
2014 - now

Task 1

Model 1

Output

Task 2

Model 2

Output

...

Unified Computer Vision
2020 - now

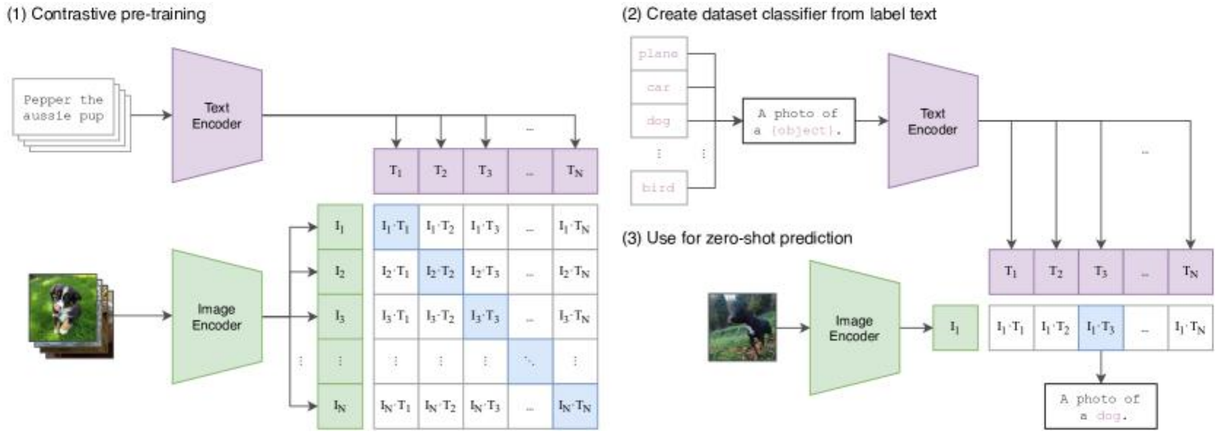
Task

Unified Model

Output

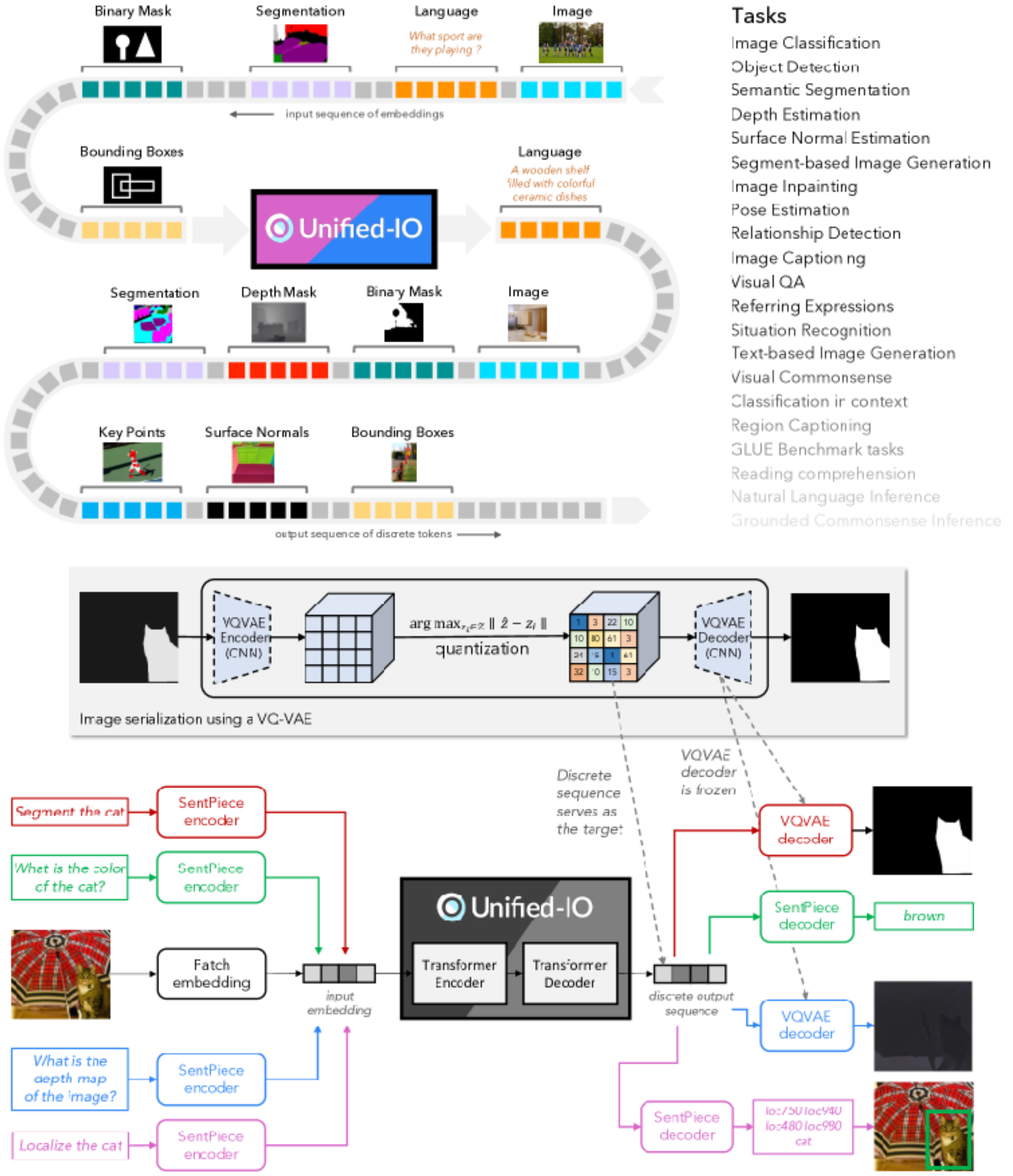
CLIP as a VLM

- Clip maps
  - Images to text
  - Text to images
- Primitive image and text models
- No dialogue



Unified-IO

- Tokenize everything
  - Image, Text, Sparse (few numbers), Dense (depth maps etc)
  - Image, Dense: VQ-GAN
  - Text: SentencePiece
  - Sparse: 1000 special tokens (coordinates)



Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks, Lu etal 2022

Unified-IO

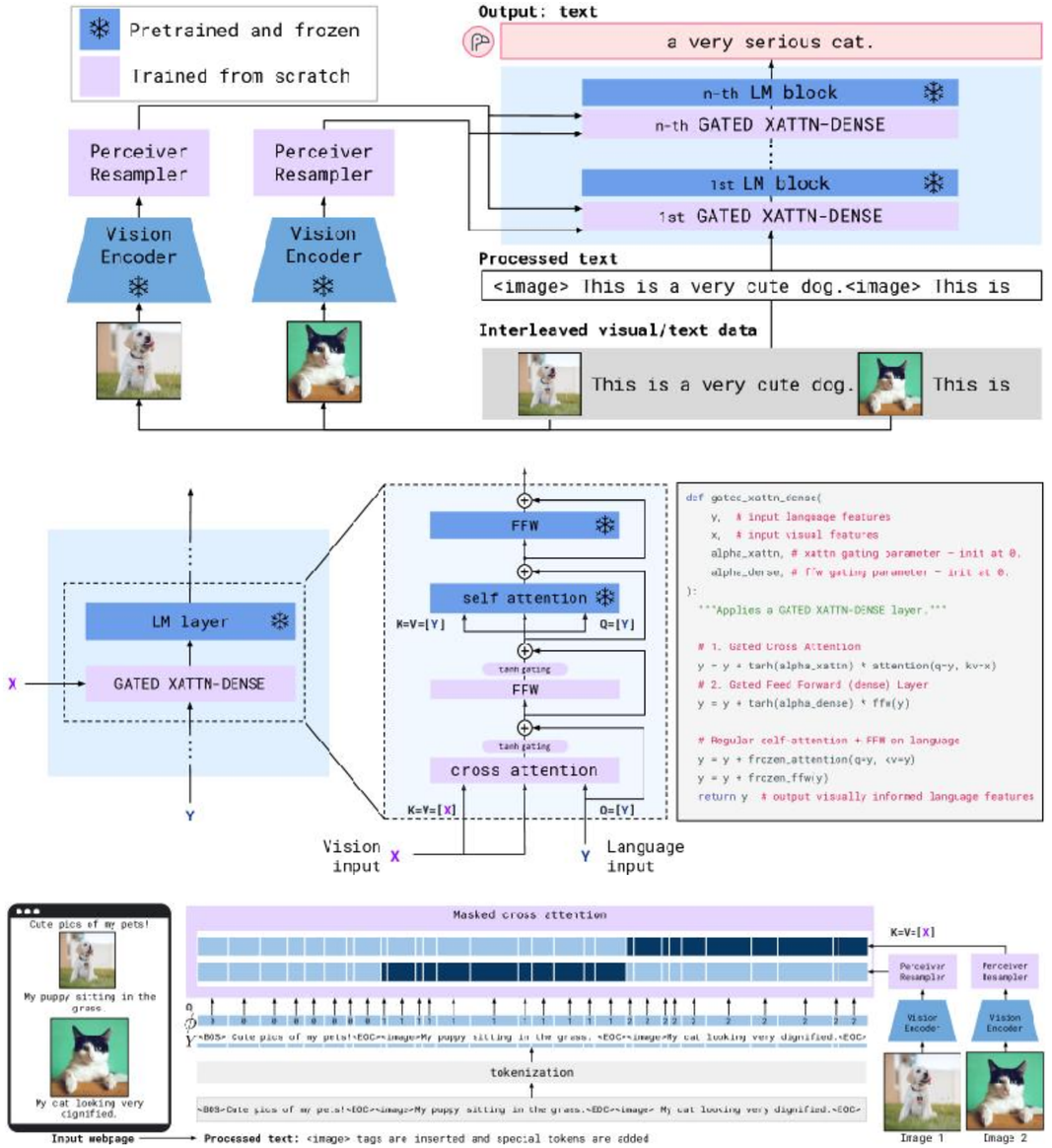
- Everything is sequence prediction (with 1D or 2D positional embeddings)
- Unsupervised pre-training of vision and language inputs (masked prediction)
- Fine-tune (train) on 95 unified datasets

	Example Source	Size				Input Modalities				Output Modalities			
		Datasets	Size	Percent	Rate	Text	Image	Sparse	Dense	Text	Image	Sparse	Dense
Image Synthesis		**14**	**56m**	**43.0**	**18.7**	✓	✓	✓	✓	-	✓	-	-
Image Synthesis from Text	*RedCaps*	9	55m	41.9	16.7	✓	-	-	-	-	✓	-	-
Image Inpainting	*VG*	3	1.2m	0.9	1.5	✓	✓	✓	-	-	✓	-	-
Image Synthesis from Seg.	*LVIS*	2	220k	0.2	0.6	✓	-	-	✓	-	✓	-	-
Sparse Labelling		**10**	**8.2m**	**6.3**	**12.5**	✓	✓	✓	-	-	-	✓	-
Object Detection	*Open Images*	3	1.9m	1.5	3.6	-	✓	-	-	-	-	✓	-
Object Localization	*VG*	3	6m	4.6	7.1	✓	✓	-	-	-	-	✓	-
Keypoint Estimation	*COCO*	1	140k	0.1	0.7	-	✓	✓	-	-	-	✓	-
Referring Expression	*RefCoco*	3	130k	0.1	1.1	✓	✓	-	-	-	-	✓	-
Dense Labelling		**6**	**2.4m**	**1.8**	**6.2**	✓	✓	-	-	-	-	-	✓
Depth Estimation	*NYU Depth*	1	48k	0.1	0.4	-	✓	-	-	-	-	-	✓
Surface Normal Estimation	*Framenet*	2	210k	0.2	1.1	-	✓	-	-	-	-	-	✓
Object Segmentation	*LVIS*	3	2.1m	1.6	4.7	✓	✓	-	-	-	-	-	✓
Image Classification		**9**	**22m**	**16.8**	**12.5**	-	✓	✓	-	✓	-	-	-
Image Classification	*ImageNet*	6	16m	12.2	8.1	✓	✓	-	-	✓	-	-	-
Object Categorization	*COCO*	3	6m	4.6	4.4	-	✓	✓	-	✓	-	-	-
Image Captioning		**7**	**31m**	**23.7**	**12.5**	-	✓	✓	-	✓	-	-	-
Webly Supervised Captioning	*CC12M*	3	26m	19.7	8.8	-	✓	-	-	✓	-	-	-
Supervised Captioning	*VizWiz*	3	1.4m	1.1	1.7	-	✓	-	-	✓	-	-	-
Region Captioning	*VG*	1	3.8m	2.9	2.0	-	✓	✓	-	✓	-	-	-
Vision & Language		**16**	**4m**	**3.0**	**12.5**	✓	✓	✓	-	✓	-	-	✓
Visual Question Answering	*VQA 2.0*	13	3.3m	2.5	10.1	✓	✓	✓	-	✓	-	-	-
Relationship Detection	*VG*	2	640k	0.5	1.9	-	✓	✓	-	✓	-	-	-
Grounded VQA	*VizWiz*	1	6.5k	0.1	0.1	✓	✓	-	-	✓	-	-	✓
NLP		**31**	**7.1m**	**5.4**	**12.5**	✓	-	-	-	✓	-	-	-
Text Classification	*MNLI*	17	1.6m	1.2	4.8	✓	-	-	-	✓	-	-	-
Question Answering	*SQuAD*	13	1.7m	1.3	5.2	✓	-	-	-	✓	-	-	-
Text Summarization	*Gigaword*	1	3.8m	2.9	2.5	✓	-	-	-	✓	-	-	-
Language Modelling		**2**	-	-	**12.5**	✓	-	-	-	✓	-	-	-
Masked Language Modelling	*C4*	2	-	-	12.5	✓	-	-	-	✓	-	-	-
All Tasks		**95**	**130m**	**100**	**100**	✓	✓	✓	✓	✓	✓	✓	✓

Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks, Lu etal 2022

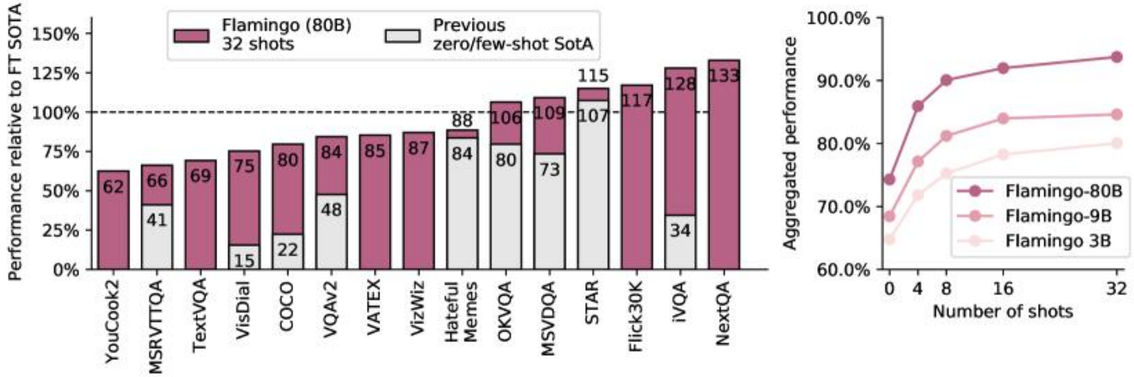
Flamingo

- One of first big VLM
- Based on Chinchilla
  - Frozen vision model spliced in (Gated X-Attn)
- Multi-image support
  - Masked attention only to preceding image



Flamingo: a Visual Language Model for Few-Shot Learning, Alayrac etal 2022

Flamingo

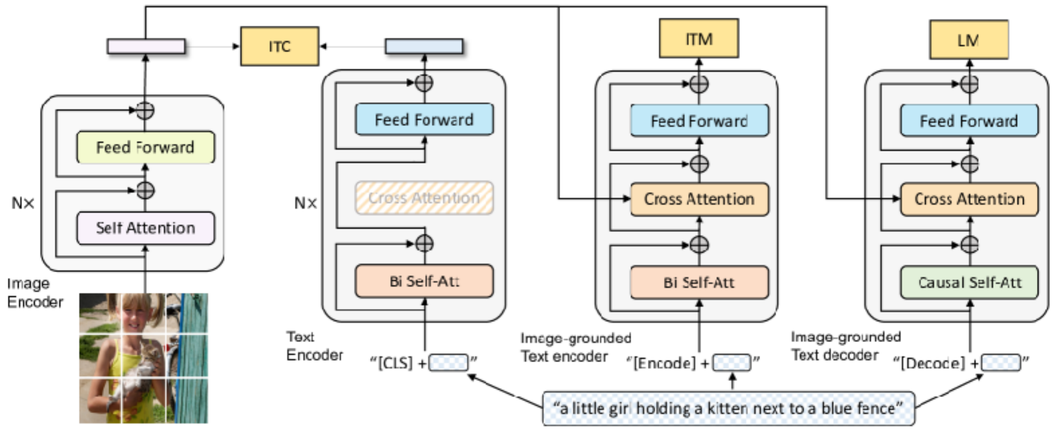


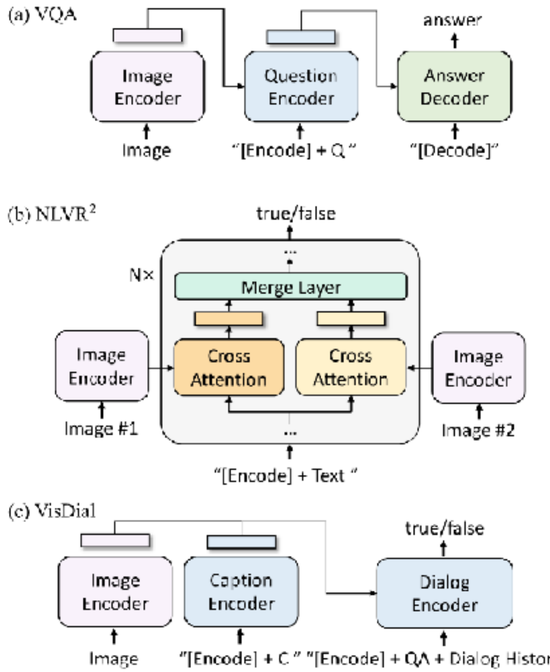
- Started zero-shot eval trend
- Good results on VQA
  - Under VQA metrics

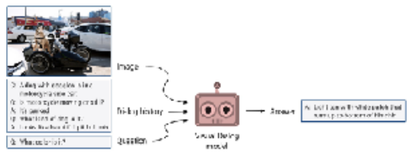
Flamingo: a Visual Language Model for Few-Shot Learning, Alayrac etal 2022

BLIP

- One network for 3 tasks
  - CLIP-style embeddings (ITC)
  - Image-language matching (ITM)
  - Language modeling / Captioning (LM)
- Different structure, shared weights (except SA)



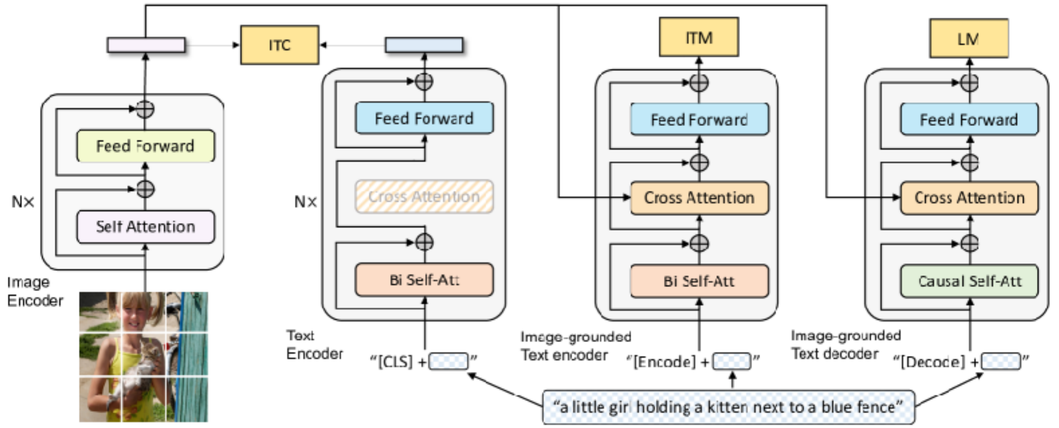


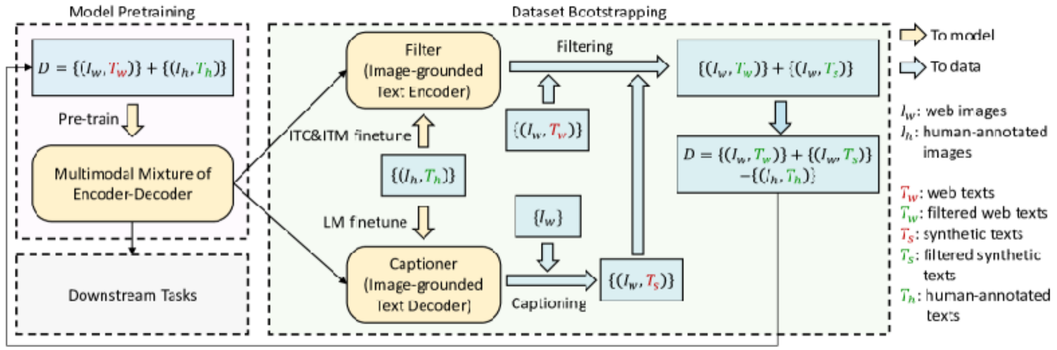


BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, Li etal 2022

BLIP

- Pre-Training
  - Image captioning data (COCO etal)
- CapFilt
  - Use ITM, LM fine-tuned separately on COCO
  - Clean up web-scale image-text data
    - Create new captions (LM)
    - Filter new + original captions (ITM)





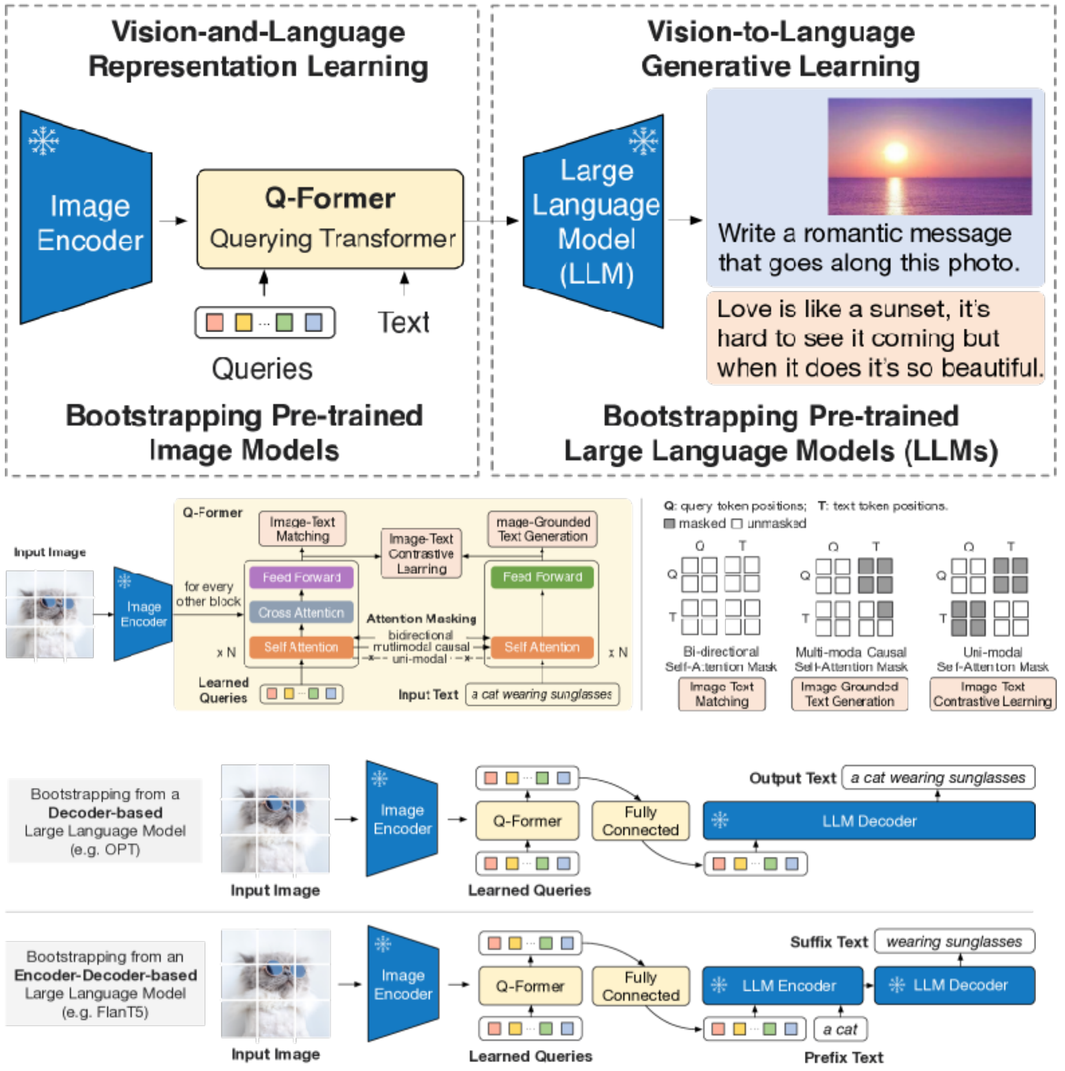
Method	Pre-train #Images	NoCaps validation in-domain C	in-domain S	near-domain C	near-domain S	out-domain C	out-domain S	overall C	overall S	COCO Caption Karpathy test B@4	COCO Caption Karpathy test C
Enc-Dec (Changpinyo et al., 2021)	15M	92.6	12.5	88.3	12.1	94.5	11.9	90.2	12.1	-	110.9
VinVL† (Zhang et al., 2021)	5.7M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
$\text{LEMON}_{\text{base}}$† (Hu et al., 2021)	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
$\text{LEMON}_{\text{base}}$† (Hu et al., 2021)	200M	107.7	14.7	106.2	14.3	107.9	13.1	106.8	14.1	**40.3**	**133.3**
BLIP	14M	111.3	15.1	104.5	14.4	102.4	13.7	105.1	14.4	38.6	129.7
BLIP	129M	109.1	14.8	105.8	14.4	105.7	13.7	106.3	14.3	39.4	131.4
$\text{BLIP}_{\text{CapFilt-L}}$	129M	**111.8**	**14.9**	**108.6**	**14.8**	**111.5**	**14.2**	**109.6**	**14.7**	39.7	**133.3**
$\text{LEMON}_{\text{large}}$† (Hu et al., 2021)	200M	116.9	15.8	113.3	15.1	111.3	14.0	113.4	15.0	40.6	135.7
$\text{SimVLM}_{\text{huge}}$ (Wang et al., 2021)	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
$\text{BLIP}_{\text{ViT-L}}$	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7

Table 7. Comparison with state-of-the-art image captioning methods on NoCaps and COCO Caption. All methods optimize the cross-entropy loss during finetuning. C: CIDEr, S: SPICE, B@4: BLEU@4. $\text{BLIP}_{\text{CapFilt-L}}$ is pre-trained on a dataset bootstrapped by captioner and filter with ViT-L. VinVL† and LEMON† require an object detector pre-trained on 2.5M images with human-annotated bounding boxes and high resolution (800×1333) input images. $\text{SimVLM}_{\text{huge}}$ uses 13× more training data and a larger vision backbone than ViT-L.

BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, Li etal 2022

BLIP v2

- Captioning -> Language modeling
- Q-Former
  - Image-text embedder / tokenizer
  - Fed as context to LLM

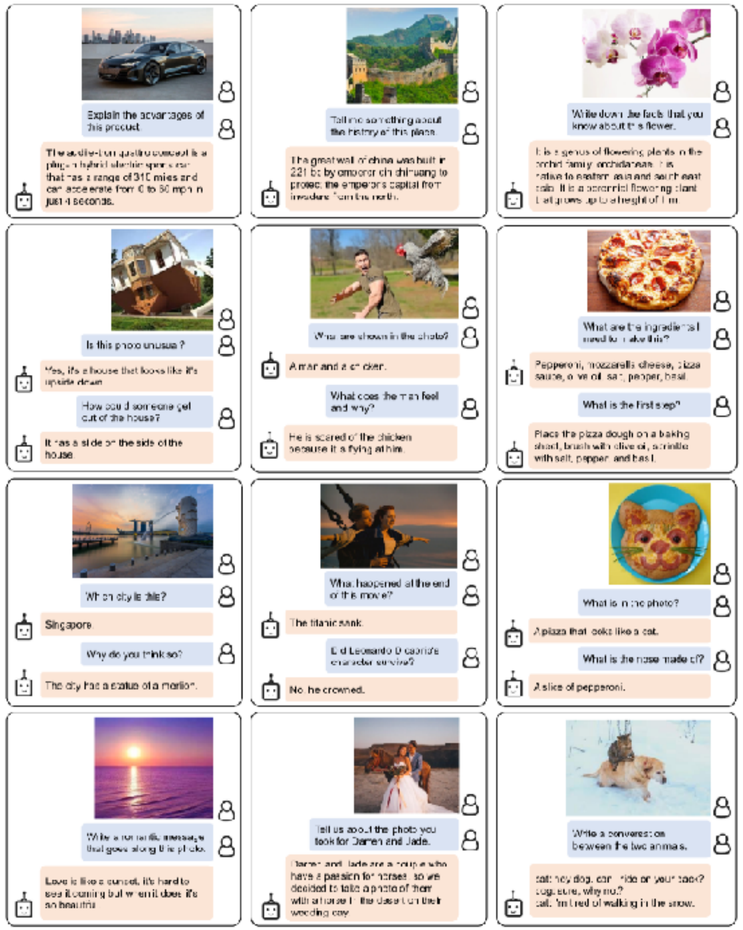


BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, Li etal 2023

BLIP v2

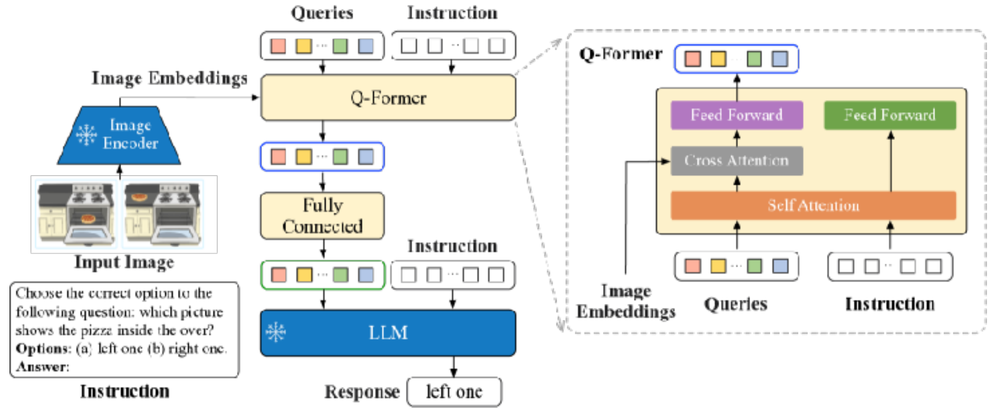
- Great results
  - Captioning
  - VQA
  - Retrieval
- Early visual dialogues

Models	#Trainable Params	Open-sourced?	Visual Question Answering VQAv2 (test-dev) VQA acc.	Image Captioning NoCaps (val) CIDEr	SPICE	Image-Text Retrieval 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	✓	**65.0**	**121.6**	**15.8**	**97.6**	**89.7**

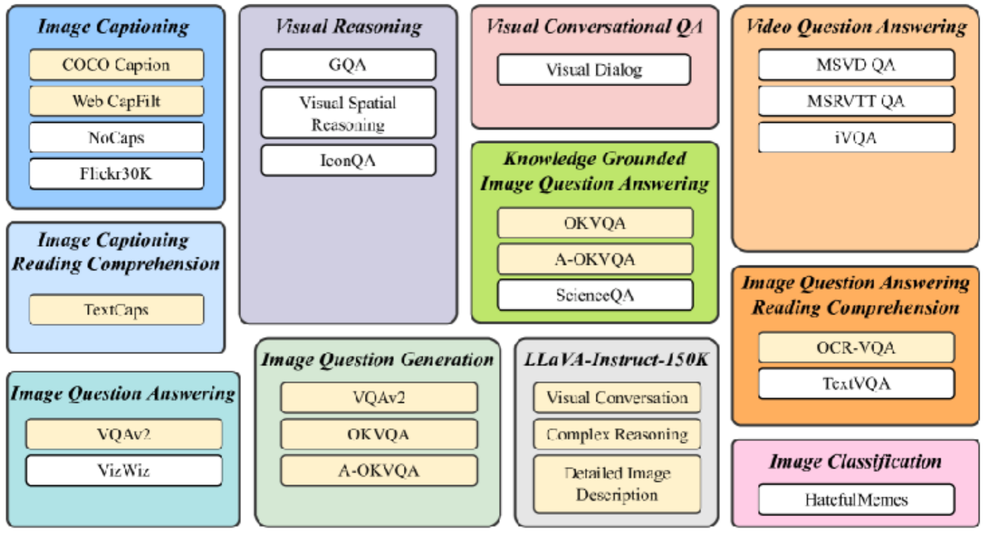


BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, Li etal 2023

InstructBLIP



- BLIP v2 architecture
- Instruction tuning for visual reasoning benchmarks
  - 26 datasets (13 train)
  - *Meticulously* crafted instruction templates
- Good results



InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, Dai etal 2023

InstructBLIP

- BLIP v2 architecture
- Instruction tuning for visual reasoning benchmarks
  - 26 datasets (13 train)
  - *Meticulously* crafted instruction templates
- Good results

Task	Instruction Template
Image Captioning	<Image>A short image caption: <Image>A short image description: <Image>A photo of <Image>An image that shows <Image>Write a short description for the image. <Image>Write a description for the photo. <Image>Provide a description of what is presented in the photo. <Image>Briefly describe the content of the image. <Image>Can you briefly explain what you see in the image? <Image>Could you use a few words to describe what you perceive in the photo? <Image>Please provide a short depiction of the picture. <Image>Using language, provide a short account of the image. <Image>Use a few words to illustrate what is happening in the picture.
VQA	<Image>{Question} <Image>Question: {Question} <Image>{Question} A short answer to the question is <Image>Q: {Question} A: <Image>Question: {Question} Short answer: <Image>Given the image, answer the following question with no more than three words. {Question} <Image>Based on the image, respond to this question with a short answer: {Question}. Answer: <Image>Use the provided image to answer the question: {Question} Provide your answer as short as possible: <Image>What is the answer to the following question? "{Question}" <Image>The question "{Question}" can be answered using the image. A short answer is
VQG	<Image>Given the image, generate a question whose answer is: {Answer}. Question: <Image>Based on the image, provide a question with the answer: {Answer}. Question: <Image>Given the visual representation, create a question for which the answer is "{Answer}". <Image>From the image provided, craft a question that leads to the reply: {Answer}. Question: <Image>Considering the picture, come up with a question where the answer is: {Answer}. <Image>Taking the image into account, generate an question that has the answer: {Answer}. Question:

Table 5: Instruction templates used for transforming held-in datasets into instruction tuning data. For datasets with OCR tokens, we simply add "OCR tokens:" after the image query embeddings.

GQA, VizWiz, IVQA, MSVD, MSRVTT <Image> Question: {} Short answer:

NoCaps, Flickr30k <Image> A short image description:

TextVQA <Image> OCR tokens: {}. Question: {} Short answer:

IconQA <Image> Question: {} Options: {}. Short answer:

ScienceQA <Image> Context: {} Question: {} Options: {}. Answer:

HatefulMemes <Image> This is an image with: "{}" written on it. Is it hateful? Answer:

VSR <Image> Based on the image, is this statement true or false? "{}" Answer:

Visual Dialog <Image> Dialog history: {}\n Question: {} Short answer:

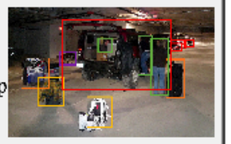
	ScienceQA IMG	OCR-VQA	OKVQA	A-OKVQA Direct Answer Val	A-OKVQA Direct Answer Test	A-OKVQA Multi-choice Val	A-OKVQA Multi-choice Test
Previous SOTA	LLaVA [25] 89.0	GIT [43] 70.3	PaLM-E(562B) [9] **66.1**	[15] 56.3	[37] 61.6	[15] 73.2	[37] 73.6
BLIP-2 (FlanT5$_{XXL}$)	89.5	72.7	54.7	57.6	53.7	80.2	76.2
InstructBLIP (FlanT5$_{XXL}$)	**90.7**	**73.3**	55.5	57.1	54.8	**81.0**	**76.7**
BLIP-2 (Vicuna-7B)	77.3	69.1	59.3	60.0	58.7	72.1	69.0
InstructBLIP (Vicuna-7B)	79.5	72.8	62.1	**64.0**	**62.1**	75.7	73.4

InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, Dai etal 2023

LLAVA v1

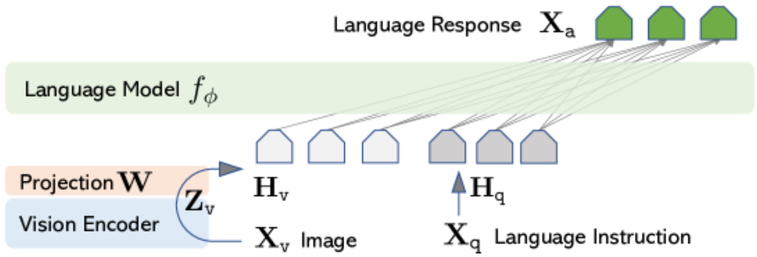
- Instruction tuning (visual dialogue) data is scarce
  - Let's use GPT4 (not GPT4-V) to create more data
  - Subset of CC3M (<600k)
  - Learn a single projection W on top of CLIP-VIT into Vicuna (llama2)

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], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
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. ...<omitted>
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 ...<omitted>



Visual Instruction Tuning, Liu etal 2023

LLAVA v1

- Stage 1 training: Adapter only
  - CC3M subset
  - GPT-4 ask question for caption
- Stage 2 training: Adapter + LLM
  - GPT-4 data (ask GPT-4 to do QA using caption and box inputs)
  - ScienceQA
- Quite easy to train (<1 day on 8 GPUs)

Visual Instruction Tuning, Liu etal 2023

```
messages = [ {"role":"system", "content": f"""You are an AI visual assistant, and you are seeing a single image. What you see are provided with five sentences, describing the same image you are looking at. Answer all questions as you are seeing the image.

Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the image, including the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Only include questions that have definite answers:
(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently.

Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary."""}
]
for sample in fewshot_samples:
    messages.append({"role":"user", "content":sample['context']})
    messages.append({"role":"assistant", "content":sample['response']} )
messages.append({"role":"user", "content":'\n'.join(query)})
```

Table 13: For each query, we illustrate the prompt construction process for ChatGPT/GPT 4 to collect `query['response']` from `query['context']`, using few-shot in-context-learning, where examples are from `fewshot_samples`, each example including input `sample['context']` and output `sample['response']`. Note that `messages` is the final prompt. In this example, we provide the prompt used to generate the conversation response, please see also see its in-context-learning examples in Table 15 and Table 16 for details. We recommend readers to check out the codebase for the prompts to generated two other types of responses, including detailed decription and complex reasoning.

Captions
There is a movie theater that displays the show times above the doors
A red fire hydrant is deep in the snow.
The fire hydrant is in the snow near a recently plowed sidewalk.
This city has had a very hard winter with snow.
A hotel for dogs in the snow in winter.

Conversation
Question:
What color is the fire hydrant in the image?
Answer:
The fire hydrant in the image is red.
Question:
Is the fire hydrant fully covered by snow or partially visible?
Answer:
The fire hydrant is partially visible, with a significant portion of it covered by the deep snow.
Question:
What is the condition of the sidewalk near the fire hydrant?
Answer:
The sidewalk near the fire hydrant has been recently plowed, so it is relatively clear of snow compared to the surrounding area.
Question:
What does the image suggest about the winter season in the city?
Answer:
The image suggests that the winter season in the city has been quite harsh and challenging. The significant snow accumulation seen in the image indicates that the city has experienced heavy snowfall, which can lead to various difficulties and disruptions for its residents.

For instance, the deep snow covering the fire hydrant could potentially hinder its accessibility in case of emergencies, which poses safety concerns. The recently plowed sidewalk implies that the city's maintenance crews have been actively working to keep the walkways clear and safe for pedestrians, but the sheer amount of snowfall might make it difficult to keep up with the continuous snow removal efforts.

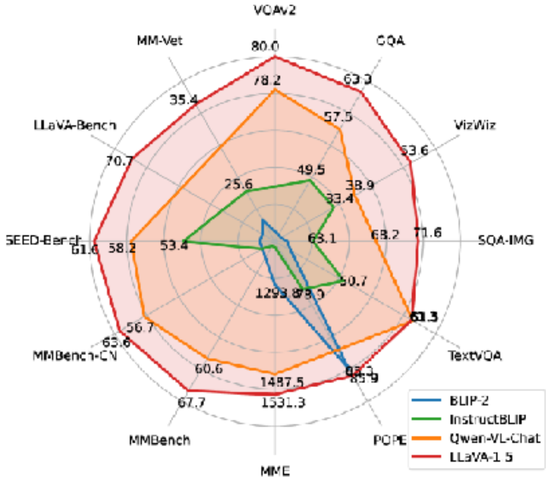
Furthermore, such extreme winter conditions can impact transportation, leading to traffic issues, delays in public transit, and increased risks of accidents due to slippery roads. It can also cause problems for businesses and schools, as people might face challenges commuting to work or attending classes. Additionally, the heavy snow can put extra strain on infrastructure, such as roofs and power lines, increasing the likelihood of structural damage or power outages.

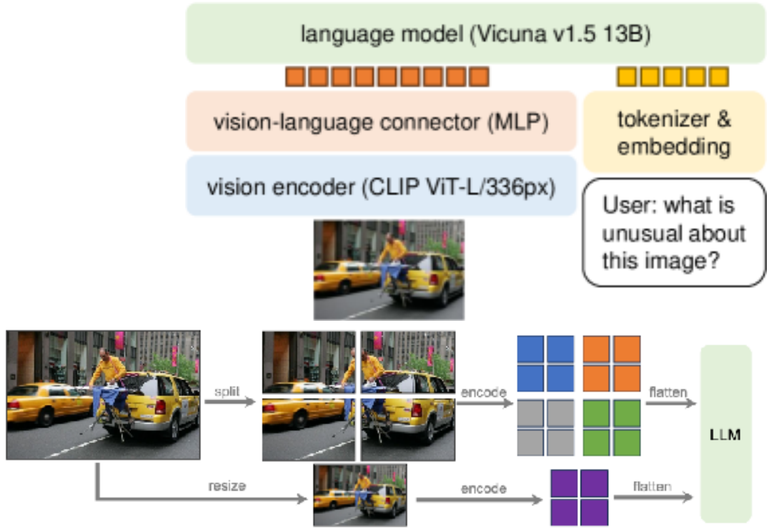
In conclusion, the image of the red fire hydrant deep in the snow and the recently plowed sidewalk suggests that the city has faced a particularly severe winter season, with substantial snowfall that has likely caused various challenges and disruptions for its residents and infrastructure.

Llava 1.5

- Better (higher resolution) image encoder
- Larger LLM
- More datasets
- Still quite fast to train
  - <1 day on 8-GPUs

	Method	LLM	Res.	GQA	MME	MM-Vet
	InstructBLIP	14B	224	49.5	1212.8	25.6
	Only using a subset of InstructBLIP training data					
0	LLaVA	7B	224	–	809.6	25.5
1	+VQA-v2	7B	224	47.0	1197.0	27.7
2	+Format prompt	7B	224	46.8	1323.8	26.3
3	+MLP VL connector	7B	224	47.3	1355.2	27.8
4	+OKVQA/OCR	7B	224	50.0	1377.6	29.6
	Additional scaling					
5	+Region-level VQA	7B	224	50.3	1426.5	30.8
6	+Scale up resolution	7B	336	51.4	1450	30.3
7	+GQA	7B	336	62.0*	1469.2	30.7
8	+ShareGPT	7B	336	62.0*	1510.7	31.1
9	+Scale up LLM	13B	336	**63.3***	**1531.3**	**36.1**





Improved Baselines with Visual Instruction Tuning, Liu etal 2023

LLAVA 1.6

AKA LLAVA-NEXT

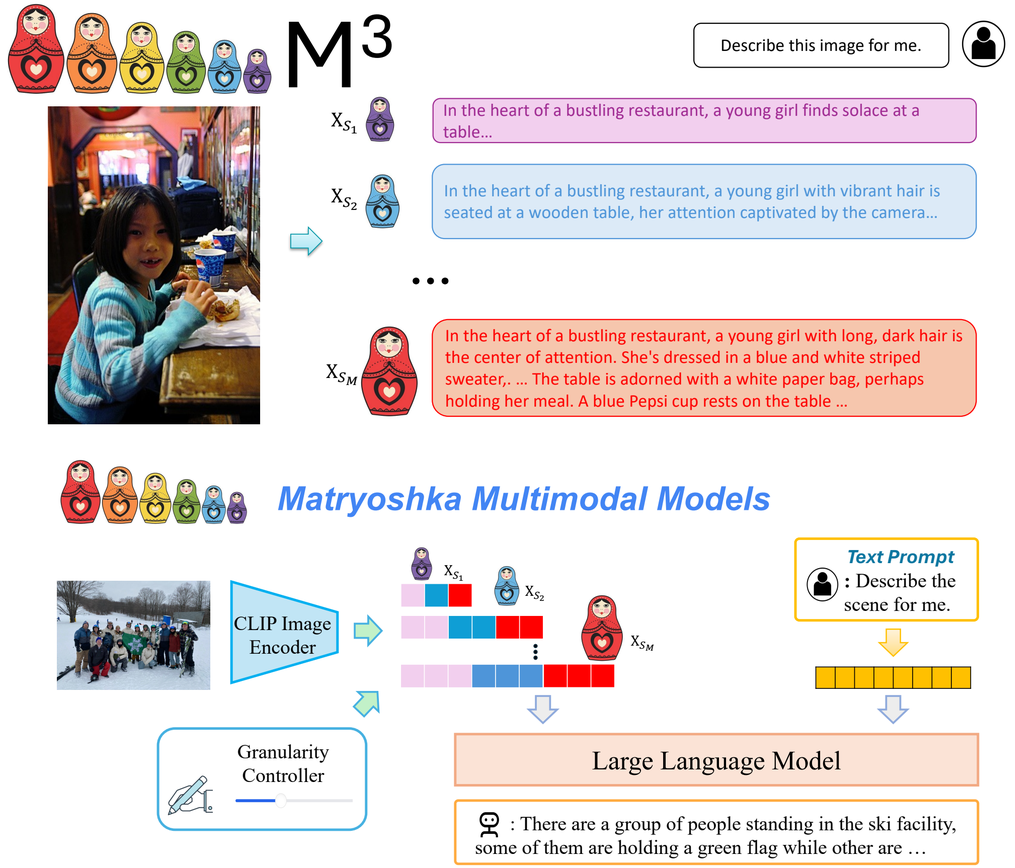
- Higher resolution, More data (OCR), Better model, full model fine-tuning, still fast (1 day with 32 A100s)

Open-Source Proprietary

Data (PT)	Data (IT)	Model	MMMU (val)	Math-Vista	MMB-ENG	MMB-CN	MM-Vet	LLaVA-Wild	SEED-IMG
N/A	N/A	GPT-4V	56.8	49.9	75.8	73.9	67.6	-	71.6
N/A	N/A	Gemini Ultra	59.4	53	-	-	-	-	-
N/A	N/A	Gemini Pro	47.9	45.2	73.6	74.3	64.3	-	70.7
1.4B	50M	Qwen-VL-Plus	45.2	43.3	-	-	55.7	-	65.7
1.5B	5.12M	CogVLM-30B	32.1	-	-	-	56.8	-	-
125M	~1M	Yi-VL-34B	45.9	-	-	-	-	-	-
558K	665K	LLaVA-1.5-13B	36.4	27.6	67.8	63.3	36.3	72.5	68.2
558K	760K	LLaVA-NeXT-34B	51.1	46.5	79.3	79	57.4	89.6	75.9

M3

- LLaVA at multiple granularities
  - Granularity Controller = 2x2 average pooling
- Good results



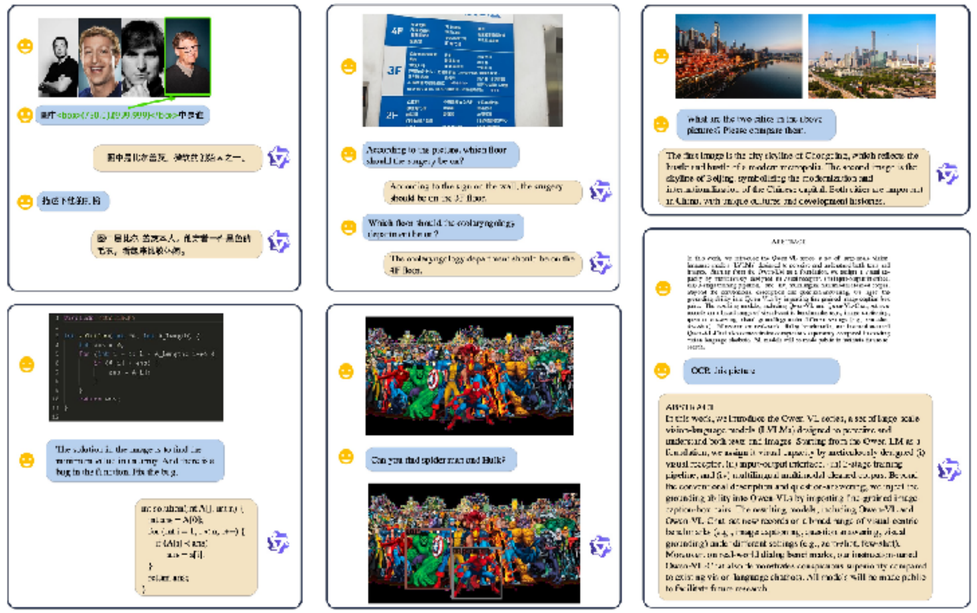
Approach	# Tokens	MMBench	GQA	POPE	VizWiz	SEEDBench
Qwen-VL [7]	256	38.2	59.3	-	35.2	56.3
Qwen-VL-Chat [7]	256	60.6	57.5	-	38.9	58.2
InstructBLIP-7B [58]	32	36.0	49.2	-	34.5	53.4
InstructBLIP-13B [58]	32	-	49.5	78.9	33.4	-
LLaVA-1.5-7B [5]	576	64.8	**62.0**	85.9	54.4	60.5
LLaVA-1.5-M^3	576	65.9	61.9	**87.4**	**54.9**	**60.6**
	144	**66.4**	61.3	87.0	53.1	59.7
	36	64.8	60.3	85.5	52.8	58.0
	9	63.1	58.0	83.4	51.9	55.4
	1	59.5	52.6	78.4	49.4	50.1

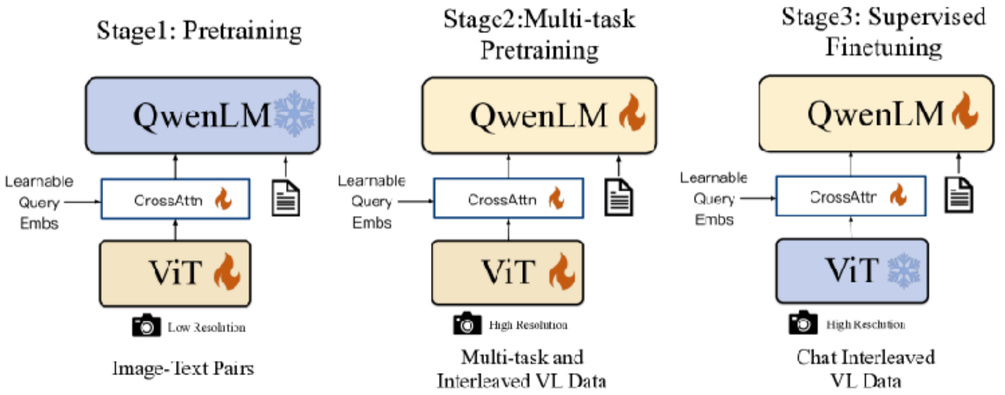
Approach	# Tokens	MSVD	MSRVTT	ActivityNet	NextQA	IntentQA	EgoSchema
Video-LLaMA [60]	-	51.6	29.6	12.4	-	-	-
LLaMA-Adapter [61]	-	54.9	43.8	34.2	-	-	-
Video-ChatGPT [62]	-	64.9	49.3	35.2	-	-	-
Video-LLaVA [63]	2048	70.7	59.2	45.3	-	-	-
InternVideo [64]	-	-	-	-	59.1	-	32.1
LLaVA-NeXT-7B [4]	2880	78.8	63.7	54.3	**63.1**	**60.3**	35.8
LLaVA-NeXT-7B-M^3	2880	78.2	**64.5**	53.9	**63.1**	58.8	36.8
	720	**79.0**	**64.5**	**55.0**	62.6	59.6	37.2
	180	77.9	63.7	**55.0**	61.4	59.3	37.6
	45	75.8	63.0	53.2	59.5	58.7	**38.8**
	5	73.5	62.7	50.8	56.5	56.7	36.2

Matryoshka Multimodal Models, Cai etal 2024

Qwen-VL

- Vision encoder (1.9B)
- Adapter (0.8B)
- LLM (7.7B)
- Inputs: Image, Bounding Box, Text
- Outputs: Bounding Box, Text
- Bounding box: Text format with special <box> and <ref> tokens





Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, Bai etal 2023

Qwen-VL

- Paper has lots of details on dataset setup and training recipes (base models etc)
- Great results for the time (Fall 2023)

Stage 1:

Language	Dataset	Original	Cleaned	Remaining%
English	LAION-en	2B	280M	14%
	LAION-COCO	600M	300M	50%
	DataComp	1.4B	300M	21%
	Coyo	700M	200M	28%
	CC12M	12M	8M	66%
	CC3M	3M	3M	100%
	SBU	1M	0.8M	80%
	COCO Caption	0.6M	0.6M	100%
Chinese	LAION-zh	108M	105M	97%
	In-house Data	220M	220M	100%
	Total	5B	1.4B	28%

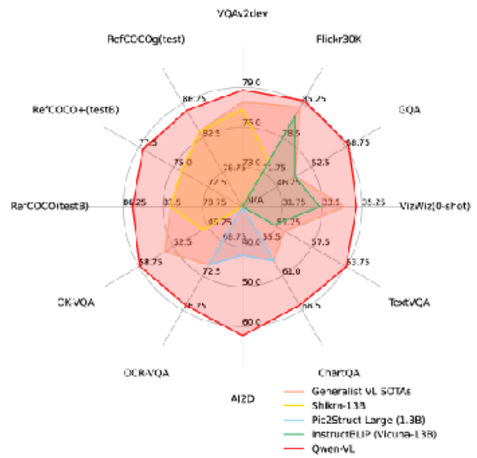
Stage 2:

Task	# Samples	Dataset
Captioning	19.7M	LAION-en & zh, DataComp, Coyo, CC12M & 3M, SBU, COCO, In-house Data
VQA	3.6M	GQA, VGQA, VQAv2, DVQA, OCR-VQA, DocVQA, TextVQA, ChartQA, AI2D
Grounding[2]	3.5M	GRIT
Ref Grounding	8.7M	GRIT, Visual Genome, RefCOCO, RefCOCO+, RefCOCOg
Grounded Cap.	8.7M	GRIT, Visual Genome, RefCOCO, RefCOCO+, RefCOCOg
OCR	24.8M	SynthDoG-en & zh, Common Crawl pdf & HTML
Pure-text Autoregression	7.8M	In-house Data

Stage 3:

The Dataset Format Example of ChatML

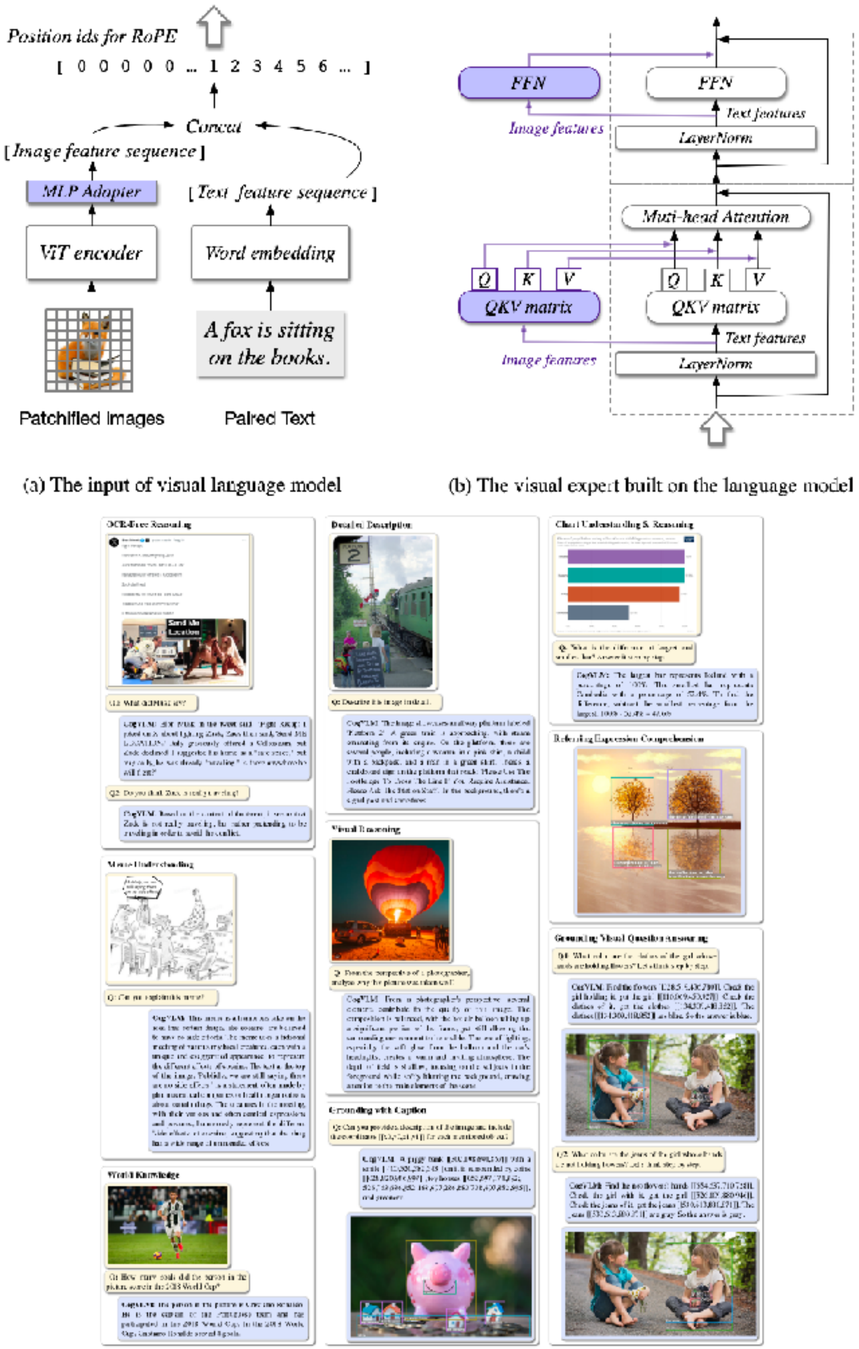
```
<im_start>user
Picture 1: <img>vg/VG_100K_2/649.jpg</img>What is the sign in the picture?<im_end>
<im_start>assistant
The sign is a road closure with an orange rhombus.<im_end>
<im_start>user
How is the weather in the picture?<im_end>
<im_start>assistant
The shape of the road closure sign is an orange rhombus.<im_end>
```



Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, Bai etal 2023

CogVLM

- VLM with Flamingo-style fusion (but on parallel not gated sequence)
- Stage 1 - Pre-training: Captioning 1.5B images (LAION-2B, COYO-700M filtered)
- Stage 2 - Pre-training: Add in Referring Expression Comprehension (same data, boxes form open-vocab detector)



CogVLM: Visual Expert for Pretrained Language Models, Wang etal 2023

CogVLM

- Alignment
  - Chat-style data (following LLaVa)
  - Grounded Captioning, Referring Expression (generation and comprehension), grounded VQA

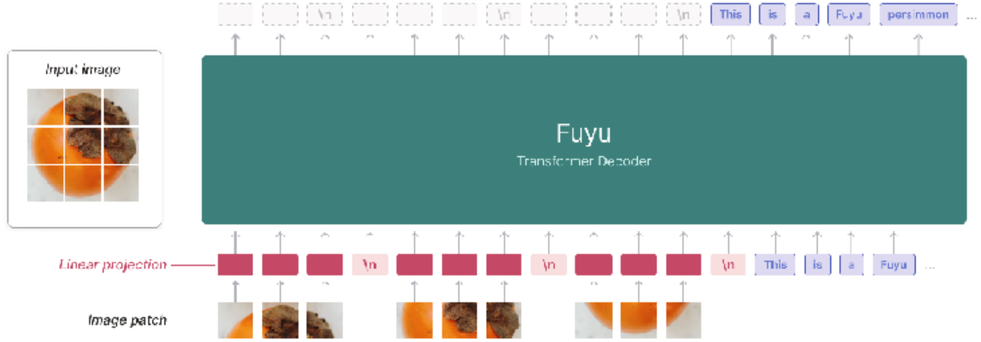
Method	Train Data	NoCaps val		NoCaps test		Flickr	COCO	TextCaps
		OOD	overall	OOD	overall	Karp.	Karp.	test
Human	-	95.7	87.1	91.6	85.3	-	-	125.1
VinVL (Zhang et al., 2021)	8.9M	83.8	94.3	78.0	92.5	-	130.8	-
SimVLM (Wang et al., 2021)	1.8B	115.2	112.2	109.5	110.3	-	143.3	-
CoCa (Yu et al., 2022)	4.8B	-	122.4	-	120.6	-	143.6	-
LEMON (Hu et al., 2022)	2B	120.2	117.3	110.1	114.3	-	139.1	-
Flamingo (Alayrac et al., 2022)	2.3B	-	-	-	-	67.2	138.1	-
Prismer (Liu et al., 2023d)	12.7M	113.5	112.9	-	110.8	-	136.5	-
BLIP-2 (Li et al., 2023b)	129M	124.8	121.6	-	-	-	144.5	-
InstructBLIP (Dai et al., 2023)	129M	-	123.1	-	-	82.4	-	-
UniversalCap (Cornia et al., 2021)	35M	123.4	122.1	114.3	119.3	-	143.4	-
GIT (Wang et al., 2022a)	0.8B	127.1	125.5	122.0	123.4	49.6	144.8	138.2
GIT2 (Wang et al., 2022a)	12.9B	130.6	126.9	122.3	124.8	50.7	145.0	145.0
Qwen-VL (Bai et al., 2023)	1.4B	-	121.4	-	-	85.8	-	-
PaLI-17B (Chen et al., 2022b)	1.6B	-	127.0	-	124.4	-	149.1	135.4
PaLI-X-55B (Chen et al., 2023b)	-	-	126.3	-	124.3	-	**149.2**	**147.0**
CogVLM (ours)	1.5B	**132.6**	**128.3**	**128.0**	**126.4**	**94.9**	148.7	144.9

Method	LLM	VQA					LVLM-Benchmark						
		VQAv2	OKVQA	TextVQA	OCRVQA	ScienceQA	MM-Vet	SEED	MMBench	LLaVA	POPE	MMMU	MathVista
MiniGPT-4	Vicuna-7B	-	-	0.6	-	39.6	22.1	47.4	23.0	45.1	-	-	23.1
IDEFICS-Instruct	LLaMA-65B	37.4	36.9	32.9	-	61.8	39.7	53.2	54.5	56.9	-	-	26.2
OpenFlamingo	MPT-7B	53.0	38.3	28.3	-	44.8	24.8	42.7	5.7	34.2	-	26.3	18.6
DreamLLM	Vicuna-7B	56.6	44.3	34.9	-	-	35.9	-	49.9	-	-	-	-
InstructBLIP	Vicuna-7B	-	-	50.1	-	60.5	26.2	58.8	33.9	59.8	53.8	-	25.3
Fuyu	Fuyu-8B	74.2*	60.6*	-	-	-	-	-	-	-	-	27.4	-
Qwen-VL-Chat	Qwen-7B	78.2*	56.6*	61.5*	70.5*	68.8	-	65.4	61.8	67.7	-	32.9	33.8
LLaVA-1.5	Vicuna-7B	78.5*	-	58.2*	-	66.8	30.5	58.6	64.3	60.7	85.9	-	23.6
mPLUG-Owl2	LLaMA2-7B	79.4*	57.7*	58.2*	-	68.7	36.2	64.1	64.5	25.0	86.2	32.1	25.3
Unified-IO2	UIO-2XXL	79.4*	55.5*	-	-	86.2*	-	65.6	71.5	-	87.7	-	-
LLaVA-1.5	Vicuna-13B	80.0*	-	61.3*	-	71.6	35.4	61.6	67.7	64.6	85.9	33.6	26.1
SPHINX-2k	LLaMA2 13B	80.7*	62.6*	61.2*	67.8*	70.6	40.2	71.6	65.9	-	87.2	32.9	27.8
Emu2-Chat	LLaMA-33B	**84.9***	64.8*	66.6*	-	-	48.5	62.8	63.6	56.4	-	34.1	-
CogVLM-Chat	Vicuna-7B	82.3*	**64.8***	**70.4***	**73.8***	**91.2***	**51.1**	**72.5**	**77.6**	**77.8**	**87.9**	**41.1**	**34.5**

Type	Model	RefCOCO			RefCOCO+			RefCOCOg		Visual7W
		val	test-A	test-B	val	test-A	test-B	val	test	test
Generalist	OFA-L* (Wang et al., 2022b)	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58	-
	VisionLLM-H (Wang et al., 2023b)	-	86.70	-	-	-	-	-	-	-
	Shikra-7B (Chen et al., 2023a)	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19	-
	Shikra-13B (Chen et al., 2023a)	87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16	85.33
	Qwen-VL (Bai et al., 2023)	89.36	92.26	85.34	83.12	88.25	77.21	85.58	85.48	-
	Ferret-13B (You et al., 2023)	89.48	92.41	84.36	82.81	88.14	75.17	85.83	86.34	-
	CogVLM-Grounding	**92.76**	**94.75**	**88.99**	**88.68**	**92.91**	**83.39**	**89.75**	**90.79**	**91.05**
Specialist	G-DINO-L (Liu et al., 2023e)	90.56	93.19	88.24	82.75	88.95	75.92	86.13	87.02	-
	UNINEXT-H (Lin et al., 2023a)	92.64	94.33	91.46	85.24	89.63	79.79	88.73	89.37	-
	ONE-PEACE (Wang et al., 2023a)	92.58	94.18	89.26	88.77	92.21	83.23	89.22	89.27	-

CogVLM: Visual Expert for Pretrained Language Models, Wang etal 2023

OtterHD



- Built on Perssimon-8B and Fuyu-8B (decoder only transformers)
  - Images fed in tokenized
  - Image new-line token
- Dynamic image resolution
- Instruction-tuned

Models	I/R Pairs	Train Res.	Eval Res.	MagBench Multi.	MagBench FF.	MME[1] Cog.	MME[1] Percep.	POPE	MM-V	MMB	M-Vista
Idefics-9B$_{\text{instruct}}$ [24]	1M	224	224	20.8	13.4	187.9	1165.0	74.6	23.7	45.5	19.8
Otter-9B [25]	150K	224	224	25.7	15.8	306.4	1292.3	72.5	24.7	48.3	19.7
InstructBLIP-7B [13]	1.2M	224	224	5.6	15.2	-	-	-	26.2	36.0	-
InstructBLIP-13B [13]	1.2M	224	224	3.8	16.3	291.8	1212.8	78.9	25.6	33.9	**25.3**
LLaVA-7B$_{1.5}$ [30]	3.6M[2]	336	336	26.8	24.7	-	**1510.7**	85.9	**30.5**	59.5	-
Qwen-VL-7B$_{\text{chat}}$ [4]	1.4B	448	448	14.5	15.9	**360.7**	1487.5	-	-	**61.8**	-
Fuyu-8B [5]	-	-	*Original*	29.3	15.2	237.5	728.6	74.1	21.4	10.7	20.6
OtterHD-8B	370K	512	512	33.5	31.4	289.8	1359.3	**86.1**	25.1	58.5	22.3
		1024	1024	37.8	37.2	288.5	1313.7	81.5	19.8	53.6	17.3
		Dynamic	*Original*	**42.7**	**39.9**	331.4	1223.4	86.0	26.3	58.3	23.5

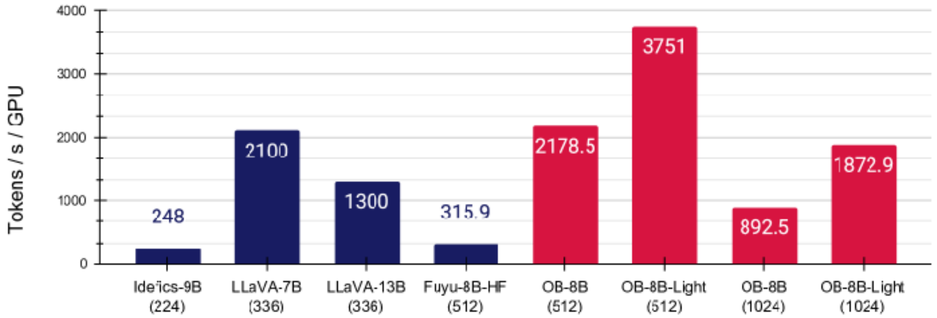
Dataset	LLaVA-DD/CR	VQAv2	GQA	OKVQA	OCRVQA	A-OKVQA	COCO-GOI
Avg. W	577	581	495	617	352	587	586
Avg. H	481	482	409	448	490	482	476
Pairs	53240	20000	30000	18018	16354	34112	20000
Dataset	**COCO-Caption**	**TextQA**	**RefCOCO**	**COCO-ITM**	**ImageNet**	**LLaVA-RLHF**	**Combined**
Avg. W	578	950	591	577	469	340	542
Avg. H	484	811	486	484	387	572	467
Pairs	20000	19293	20000	20000	50000	50000	371017

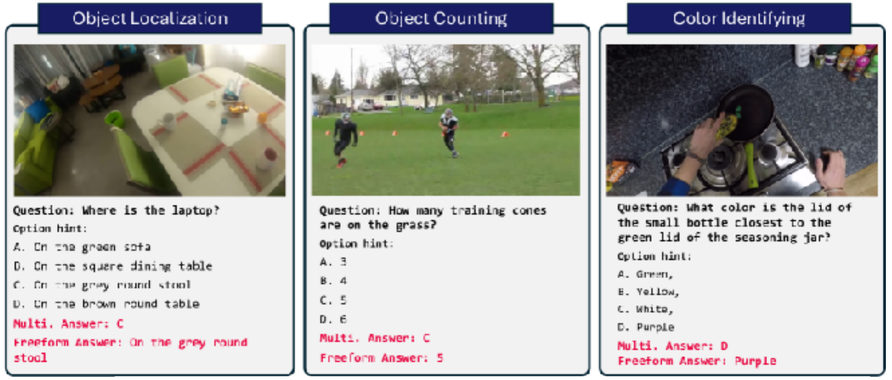
OtterHD: A High-Resolution Multi-modality Model, Li etal 2023

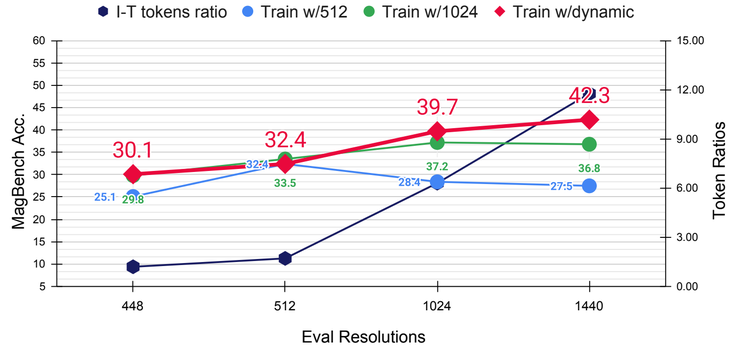
OtterHD

- More data
- Faster base-model
- New High-res VLM benchmark (MagnifierBench)

Data Mixture We compiled a total of 370K instruction/response pairs sourced from the following public datasets: LLaVA-Instruct [30], VQAv2 [2], GQA [23], OKVQA [36], OCRVQA [38], A-OKVQA [45], COCO-GOI [33], COCO-Caption [10], TextQA [48], RefCOCO [58], COCO-ITM [28], ImageNet [17], and LLaVA-RLHF [51]. The data mixture and specific prompt strategies are motivated by `LLaVA-1.5` [30] and `Idefics-Instruct` [24] to achieve better text formatting control. All the datasets were organized into instruction/response pairs, aggregated into a single dataloader and uniformly sampled during the training phase to ensure representational integrity.



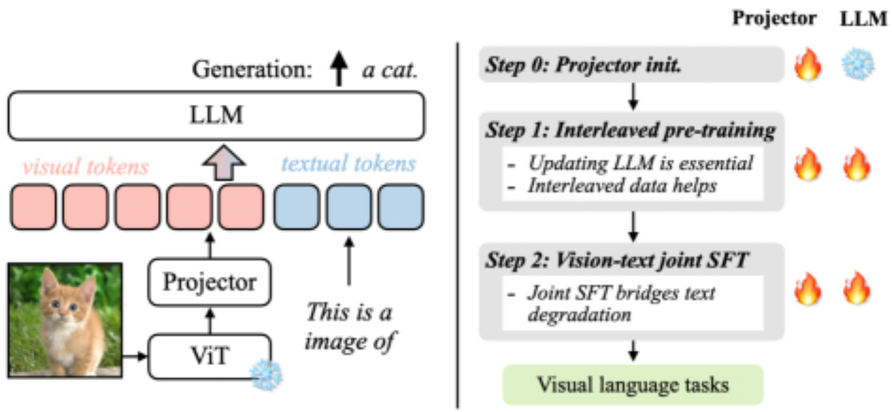




OtterHD: A High-Resolution Multi-modality Model, Li etal 2023

VILA

- Lot's of tricks and ablations
  - How to interleave multiple image-text pairs
  - Blending text-only and image-text data

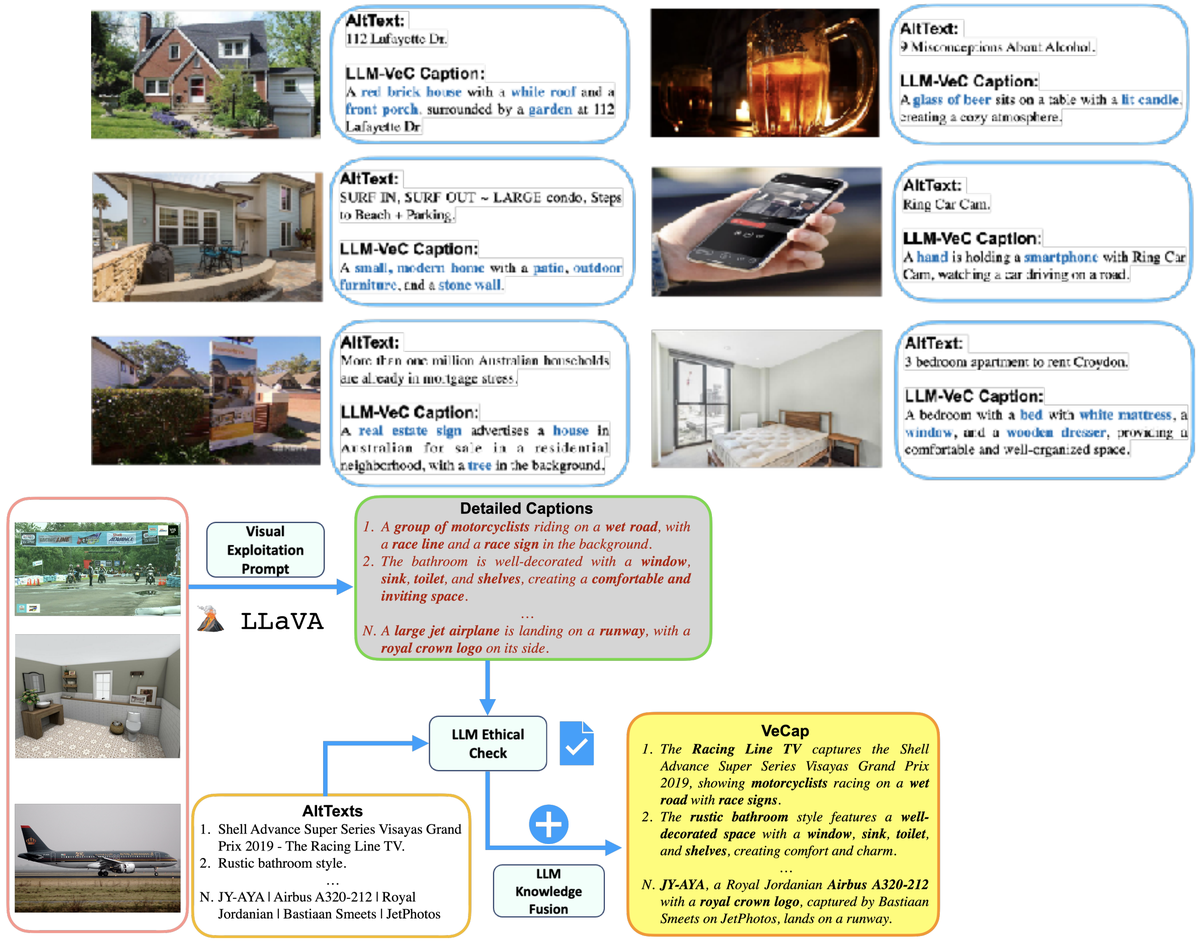


Method	LLM	Res.	PT	IT	VQA^{v2}	GQA	VisWiz	SQA^{I}	VQA^{T}	POPE	MME	MMB	MMB^{CN}	SEED	LLaVA^{W}	MM-Vet
BLIP-2 [35]	Vicuna-13B	224	129M	-	41.0	41	19.6	61	42.5	85.3	1293.8	–	–	46.4	38.1	22.4
InstructBLIP [18]	Vicuna-7B	224	129M	1.2M	–	49.2	34.5	60.5	50.1	–	–	36	23.7	53.4	60.9	26.2
InstructBLIP [18]	Vicuna-13B	224	129M	1.2M	–	49.5	33.4	63.1	50.7	78.9	1212.8	–	–	–	58.2	25.6
Shikra [12]	Vicuna-13B	224	600K	5.5M	77.4*	–	–	–	–	–	–	58.8	–	–	–	–
IDEFICS-9B [30]	LLaMA-7B	224	353M	1M	50.9	38.4	35.5	–	25.9	–	–	48.2	25.2	–	–	–
IDEFICS-80B [30]	LLaMA-65B	224	353M	1M	60.0	45.2	36.0	–	30.9	–	–	54.5	38.1	–	–	–
Qwen-VL [9]	Qwen-7B	448	1.4B	50M	78.8*	59.3*	35.2	67.1	63.8	–	–	38.2	7.4	56.3	–	–
Qwen-VL-Chat [9]	Qwen-7B	448	1.4B	50M	78.2*	57.5*	38.9	68.2	61.5	–	1487.5	60.6	56.7	58.2	–	–
LLaVA-1.5 [38]	Vicuna-1.5-7B	336	0.6M	0.7M	78.5*	62.0*	50.0	66.8	58.2	**85.9**	1510.7	64.3	58.3	58.6	63.4	30.5
LLaVA-1.5 [38]	Vicuna-1.5-13B	336	0.6M	0.7M	80.0*	**63.3***	53.6	71.6	61.3	**85.9**	1531.3	67.7	63.6	61.6	70.7	35.4
VILA-7B (ours)	Llama-2-7B	336	50M	1M	79.9*	62.3*	57.8	68.2	64.4	85.5	1533.0	68.9	61.7	61.1	69.7	34.9
VILA-13B (ours)	Llama-2-13B	336	50M	1M	**80.8***	**63.3***	**60.6**	**73.7**	**66.6**	84.2	**1570.1**	**70.3**	**64.3**	**62.8**	**73.0**	**38.8**
+ShareGPT4V	Llama-2-13B	336	50M	1M	80.6*	63.2*	62.4	73.1	65.3	84.8	1556.5	70.8	65.4	61.4	78.4	45.7

VILA: On Pre-training for Visual Language Models, Lin etal 2023

VeCLIP

- Better captioning pipeline for VLMs and dataset (VeCap)
  - Ethics check (does the LLM reply: “I am sorry that I cannot …”)
  - LLM Knowledge Fusion: Rephrase the following two sentences into one…
- Better CLIP model

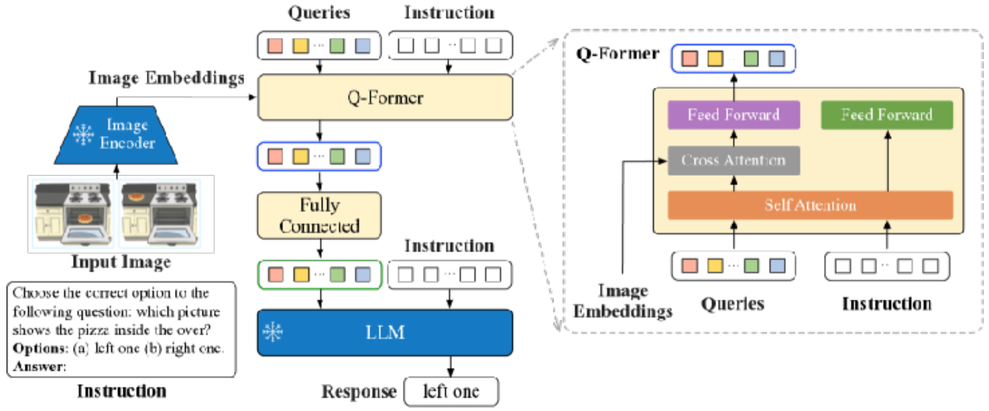


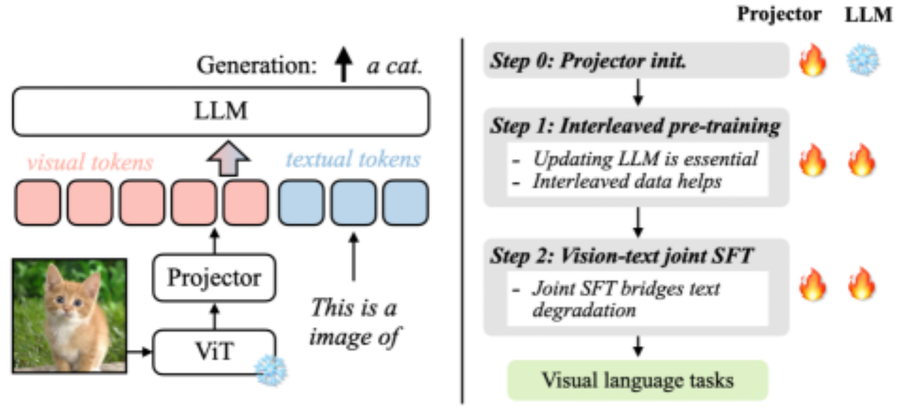
VeCLIP: Improving CLIP Training via Visual-enriched Captions, Lai etal 2023

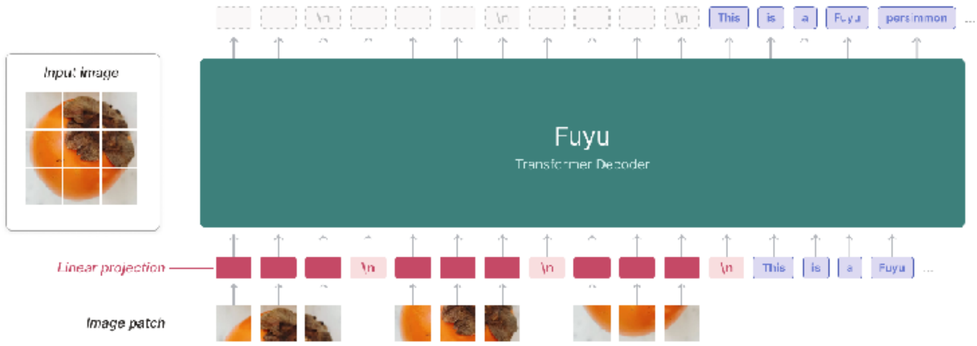
Vision language models

Trend 1

- **Architectures stabilized**
  - Image -> Sequence (vision tokens)
  - Test -> Sequence (text tokens)
  - Transformer







Vision language models

Trend 2

- **Data matters**
  - Innovations in training and data yield large improvements

References

- Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks, Lu etal. 2022
- Flamingo: a Visual Language Model for Few-Shot Learning, Alayrac etal. 2022
- BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, Li etal. 2022
- BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, Li etal. 2023
- InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, Dai etal. 2023
- Visual Instruction Tuning, Liu etal. 2023
- Improved Baselines with Visual Instruction Tuning, Liu etal. 2023
- Matryoshka Multimodal Models, Cai etal. 2024
- Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, Bai etal 2023
- CogVLM: Visual Expert for Pretrained Language Models, Wang etal. 2023
- OtterHD: A High-Resolution Multi-modality Model, Li etal. 2023
- VILA: On Pre-training for Visual Language Models, Lin etal. 2023
- VeCLIP: Improving CLIP Training via Visual-enriched Captions, Lai etal. 2023