Some Recent Image-Text Models

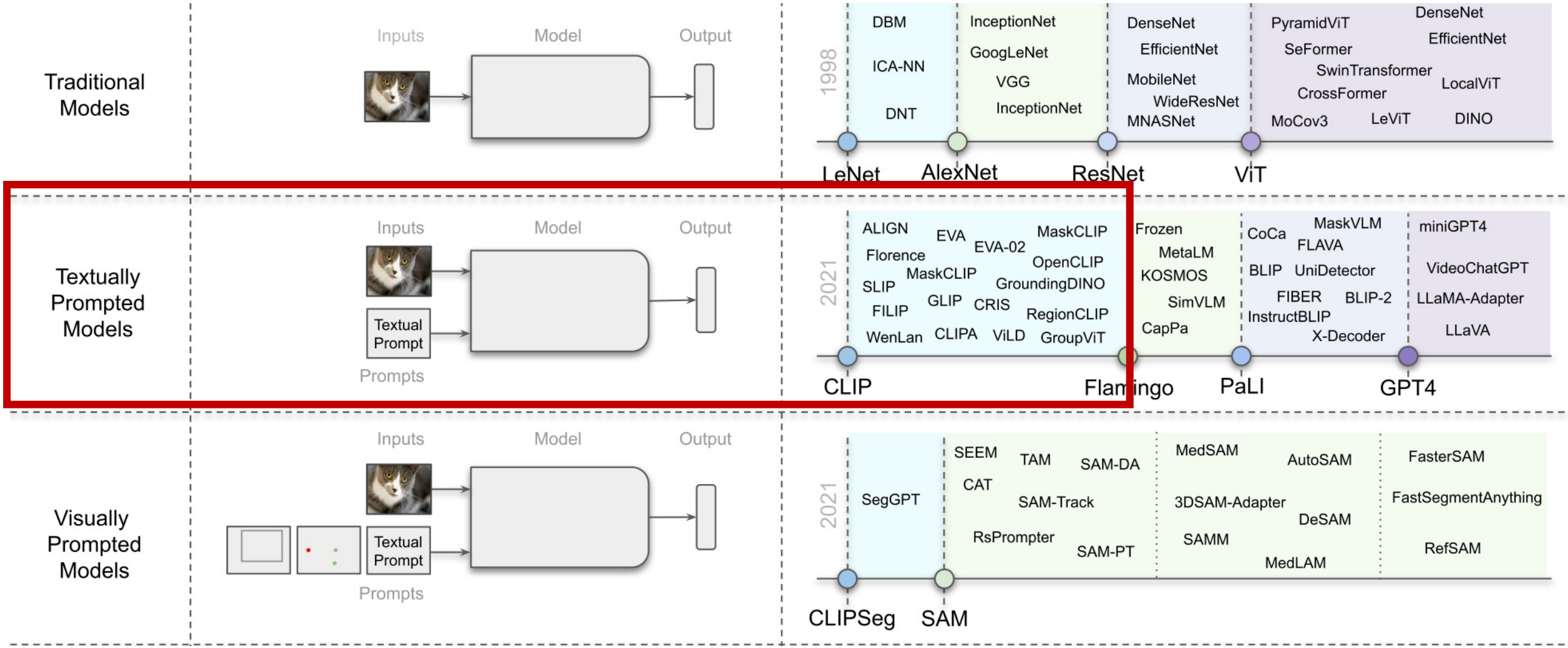
M. Soleymani

Sharif University of Technology

Spring 2026

Multi-modal data

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



Awais et al., “Foundational Models Defining a New Era in Vision: A Survey and Outlook”, 2023

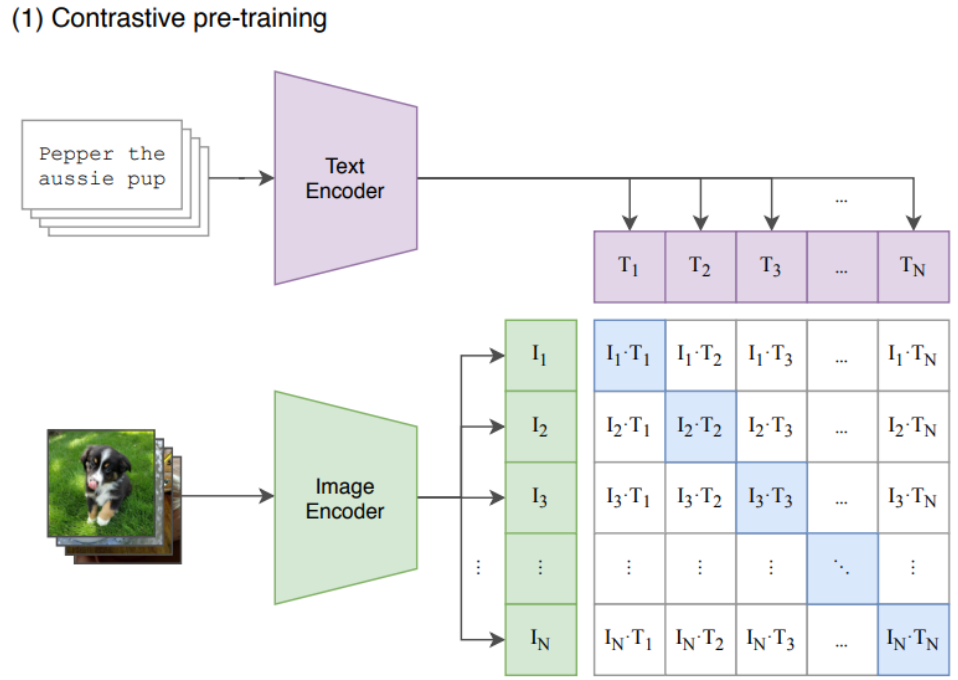
Vision-language models: Contrastive learning

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

A. Radford, J.W. Kim, et al., “Learning Transferable Visual Models From Natural Language Supervision”, 2021

CLIP (OpenAI)

- Caption as a weak supervision
- Learns a multi-modal embedding space by jointly training an image encoder and text encoder
  - maximize the cosine similarity of the image and text embeddings of the N real pairs in the batch while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairs.

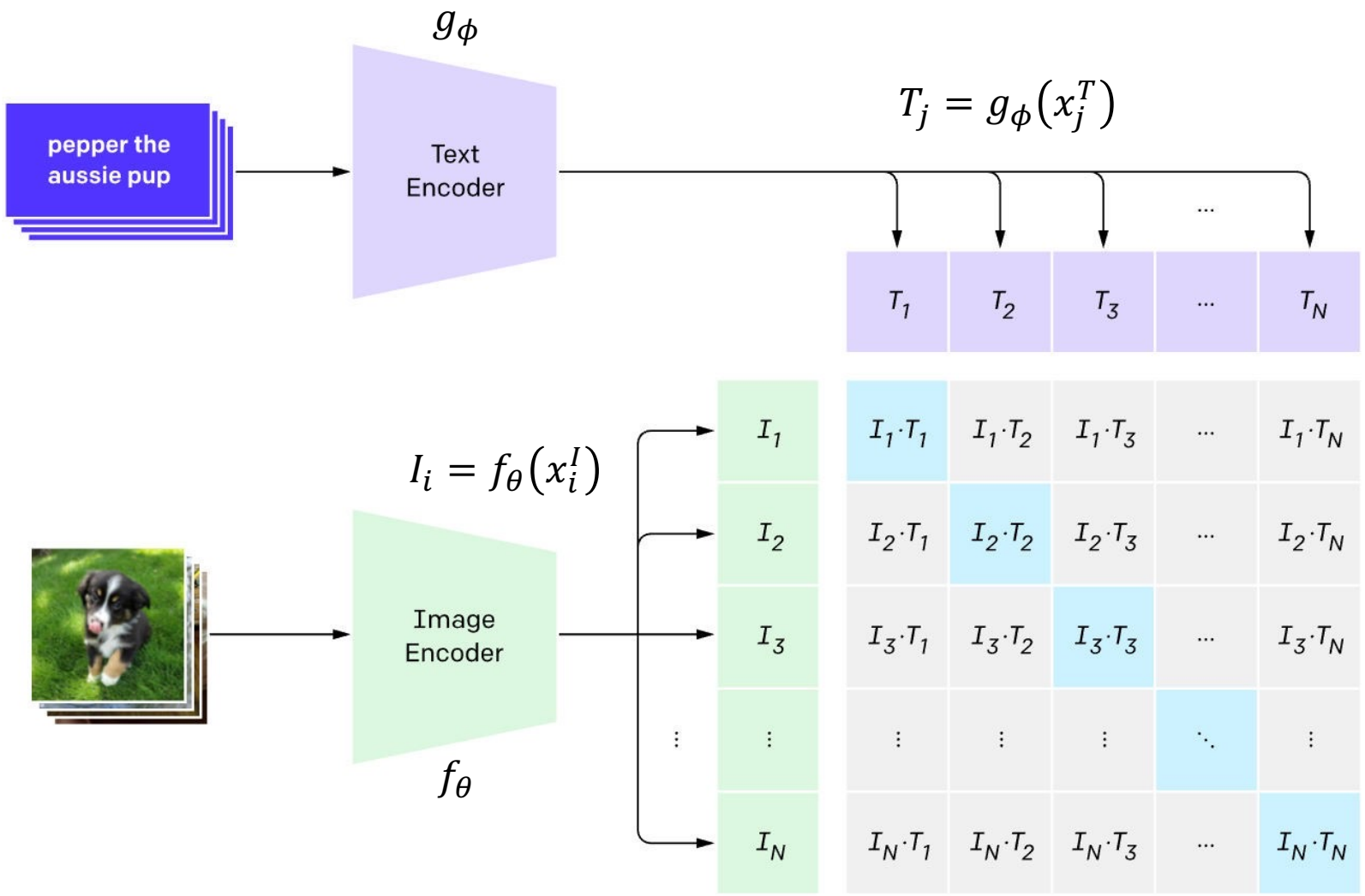


Positive pairs:
Images and corresponding texts

Negative pairs:
Random pairs

A. Radford, J.W. Kim, et al., "Learning Transferable Visual Models From Natural Language Supervision", 2021

CLIP: Contrastive Learning



$$s_{i,j}^T = s_{i,j}^I = I_i^T T_j$$

$$\mathcal{L}_i^I = -\log \frac{e^{s_{i,i}^I}}{\sum_{j=1}^N e^{s_{i,j}^I}}$$

$$\mathcal{L}_j^T = -\log \frac{e^{s_{i,i}^T}}{\sum_{i=1}^N e^{s_{i,j}^T}}$$

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^N (\mathcal{L}_i^I + \mathcal{L}_j^T)$$

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

A. Radford, J.W. Kim, et al., “Learning Transferable Visual Models From Natural Language Supervision”, 2021

CLIP: Architecture and training

- Uses only a linear projection to map from each encoder's representation to the multi-modal embedding space

- A random square crop from resized images is the only data augmentation used during training

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

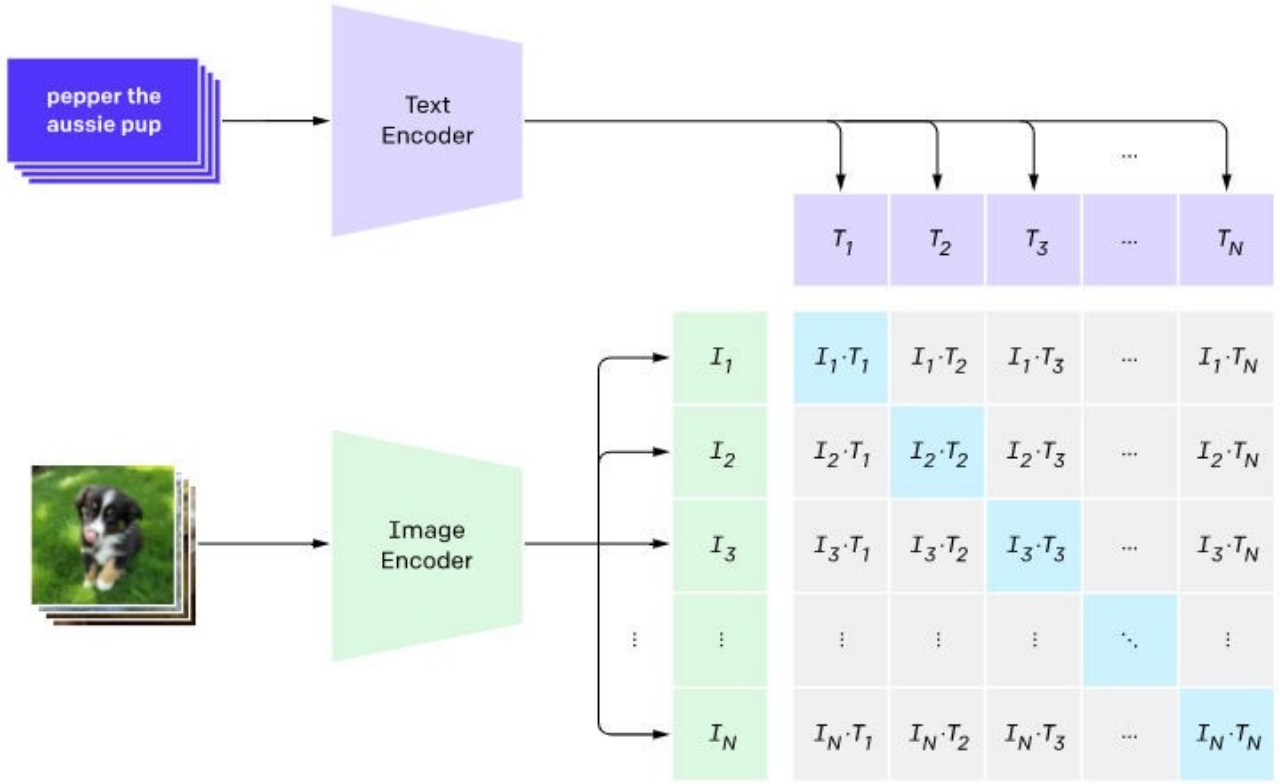
# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```


A. Radford, J.W. Kim, et al., "Learning Transferable Visual Models From Natural Language Supervision", 2021

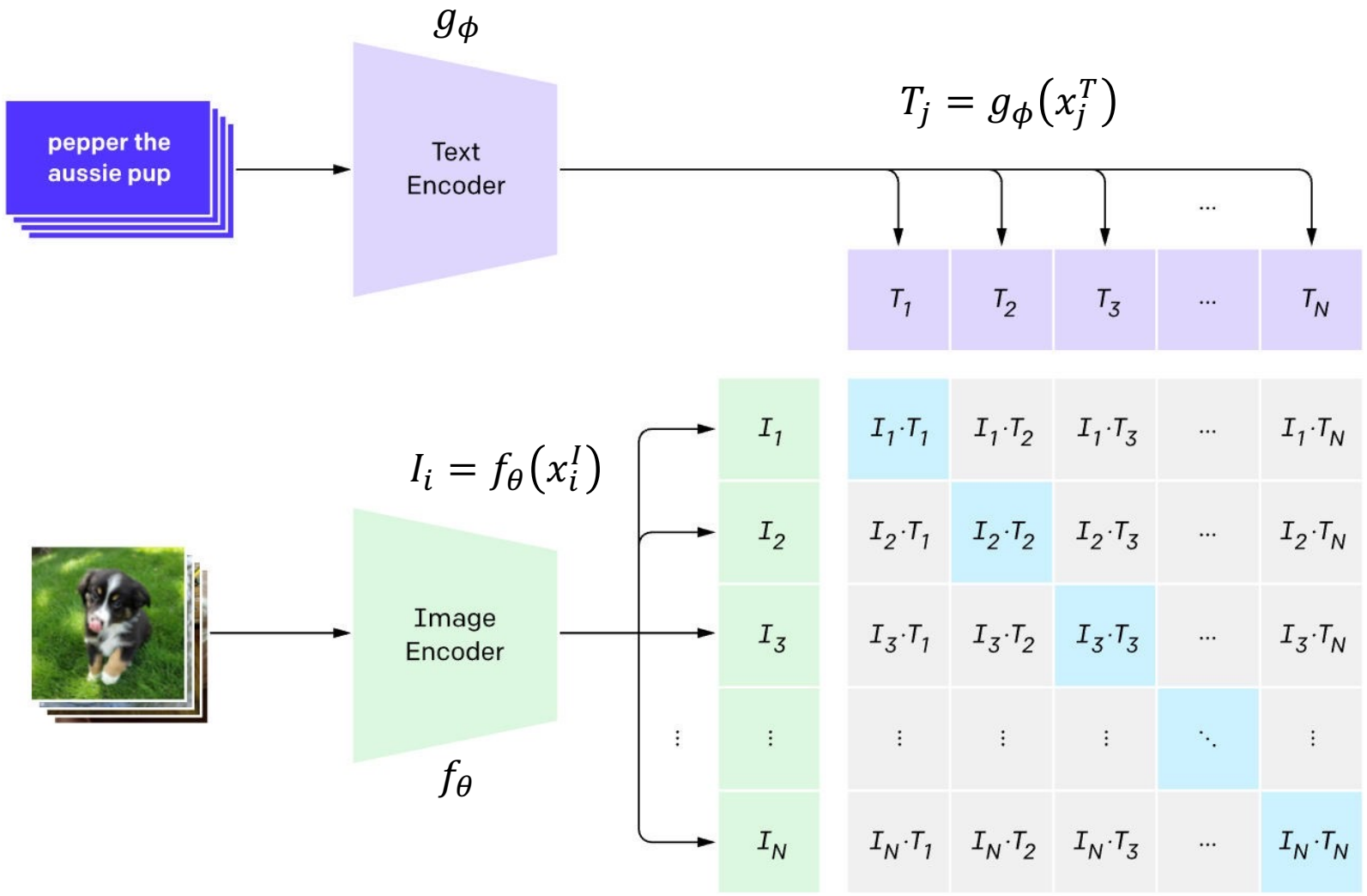
CLIP: Models and Training Complexity



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

A. Radford, J.W. Kim, et al., “Learning Transferable Visual Models From Natural Language Supervision”, 2021

1. Contrastive pre-training



$$s_{i,j}^T = s_{i,j}^I = I_i^T T_j$$

$$\mathcal{L}_i^I = -\log \frac{e^{s_{i,i}^I}}{\sum_{j=1}^N e^{s_{i,j}^I}}$$

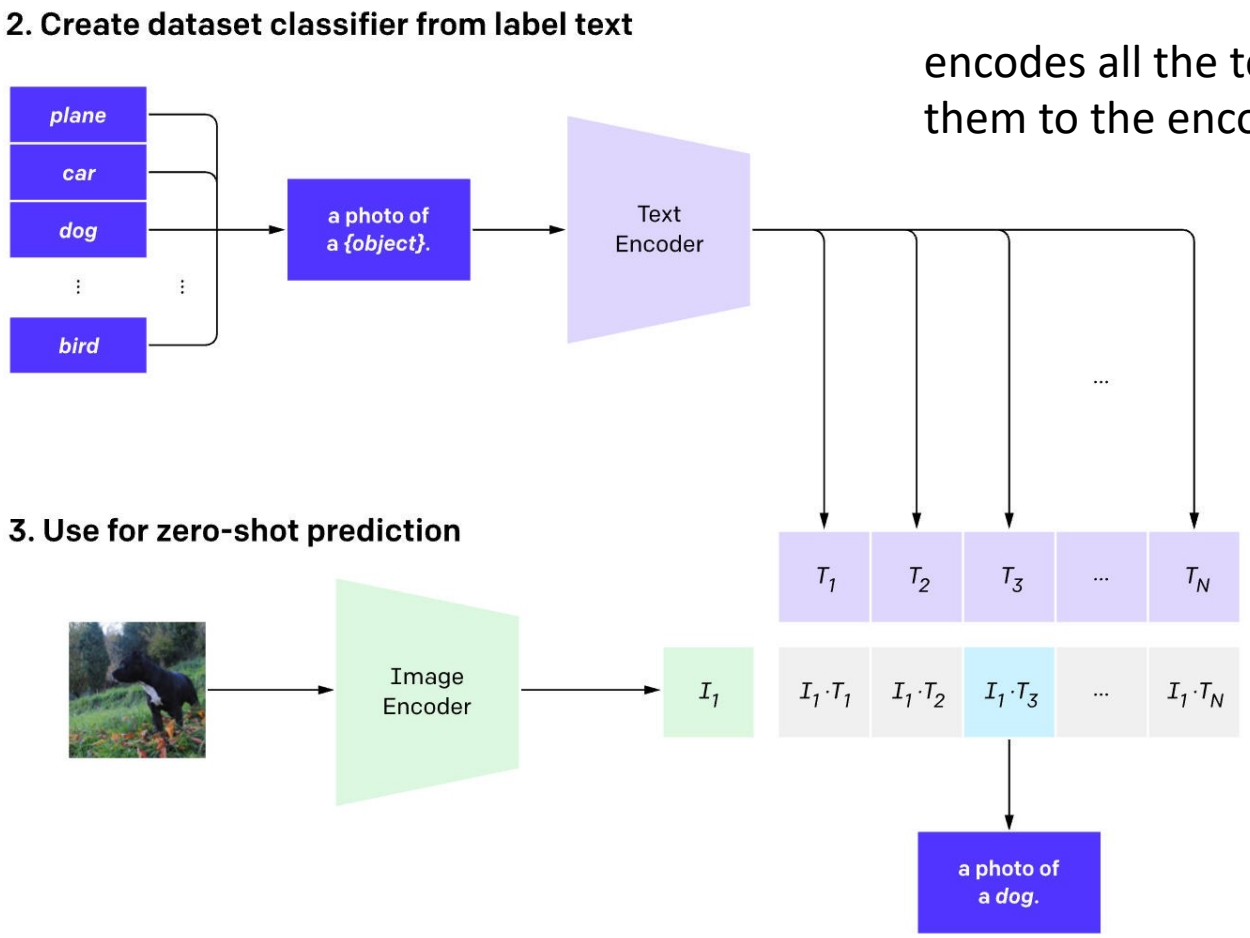
$$\mathcal{L}_j^T = -\log \frac{e^{s_{i,i}^T}}{\sum_{i=1}^N e^{s_{i,j}^T}}$$

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^{N} (\mathcal{L}_i^I + \mathcal{L}_j^T)$$

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

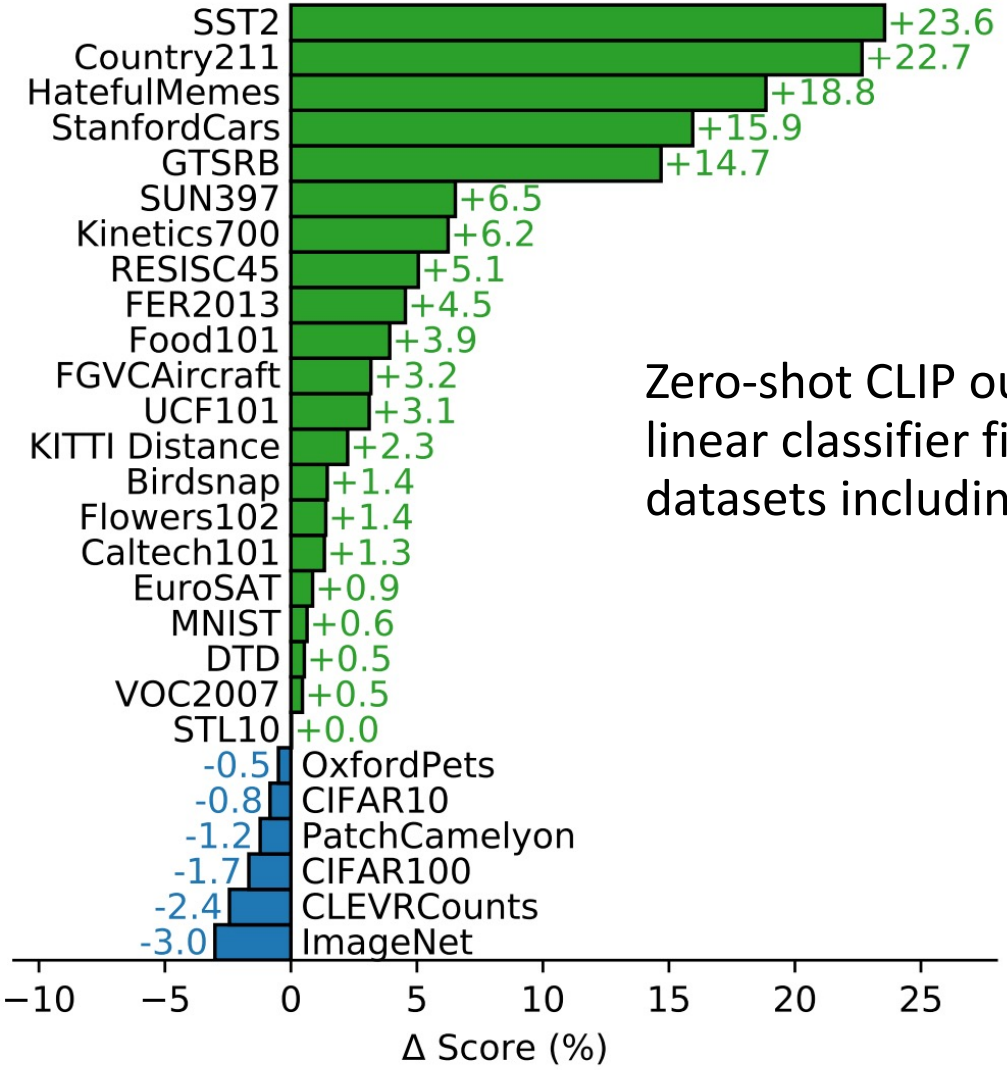
A. Radford, J.W. Kim, et al., “Learning Transferable Visual Models From Natural Language Supervision”, 2021

CLIP for zero-shot learning

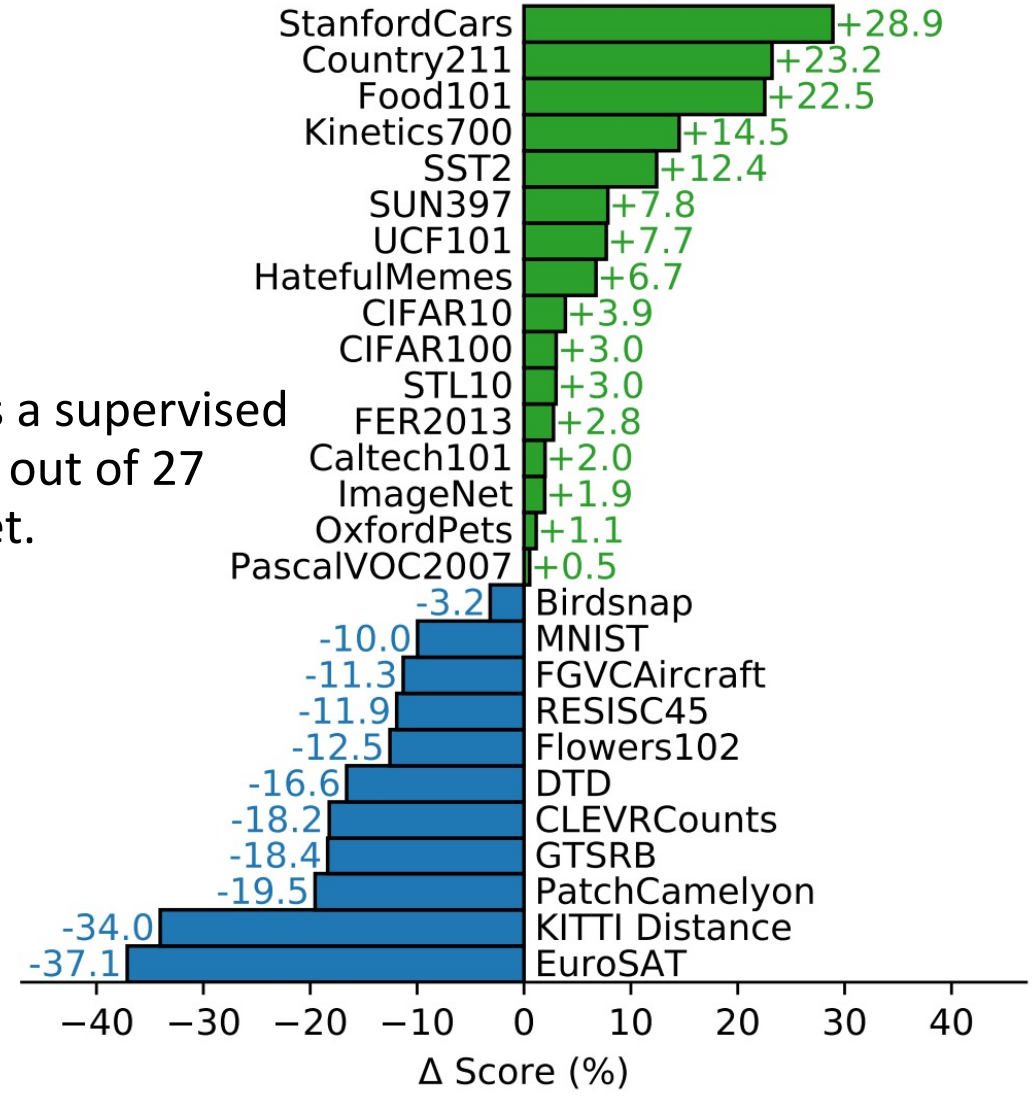


encodes all the text labels and compares them to the encoded image

A. Radford, J.W. Kim, et al., “Learning Transferable Visual Models From Natural Language Supervision”, 2021



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



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

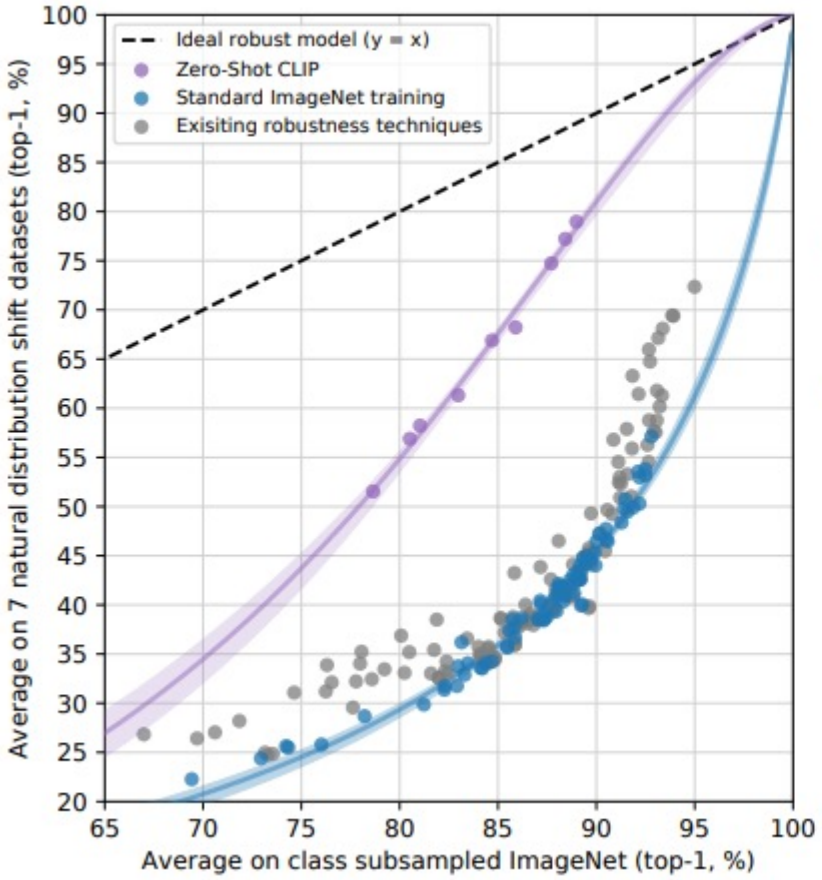
Zero-shot CLIP is competitive with a fully supervised linear-probing ResNet50

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

A. Radford, J.W. Kim, et al., "Learning Transferable Visual Models From Natural Language Supervision", 2021

Robustness to distribution shift

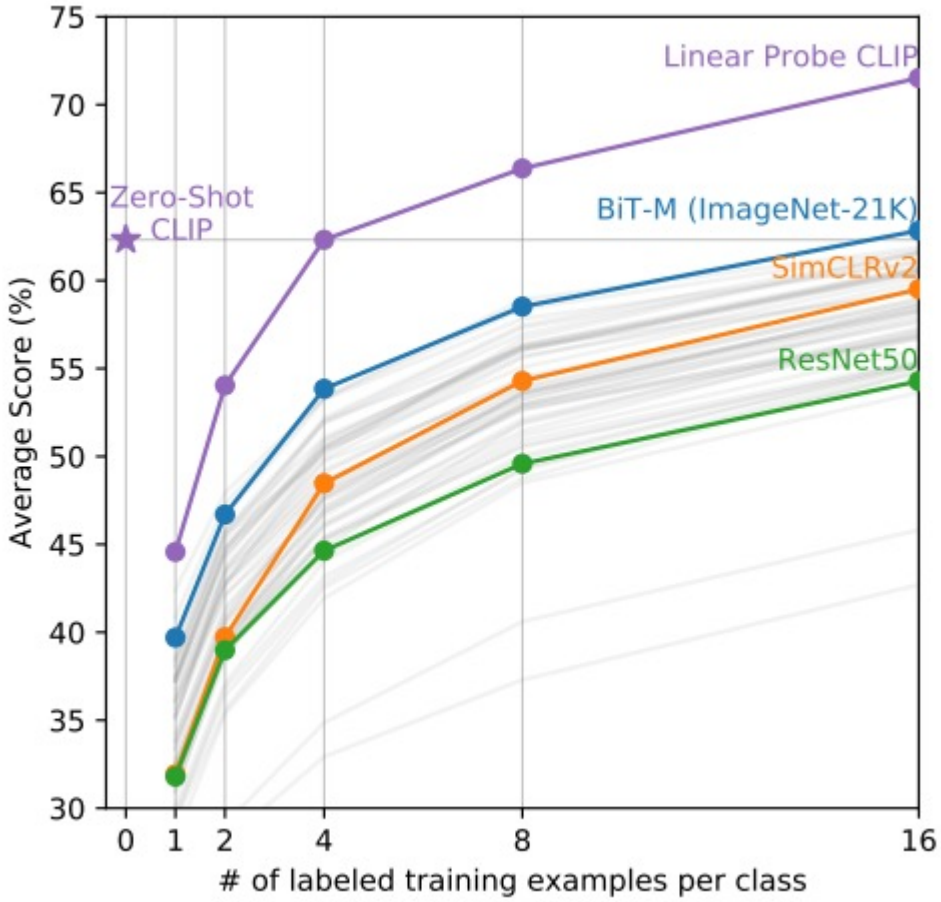
- Zero-shot CLIP is much more robust to distribution shift than standard ImageNet





A. Radford, J.W. Kim, et al., “Learning Transferable Visual Models From Natural Language Supervision”, 2021

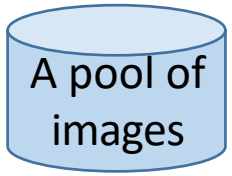
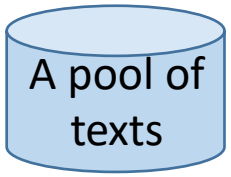
CLIP: Results



A. Radford, J.W. Kim, et al., “Learning Transferable Visual Models From Natural Language Supervision”, 2021

Vision Language Tasks

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

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

Source: P. Zhang and L. Zhang, A Tutorial On Vision Language Intelligence

CLIP: Summary

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

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

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

CLIP Variants

- Objective function or pretraining
  - Combining CLIP with label supervision
  - Contrastive + self-supervised image representation learning
    - Contrastive + Self-supervlised methods like SimCLR
    - Contrastve + Masked Image Modeling (EVA, EVA-02)
  - Fine-grained matching loss or Region-level pretraining
  - …

Learning to Prompt for VLMs

Caltech101



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

(a)

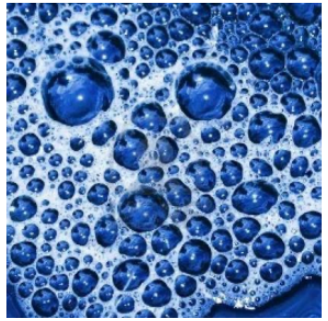
Flowers102



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

(b)

Describable Textures (DTD)



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

(c)

EuroSAT

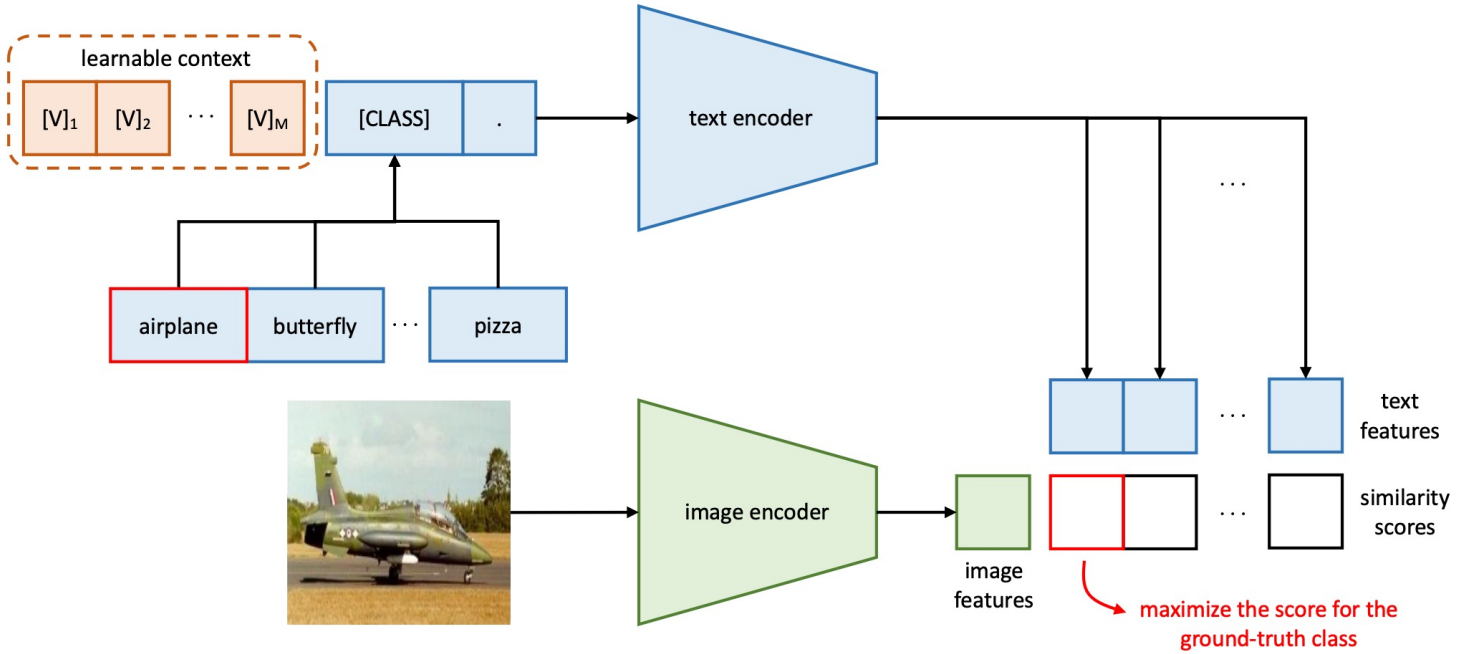


Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**83.53**

(d)

Zhu et al., "Learning to Prompt for Vision-Language Models", 2021

Learning to Prompt for VLMs

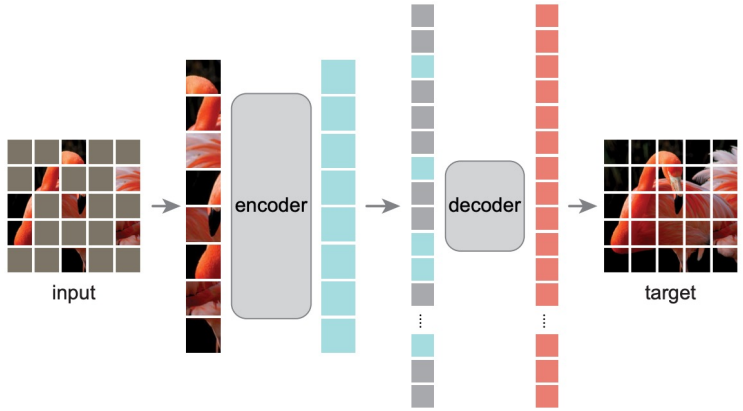


Method	Source	Target			
	ImageNet	-V2	-Sketch	-A	-R
ResNet-50					
Zero-Shot CLIP	58.18	51.34	33.32	21.65	56.00
Linear Probe CLIP	55.87	45.97	19.07	12.74	34.86
CLIP + CoOp ($M=16$)	62.95	55.11	32.74	22.12	54.96
CLIP + CoOp ($M=4$)	**63.33**	**55.40**	**34.67**	**23.06**	**56.60**
ResNet-101					
Zero-Shot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

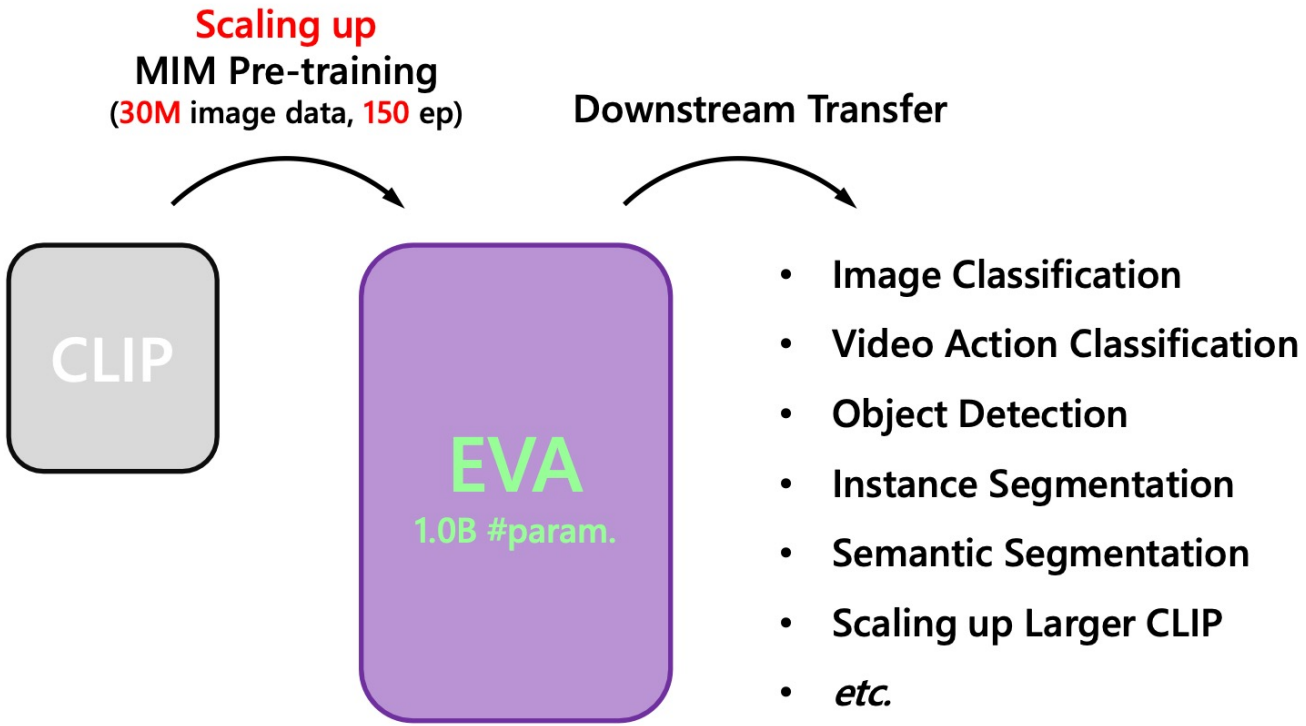
Zhu et al., "Learning to Prompt for Vision-Language Models", 2021

EVA-CLIP

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



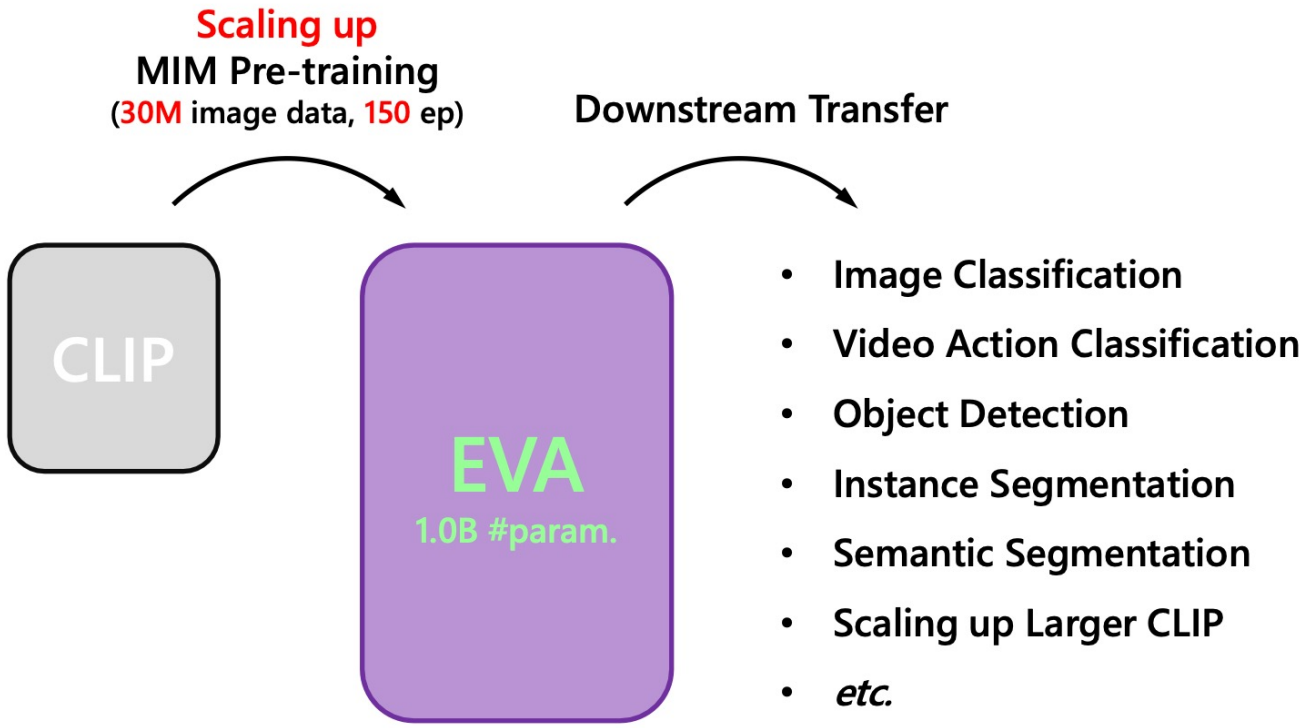
He et al., "Masked Autoencoders Are Scalable Vision Learners", 2021



Fang et al., "EVA: Exploring the Limits of Masked Visual Representation Learning at Scale", 2022.

EVA-CLIP

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



Fang et al., "EVA: Exploring the Limits of Masked Visual Representation Learning at Scale", 2022.

EVA-CLIP

- The merged dataset for pre-training has 29.6 million images in total
- The output feature of EVA is first normalized and then projected to the same dimension as the CLIP feature via a linear layer.
  - negative cosine similarity as the loss function
  - The CLIP features used as MIM prediction targets are trained on a 400 million image-text dataset

Vision-Language Models: Toward generative models

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

Representations: Fusion

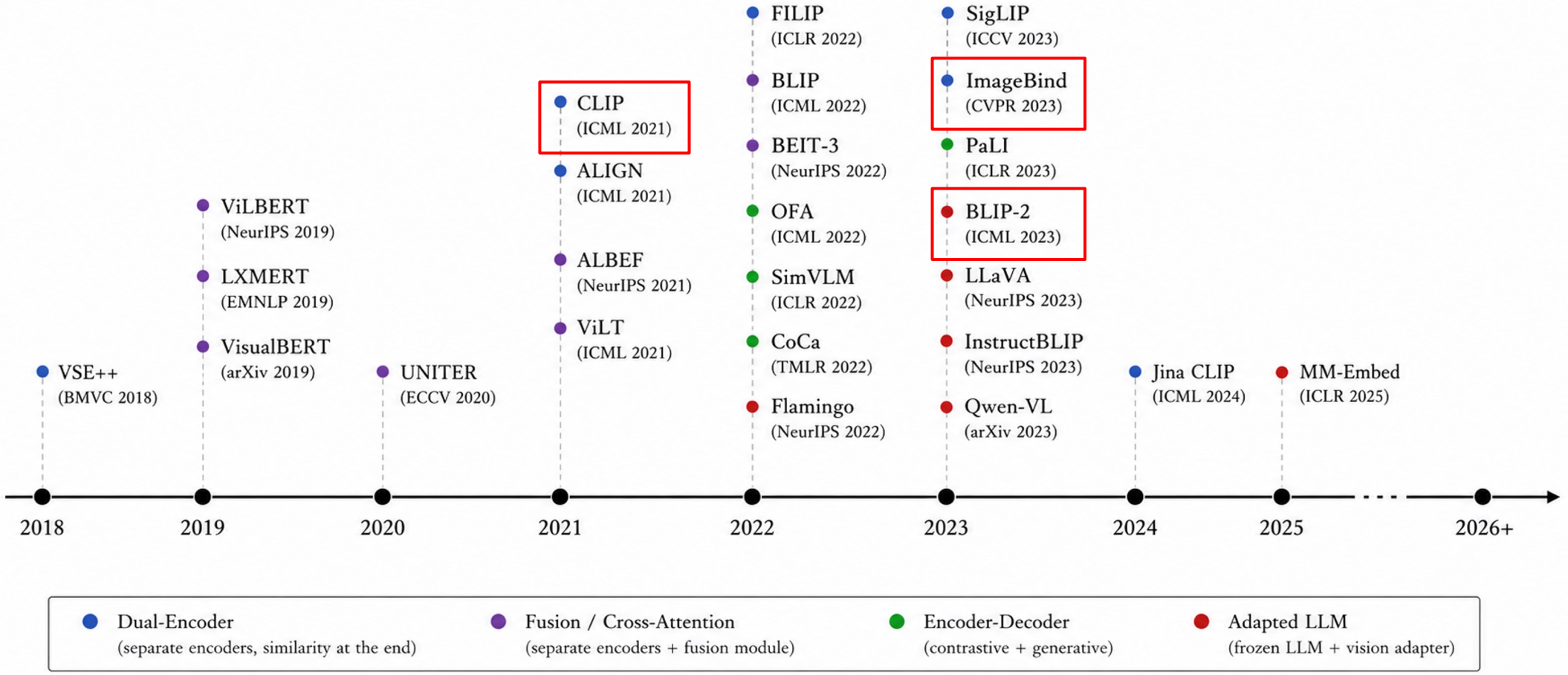
Early fusion concatenates raw inputs before any processing (max interaction, but struggles with heterogeneity),

Late fusion keeps modalities entirely separate until the final output (scalable but misses cross-modal interactions)

Middle fusion encodes each modality independently first then fuses via cross-attention or gating (best balance, dominant in modern systems like LLaVA and BLIP-2)

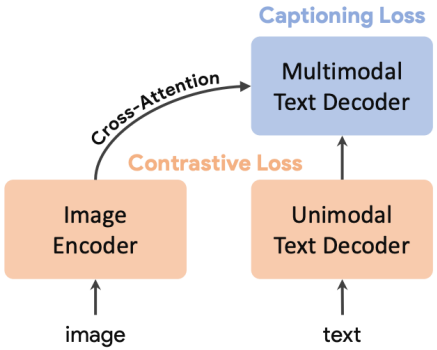
Source: Prof. Paul Liang, MIT Media Lab - How to AI (Almost) Anything

Architecture of Multimodal Models



Source: Prof. Paul Liang, MIT Media Lab - How to AI (Almost) Anything

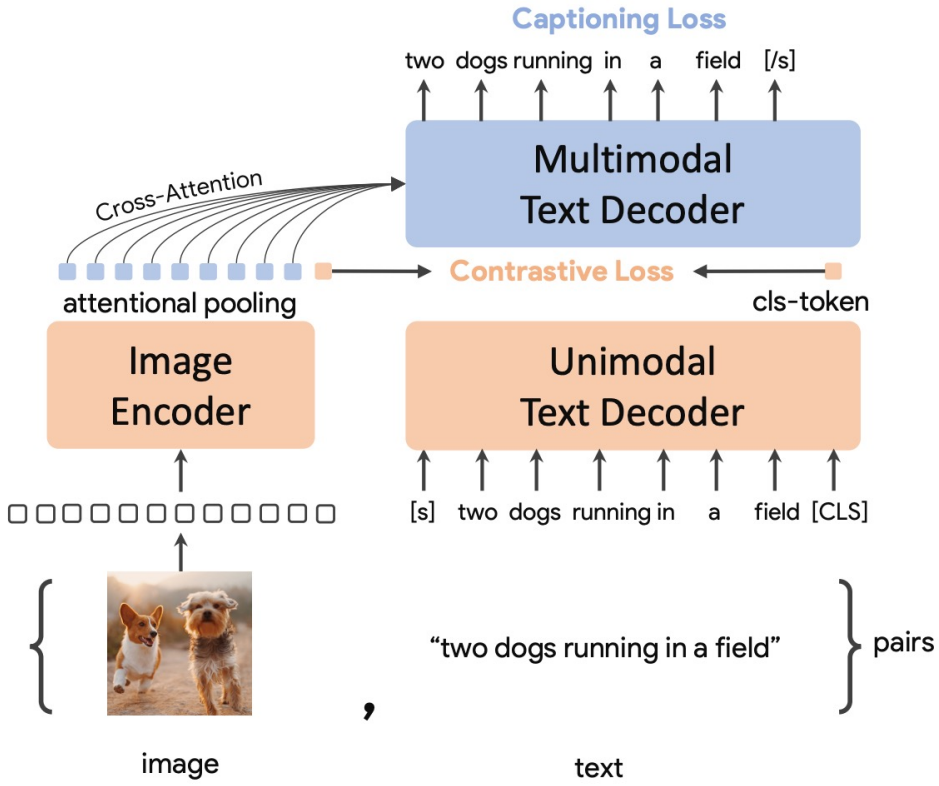
CoCa: Contrastive Captioner



Multi-modal fusion

CoCa

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

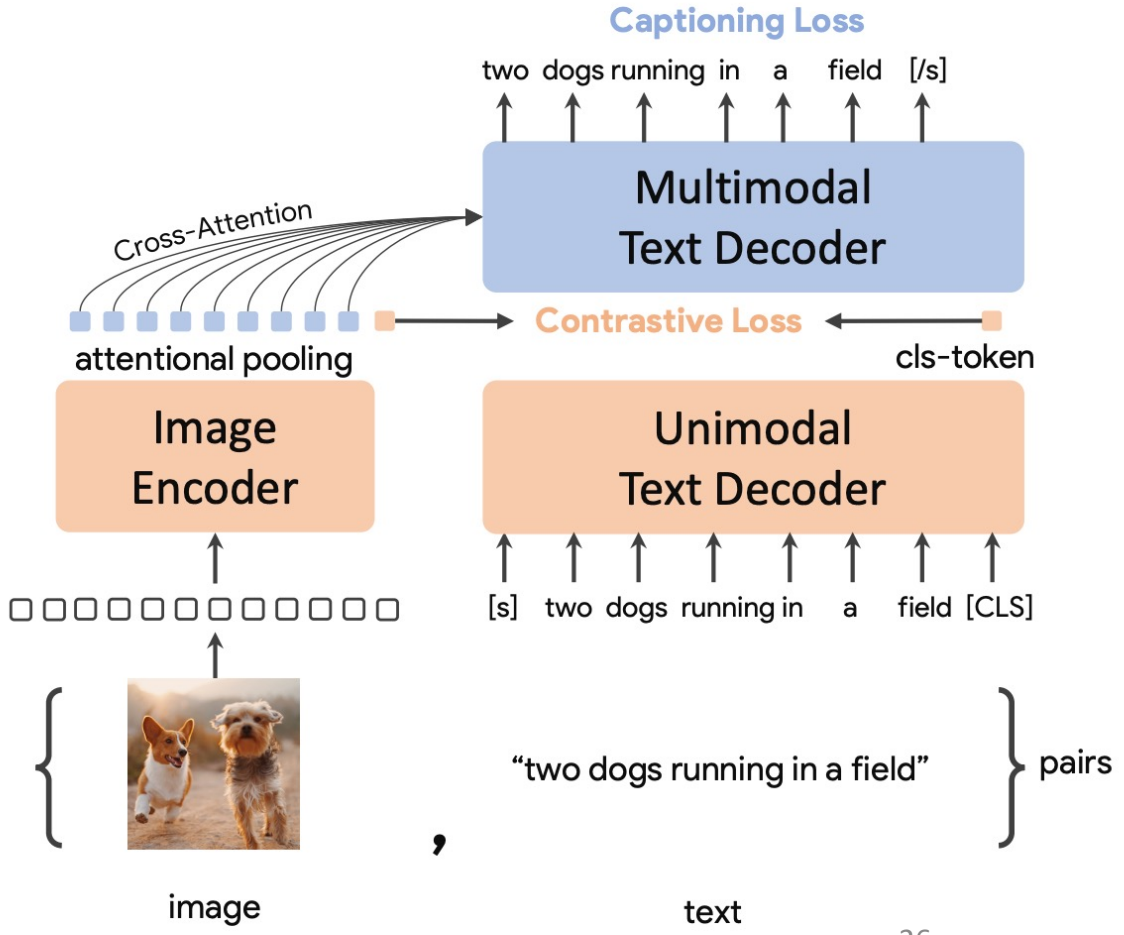


Yu et al., "Coca: Contrastive captioners are image-text foundation models", 2022

CoCa: Contrastive Captioner

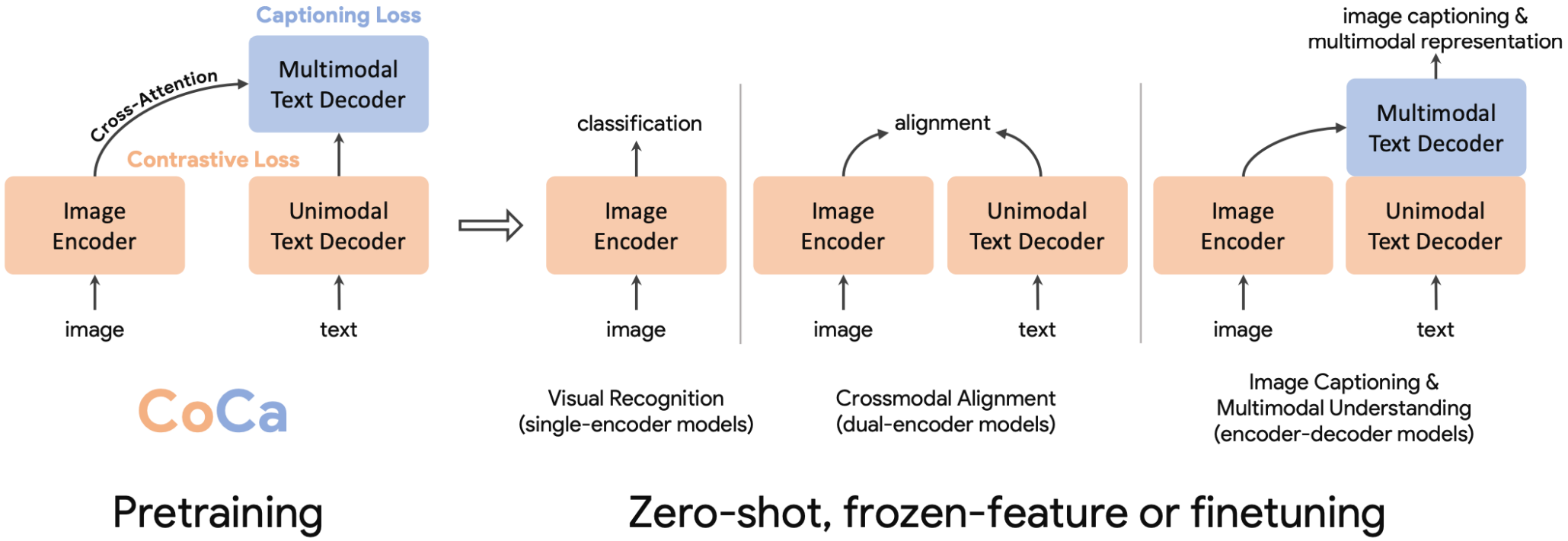
- Use mixed image-text and image-label (JFT-3B) data for pre-training
- A generative branch for enhanced performance and enabling new capabilities (image captioning and VQA)
- CoCa aims to learn a better image encoder from scratch

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



Yu et al., "Coca: Contrastive captioners are image-text foundation models", 2022

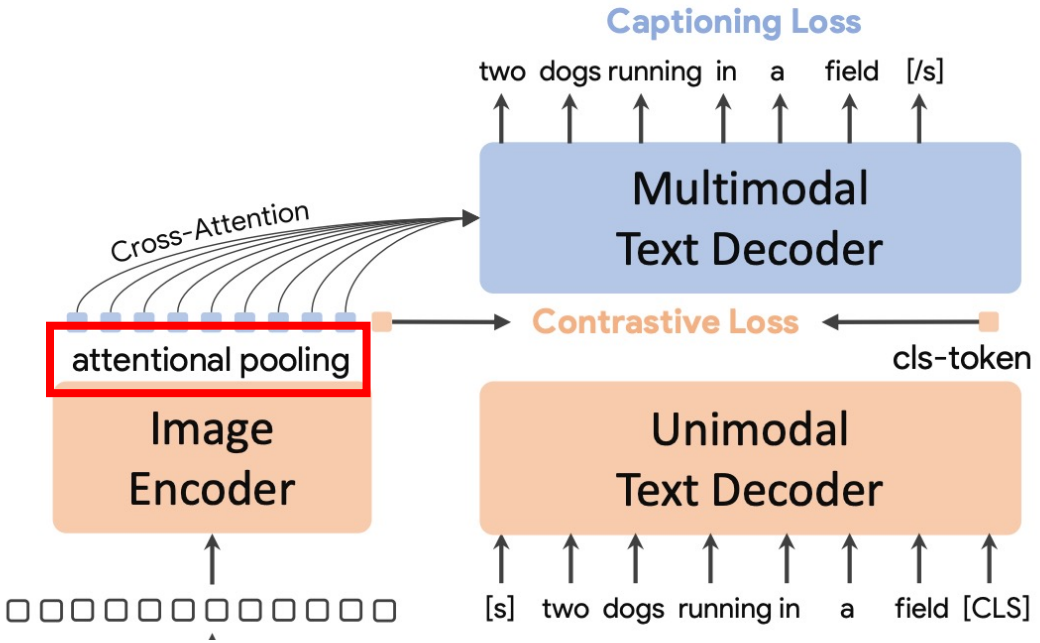
CoCa: Contrastive Captioner



Yu et al., “Coca: Contrastive captioners are image-text foundation models”, 2022

CoCa Architecture: Summary

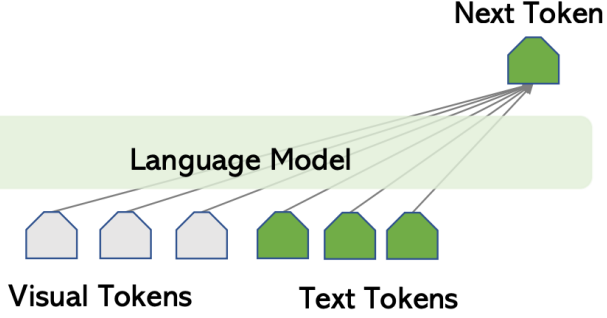
- Unified single-encoder, dual-encoder, and encoder-decoder paradigms
  - one image-text foundation model with the capabilities of all three approaches
- Cross-attention is omitted in unimodal decoder layers to encode text-only representations
- Multimodal decoder cross-attending to image encoder outputs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



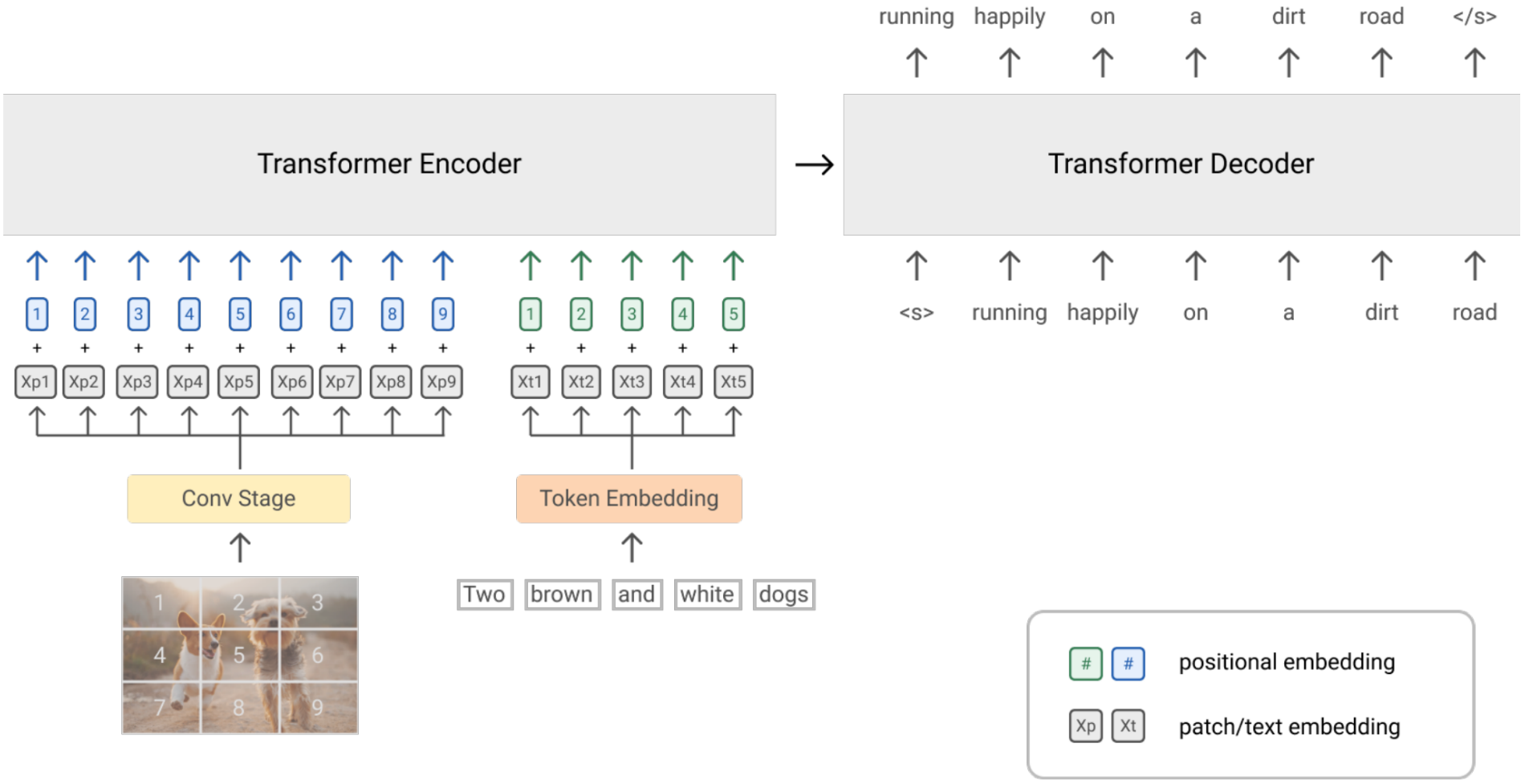
Yu et al., “Coca: Contrastive captioners are image-text foundation models”, 2022

SimVLM

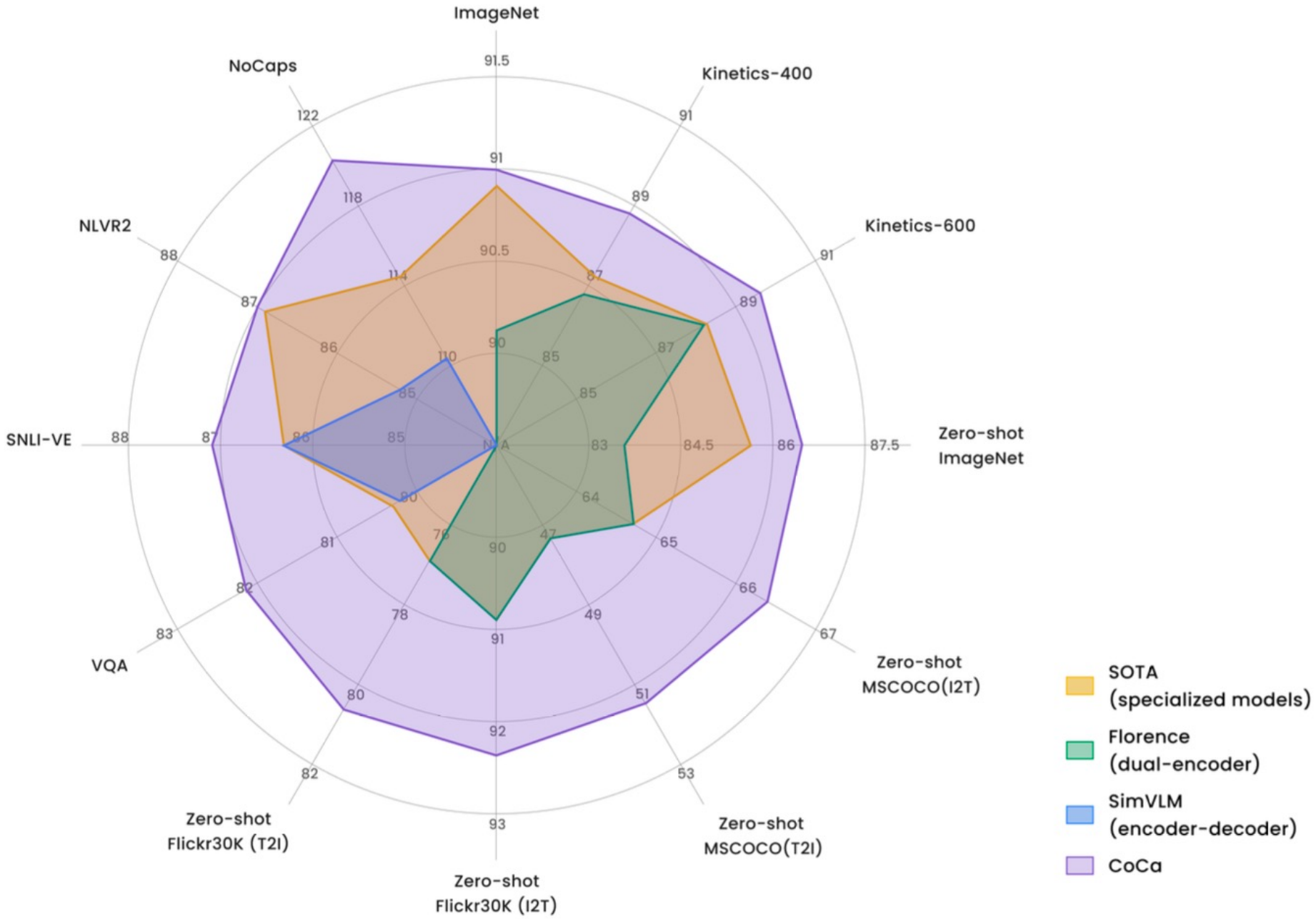
- PrefixLM



Chunyuan Li's CVPR 2023
Tutorial: Large Multimodal Models

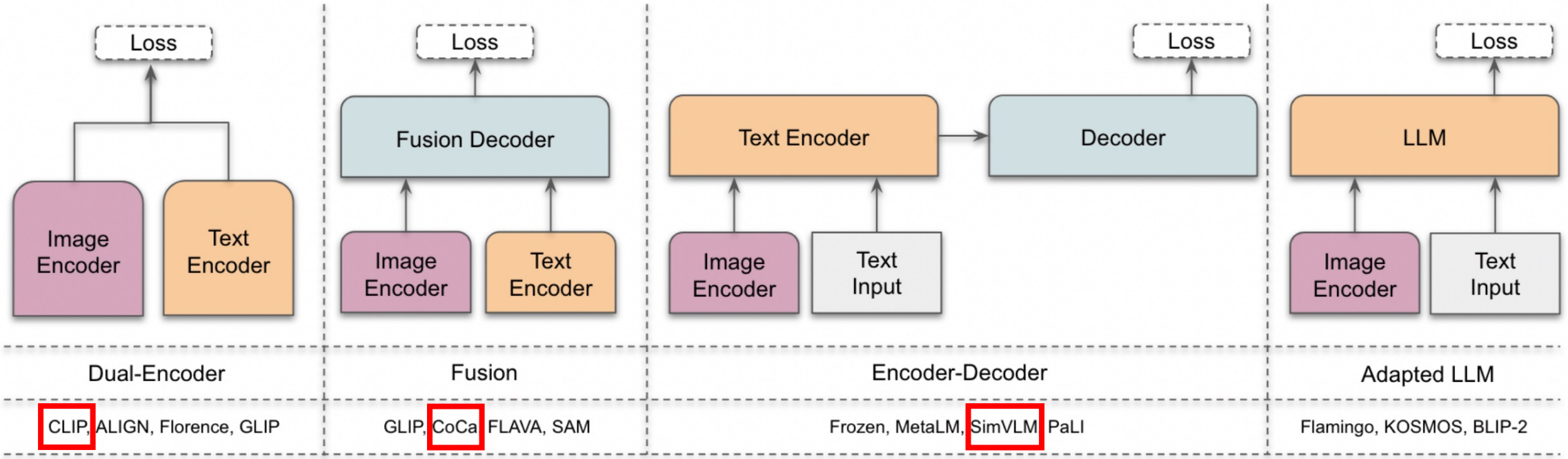


Wang et al., "SimVLM: Simple Visual Language Model Pretraining with Weak Supervision", ICLR 2022



Yu et al., “Coca: Contrastive captioners are image-text foundation models”, 2022

Architecture of Multimodal Models

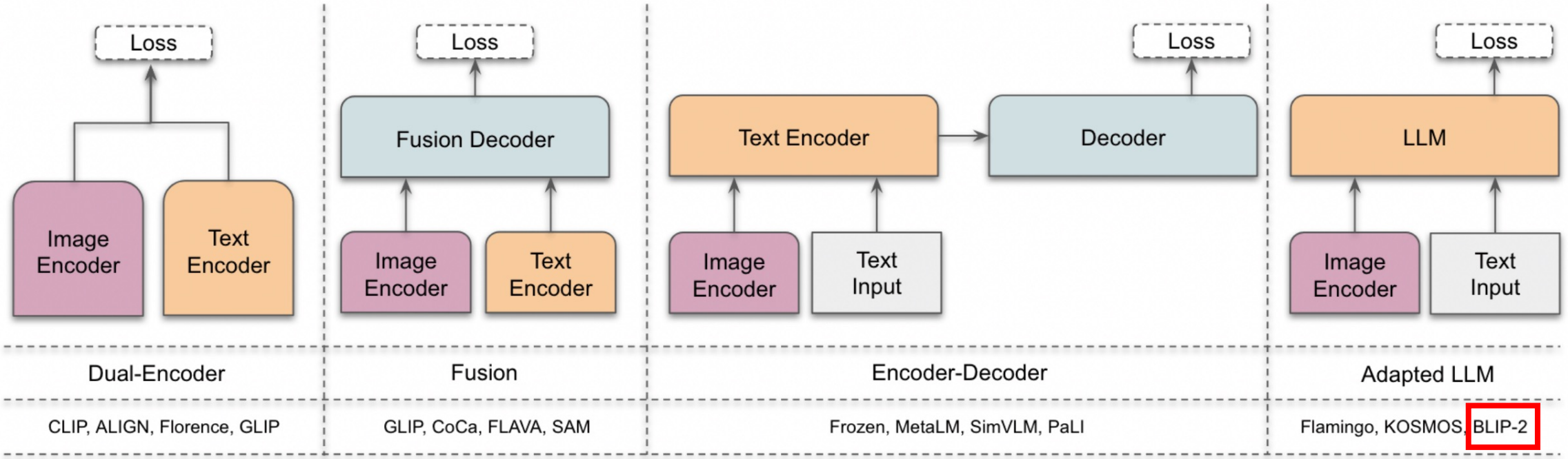


Li et al., “Multimodal Foundation Models: From Specialists to General-Purpose Assistants”, 2023

Image-Text Models: Summary

- CLIP bridges the vision and language spaces
- It showcases impressive capabilities for zero-shot adaptation to unseen tasks and image2text and text2image retrieval tasks
  - However, it does not utilize for tasks like image captioning and VQA
- Encoder-decoder and fusion-decoder architectures provide ability of generating text

Architecture of Multimodal Models



Li et al., “Multimodal Foundation Models: From Specialists to General-Purpose Assistants”, 2023

LLMs as a Universal Interface for Multi-modal Systems

- Multimodal systems can provide a more flexible way of interaction
  - typing, talking, or pointing your camera at something can be used throughout the question
- In the previous part, the interface had limited interactivity and adaptability to the user's instructions
- How can we utilize the potentials of LLMs for creating a general-purpose assistant?
  - various tasks can be explicitly represented in language
- They must be able to ingest a multimodal prompt containing images and/or videos interleaved with text.

Large Multi-modal Models (LMMs)

- Equip LLMs with eyes to see the world
  - by training them on vision-conditioned language generation tasks
  - and so use LLMs as a general interface for other modalities
  - and thus make use of facts that it has learned during language-only pre-training

- Multimodal In-Context Learning (M-ICL)
  - use demonstrations of some examples to conduct few-shot learning

- Multimodal Instruction-Tuning (M-IT)
  - Finetune the model on instruction-following & use task instructions to repurpose the model

GPT 4: Example

User	What is funny about this image? Describe it panel by panel.



Source: https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/

GPT-4	The image shows a package for a "Lightning Cable" adapter with three panels. Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port. Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it. Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end. The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

https://openai.com › gpt-4

(2) Prompt:

Describe the pointed region in the image.

Method	Validataion set								Test set							
	in.		near.		out.		overall		in.		near.		out.		overall	
	C	S	C	S	C	S	C	S	C	S	C	S	C	S	C	S
OSCAR	85.4	11.9	84.0	11.7	80.3	10.0	83.4	11.4	84.8	12.1	82.1	11.5	73.8	9.7	80.9	11.3
Human	84.4	14.3	85.0	14.3	95.7	14.0	87.1	14.2	80.6	15.0	84.6	14.7	91.6	14.2	85.3	14.6
VIVO	92.2	12.9	87.8	12.6	87.5	11.5	88.3	12.4	89.0	12.9	87.8	12.6	80.1	11.1	86.6	12.4
VinVL	103.7	13.7	95.6	13.4	83.8	11.9	94.3	13.1	98.0	13.6	95.2	13.4	78.0	11.5	92.5	13.1
UFO	103.9	14.5	95.5	13.8	83.5	12.3	94.3	13.6	98.9	14.3	94.7	13.9	77.9	12.1	92.3	13.6
mPLUG	-	-	-	-	-	-	114.8	14.8	-	-	-	-	-	-	-	-
SimVLM	113.7	-	110.9	-	115.2	-	115.2	-	113.7	-	110.9	-	115.2	-	115.2	-
LEMON	118.0	15.4	116.3	15.1	120.2	14.5	117.3	15.0	112.8	15.2	115.5	15.1	110.1	13.7	114.3	14.9
UniversalCap	123.2	15.0	121.5	15.3	123.4	14.4	122.1	15.0	118.9	15.4	120.6	15.3	114.3	14.1	119.3	15.1
CoCa	-	-	-	-	-	-	122.4	15.5	-	-	-	-	-	-	120.6	15.5
GIT_B	100.7	13.8	97.7	13.5	89.6	12.5	96.6	13.4	-	-	-	-	-	-	-	-
GIT_L	107.7	14.9	107.8	14.5	102.5	13.7	106.9	14.4	-	-	-	-	-	-	-	-
GIT	**129.8**	**16.3**	124.1	16.0	127.1	15.7	125.5	16.0	122.4	16.2	123.9	16.0	122.0	**15.7**	123.4	15.9
GIT2	126.9	16.1	**125.8**	**16.2**	**130.6**	**15.8**	**126.9**	**16.1**	**124.2**	**16.4**	**125.5**	**16.1**	**122.3**	15.6	**124.8**	**16.1**

GPT-4V:

The highlighted red region in the image is a table cell that contains a numerical value of 122.3. It is located in the "Test set" section of the table, under the "out" column and the "GIT2" row. This value represents the overall performance metric for the GIT2 method on the test set, in the "out" category.

Prompt:

Localize each person in the image using bounding box. What is the image size of the input image?



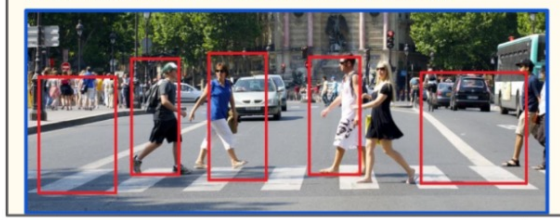
GPT-4V:

The image size of the input image is 1000 x 333 pixels.

Bounding boxes for each person:
- Person 1 (leftmost): (x1, y1) = (22, 108), (x2, y2) = (170, 300)
- Person 2: (x1, y1) = (200, 80), (x2, y2) = (290, 270)
- Person 3: (x1, y1) = (350, 70), (x2, y2) = (460, 280)
- Person 4: (x1, y1) = (540, 80), (x2, y2) = (640, 270)
- Person 5 (rightmost): (x1, y1) = (760, 100), (x2, y2) = (920, 290)

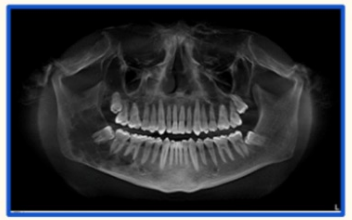
Note: These bounding box coordinates are approximate and may not be exact.

Visualization of Detected Bbox:



Prompt:

Which wisdom tooth is hardest to remove?



Note: Same query image as shown above, resized to save space

GPT-4V:

The lower wisdom teeth, especially those that are impacted, are typically considered the hardest to remove. This is because the lower jawbone is denser than the upper jawbone, and the impacted wisdom teeth may be more difficult to access. Additionally, the roots of the lower wisdom teeth are often longer and more curved than those of the upper wisdom teeth, making them more challenging to remove.

Prompt:

Count the number of apples in the image.

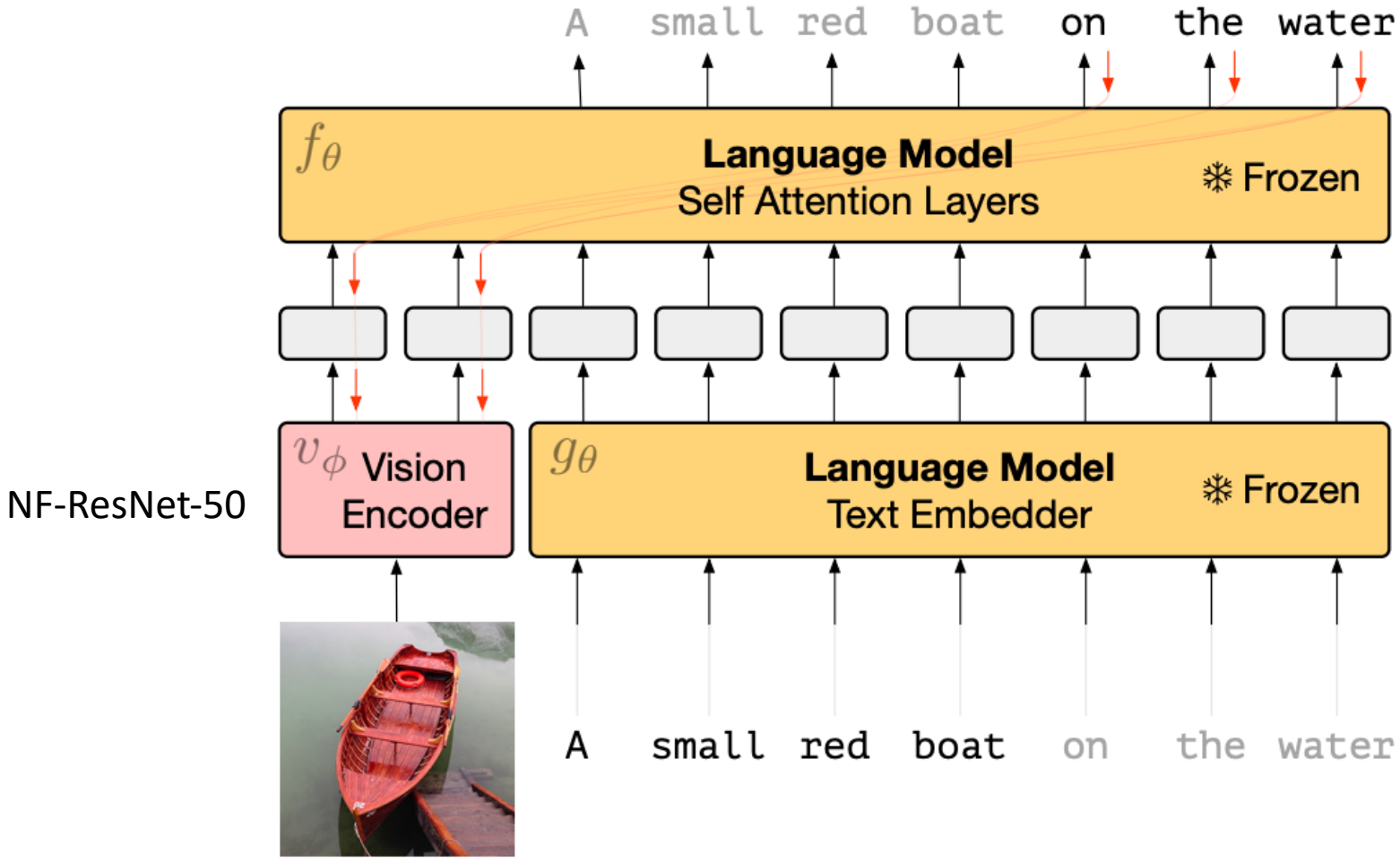


GPT-4V:

There are 7 apples in the image.

https://huyenchip.com/2023/10/10/multimodal.html

Frozen LM Prefix



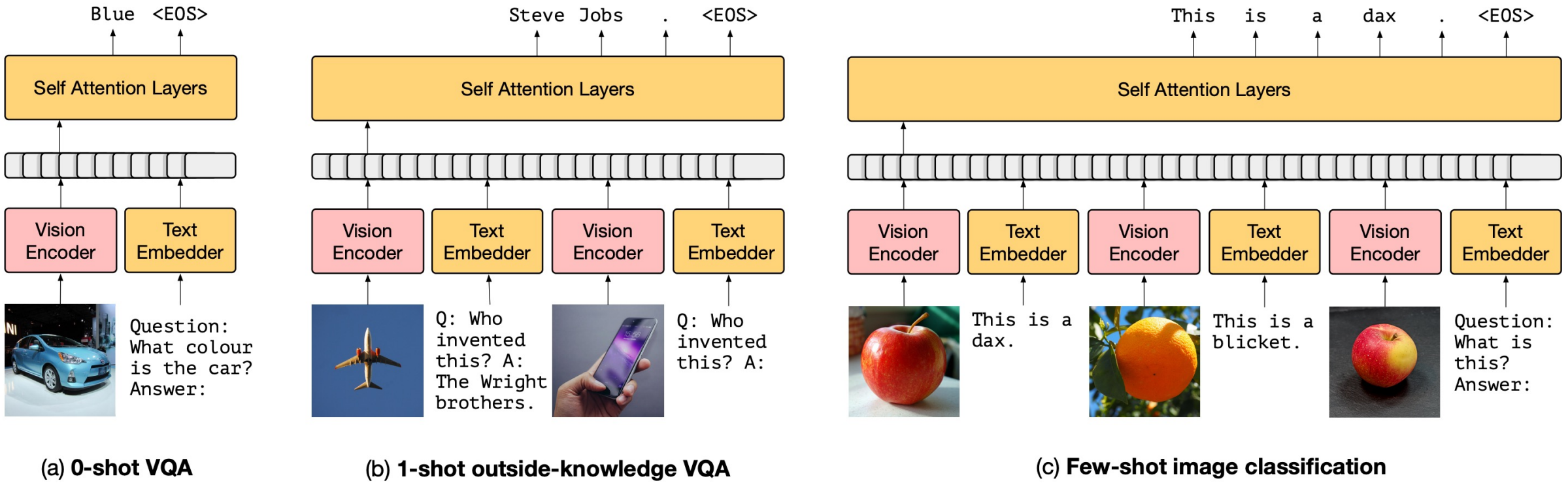
Tsimpoukelli et al., “Multimodal Few-Shot Learning with Frozen Language Models”, ICLR 2022

Frozen LM Prefix

- Frozen: Visual modalities are incorporated as input of LLMs without needing to update LLMs weights
- Concatenated textual and visual embeddings are fed to the decoder of the LLM, which generates a textual output autoregressively
- Finetunes an image encoder whose outputs are directly used as soft prompts for the LLM.

Tsimpoukelli et al., “Multimodal Few-Shot Learning with Frozen Language Models”, ICLR 2022

Frozen: Multimodal few shot learning

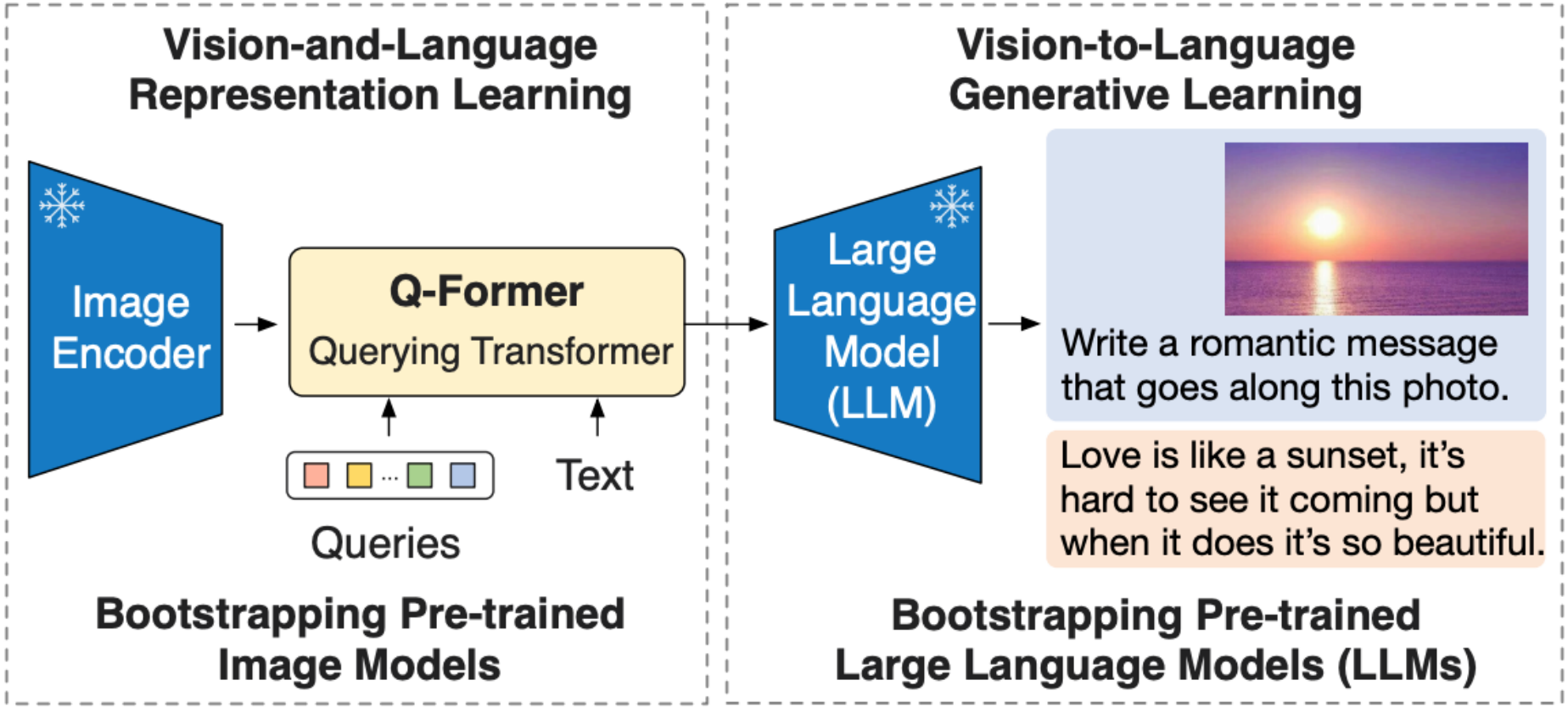


Tsimpoukelli et al., “Multimodal Few-Shot Learning with Frozen Language Models”, ICLR 2022

Architectures

- Dual-encoder
- Encoder-decoder
- Fusion-encoder
- **Unified transformer**

BLIP-2



Aligns pretrained and frozen image encoders and language models by a lightweight Q-Former

Li et al., "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models", ICML 2023

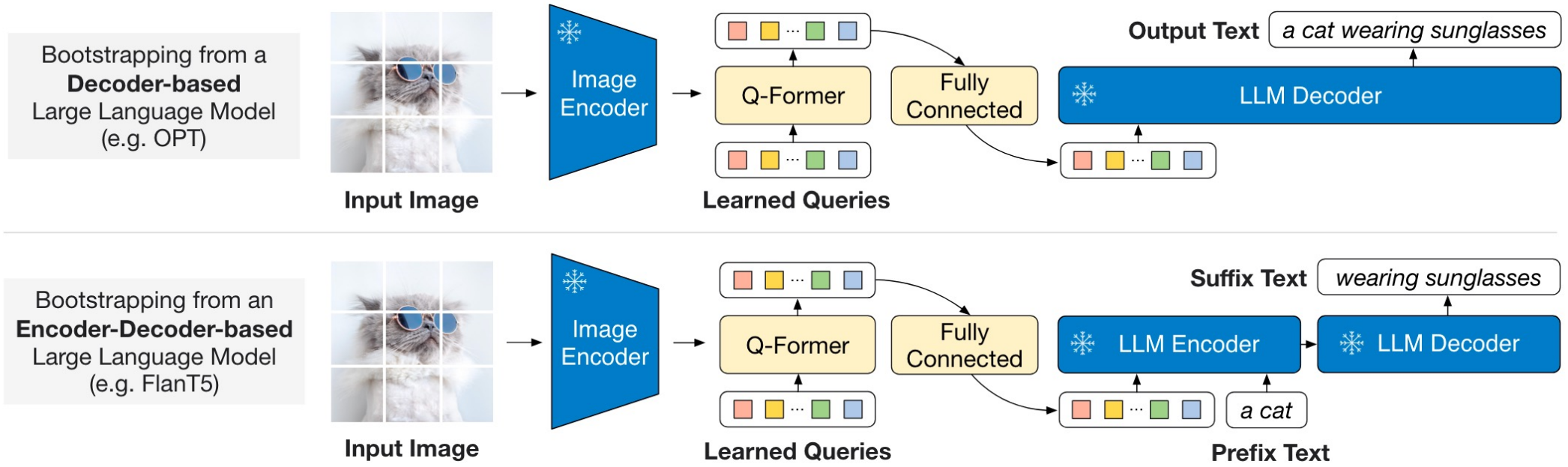
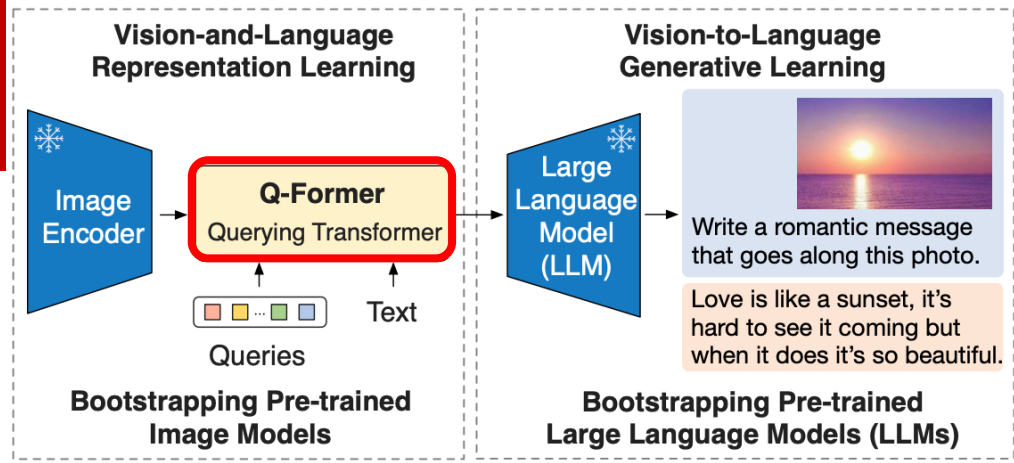
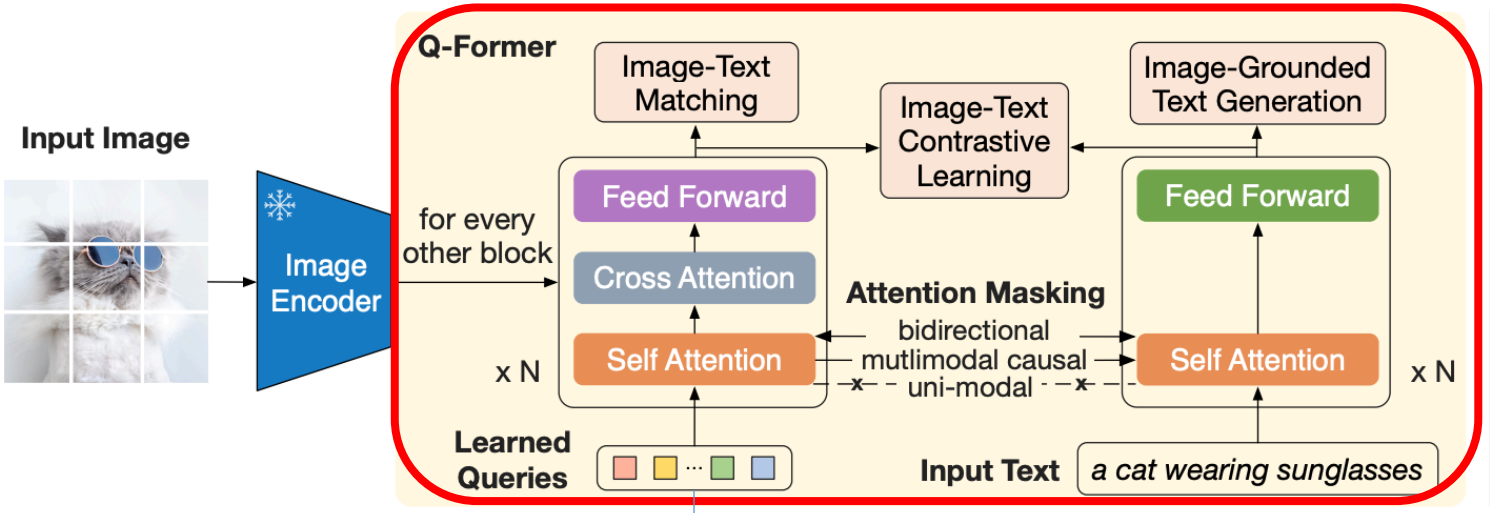


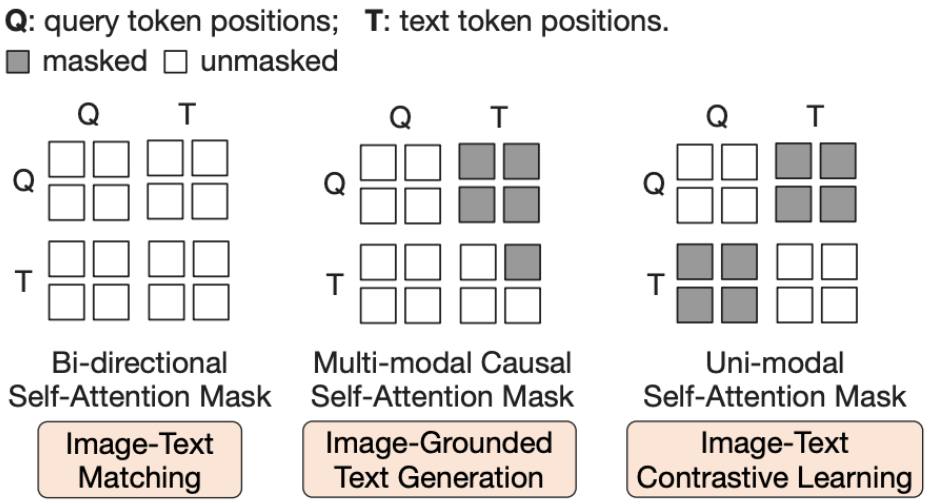
Figure 3. BLIP-2's second-stage vision-to-language generative pre-training, which bootstraps from frozen large language models (LLMs). (**Top**) Bootstrapping a decoder-based LLM (e.g. OPT). (**Bottom**) Bootstrapping an encoder-decoder-based LLM (e.g. FlanT5). The fully-connected layer adapts from the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

BLIP-2: Q-former





a set of learnable query vectors to extract visual features from the frozen image encoder

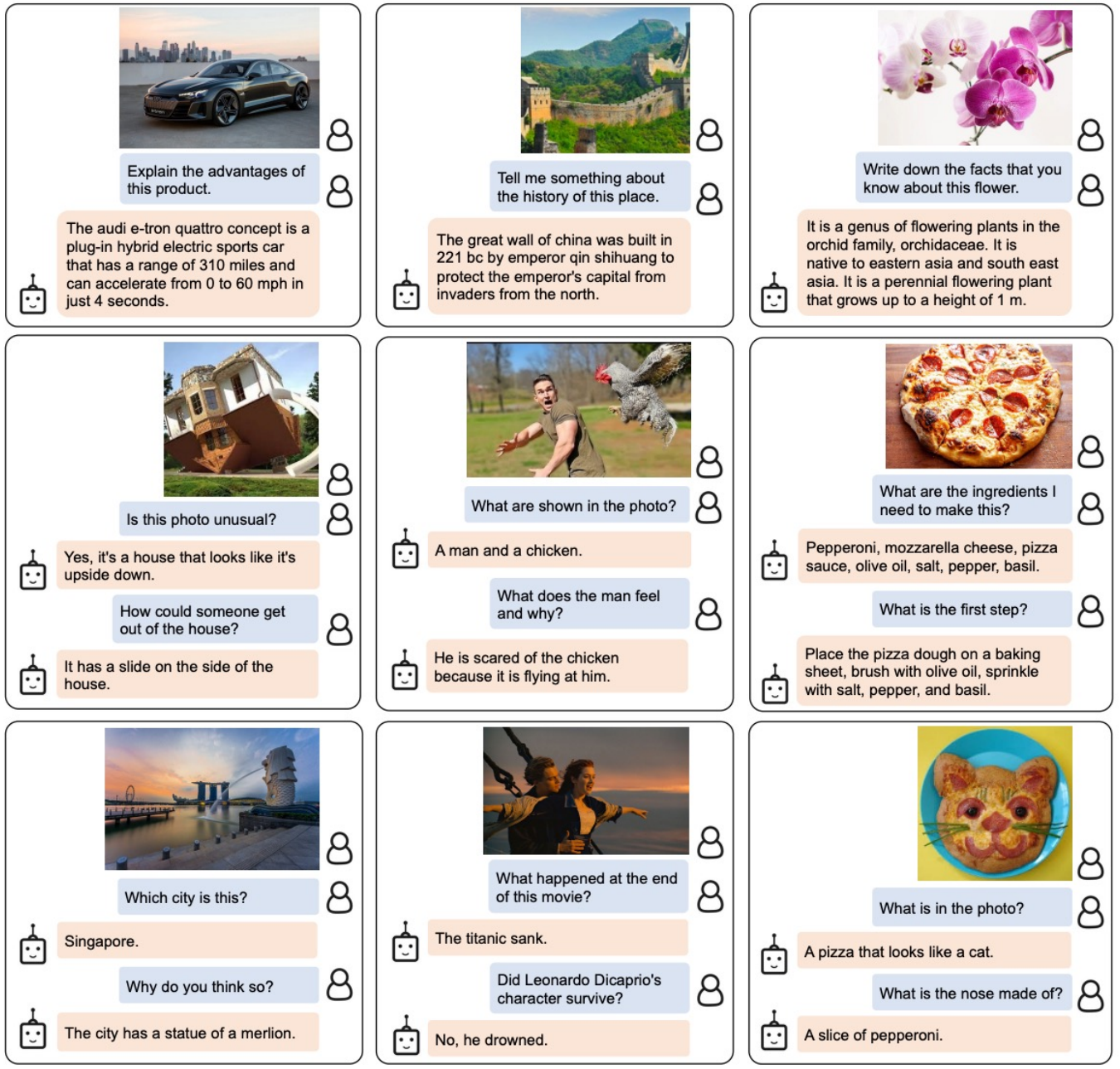


Li et al., "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models", ICML 2023

BLIP-2: Emerging capabilities

- Powered by LLMs (e.g. OPT and FlanT5), BLIP-2 can be prompted to perform zero-shot image-to-text generation
  - follows natural language instructions
  - enables emerging capabilities such as visual knowledge reasoning and visual conversation

Li et al., “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models”, ICML 2023



Li et al., “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models”, ICML 2023

Zero-shot Results on Various Tasks

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**

Li et al., "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models", ICML 2023

LLaVA: Large Language-and-Vision Assistant

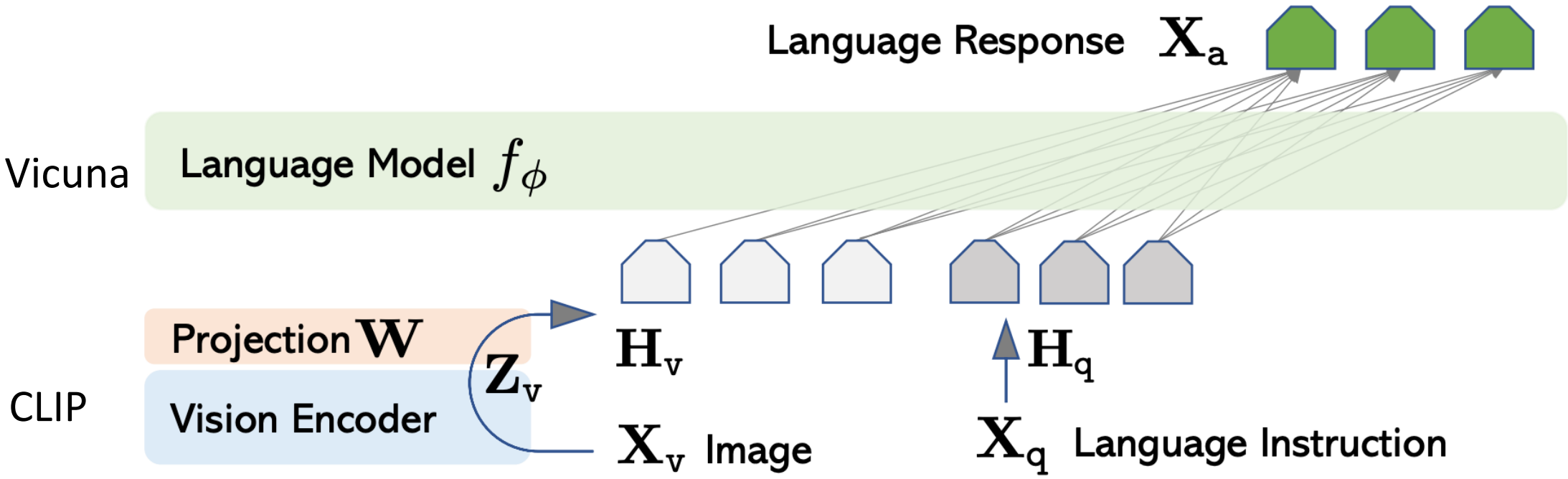
- The available multimodal instruction-following data is limited
  - time-consuming collection
  - less well-defined when human crowd-scouring is considered.

- LLaVA is the first visual instruction-following method

- LLaVA fine-tunes the whole LLM on 150K high-quality multi-modal instruction data generated by GPT-4.
  - expanding the original captions of COCO with handwritten seed instructions using GPT-4 to provide descriptions and multi-round conversations.

Liu et al., “Visual Instruction Tuning”, NeurIPS 2023

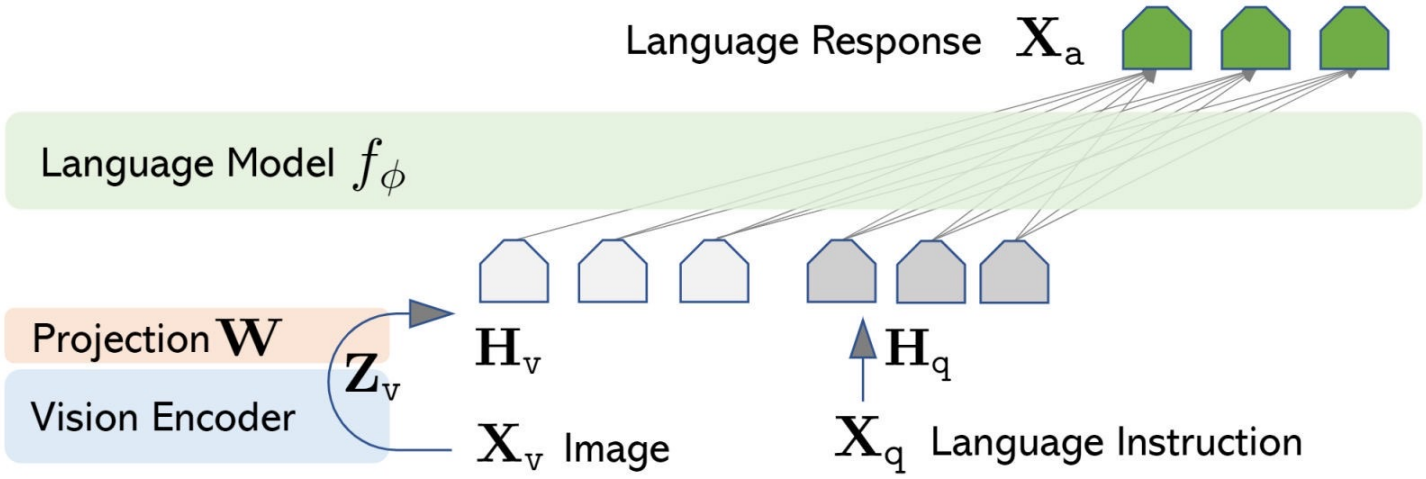
LLaVA: Architecture



The vision encoder converts input images into features
Linear projection layer convert these features into a space compatible with the LLM.

Liu et al., "Visual Instruction Tuning", NeurIPS 2023

LLaVA: Training



Two-stage Training

- **Stage 1: Pre-training for Vision-language Alignment**.

  Only the projection matrix is updated, based on a subset of CC3M.

- **Stage 2: Fine-tuning End-to-End.** Both the projection matrix and LLM are finetuned on curated dataset
  - **Visual Chat**: Generated multimodal instruction data for daily user-oriented applications.
  - **Science QA**: Multimodal reasoning dataset for the science domain.

Liu et al., "Visual Instruction Tuning", NeurIPS 2023

Strong Visual Reasoning Ability

Example 1: Extreme Ironing

Visual input example, Extreme Ironing:



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

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

Liu et al., "Visual Instruction Tuning", NeurIPS 2023

Strong Visual Reasoning Ability

Example 2: Chicken Nugget Map

Visual input example, Chicken Nugget Map:



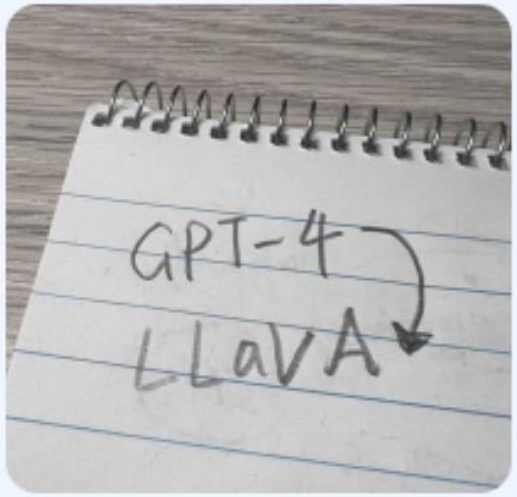
User	Can you explain this meme in detail?
LLaVA	The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all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

Liu et al., "Visual Instruction Tuning", NeurIPS 2023

Strong OCR Ability

Example 3: LLaVA

what is written in the photo?



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

Liu et al., “Visual Instruction Tuning”, NeurIPS 2023

Results

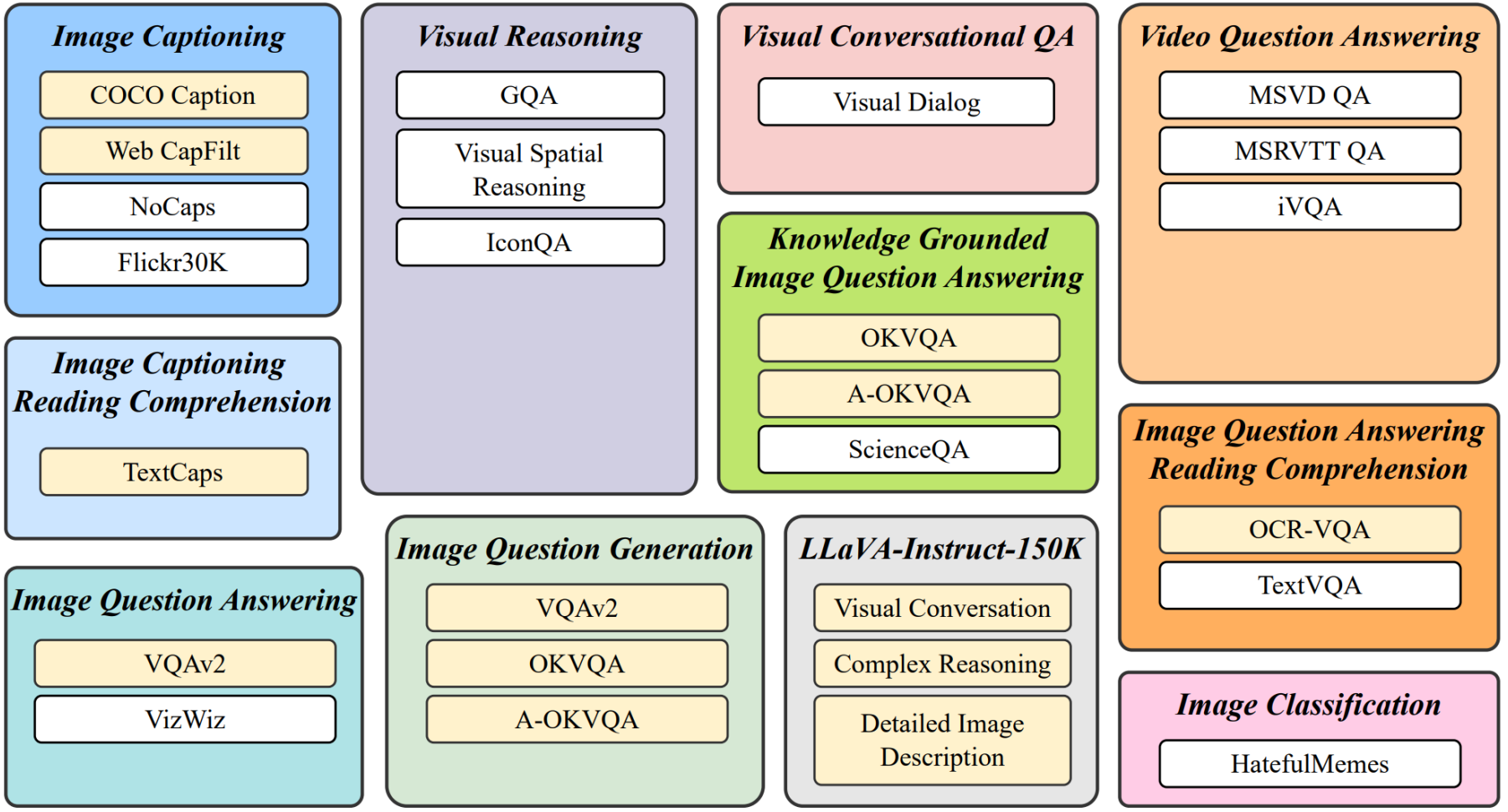
Method	Subject			Context Modality			Grade		Average
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Representative & SoTA methods with numbers reported in the literature									
Human [30]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [30]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [30]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
LLaMA-Adapter [55]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT$_{Base}$ [57]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{Large}$ [57]	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
Results with our own experiment runs									
GPT-4	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
LLaVA	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaVA+GPT-4 (complement)	90.36	95.50	88.55	89.05	87.80	91.08	92.22	88.73	90.97
LLaVA+GPT-4 (judge)	91.56	96.74	91.09	90.62	88.99	93.52	92.73	92.16	**92.53**

Table 6: Results (accuracy %) on Science QA dataset. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12.

Liu et al., “Visual Instruction Tuning”, NeurIPS 2023

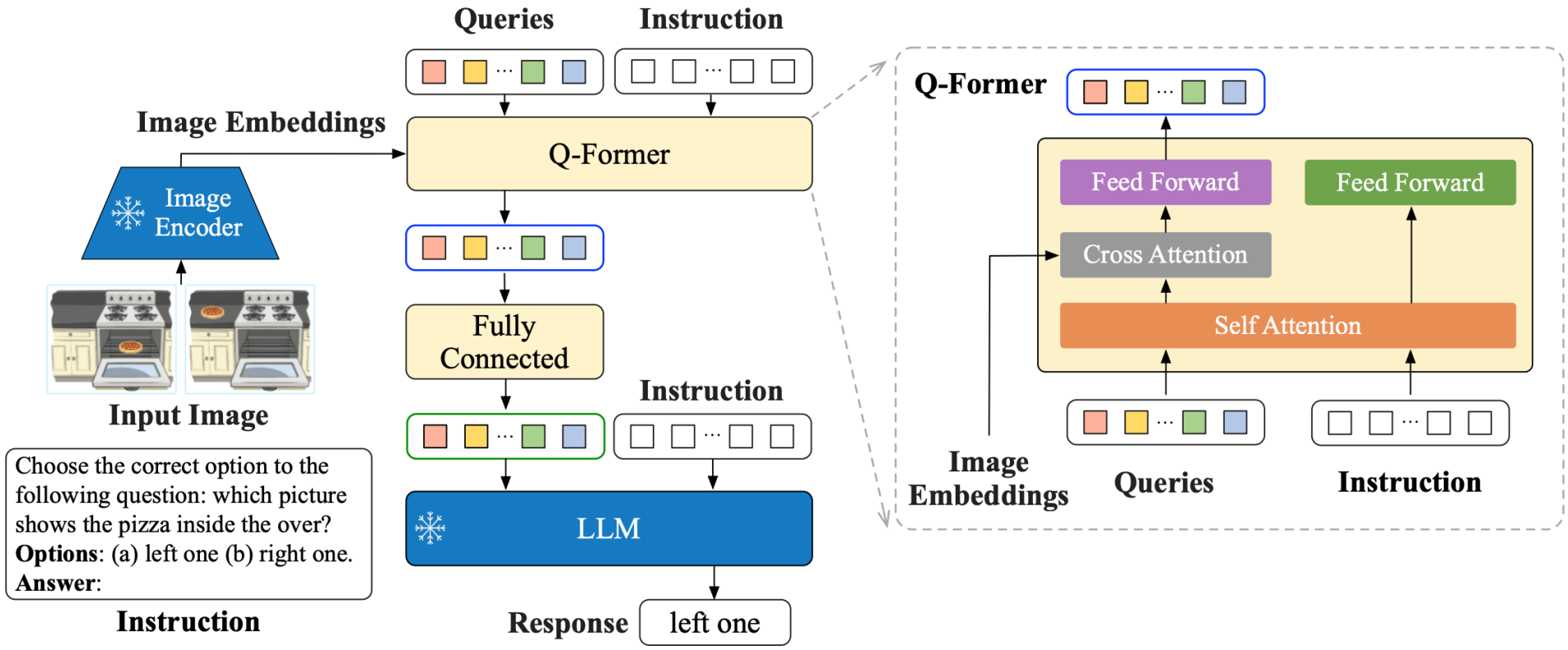
InstructBLIP

- Multitask Instruct with Established Datasets



Dai et al., InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, 2023

InstructBLIP: Architecture



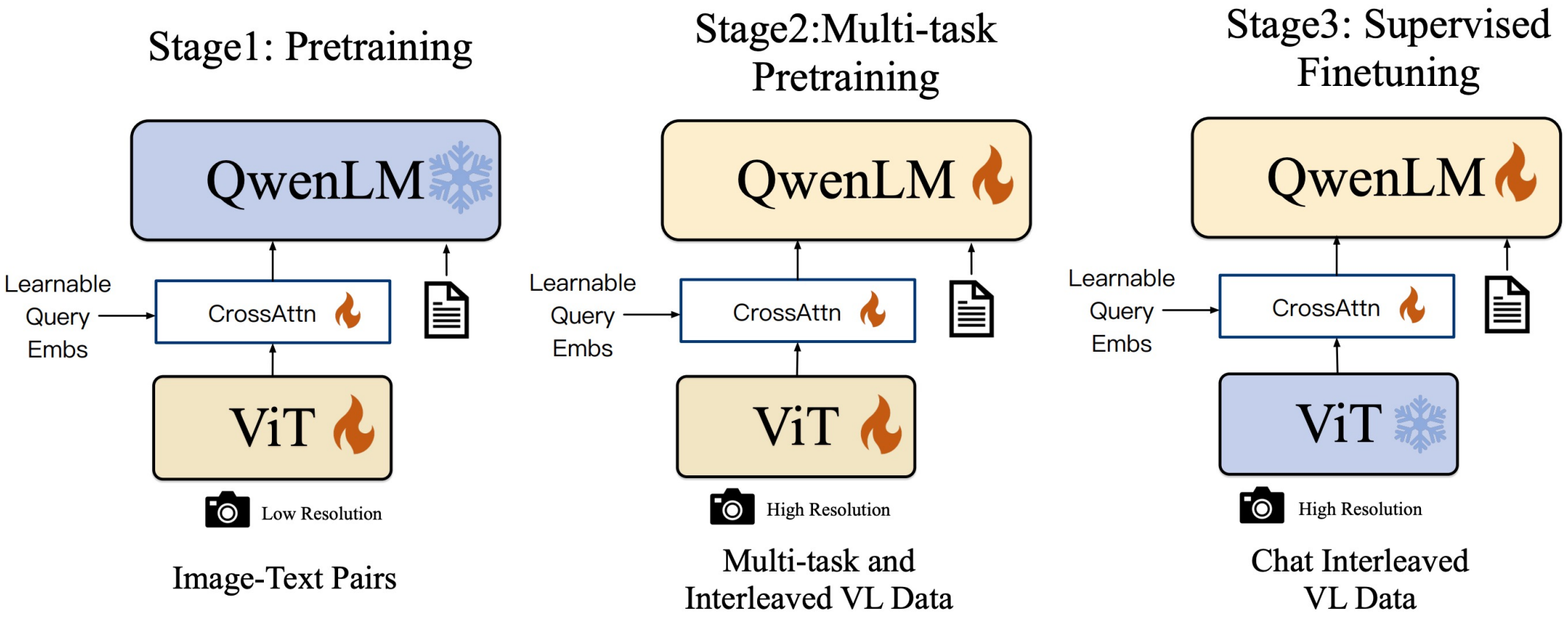
Dai et al., InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, 2023

Zero-shot results on the held-out datasets

	NoCaps	Flickr 30K	GQA	VSR	IconQA	TextVQA	Visdial	HM	VizWiz	SciQA image	MSVD QA	MSRVTT QA	iVQA
Flamingo-3B [4]	-	60.6	-	-	-	30.1	-	53.7	28.9	-	27.5	11.0	32.7
Flamingo-9B [4]	-	61.5	-	-	-	31.8	-	57.0	28.8	-	30.2	13.7	35.2
Flamingo-80B [4]	-	67.2	-	-	-	35.0	-	46.4	31.6	-	35.6	17.4	40.7
BLIP-2 (FlanT5$_{XL}$) [20]	104.5	76.1	44.0	60.5	45.5	43.1	45.7	53.0	29.8	54.9	33.7	16.2	40.4
BLIP-2 (FlanT5$_{XXL}$) [20]	98.4	73.7	44.6	68.2	45.4	44.1	46.9	52.0	29.4	64.5	34.4	17.4	45.8
BLIP-2 (Vicuna-7B)	107.5	74.9	38.6	50.0	39.7	40.1	44.9	50.6	25.3	53.8	18.3	9.2	27.5
BLIP-2 (Vicuna-13B)	103.9	71.6	41.0	50.9	40.6	42.5	45.1	53.7	19.6	61.0	20.3	10.3	23.5
InstructBLIP (FlanT5$_{XL}$)	119.9	**84.5**	48.4	64.8	50.0	46.6	46.6	56.6	32.7	70.4	43.4	25.0	53.1
InstructBLIP (FlanT5$_{XXL}$)	120.0	83.5	47.9	**65.6**	**51.2**	46.6	**48.5**	54.1	30.9	**70.6**	**44.3**	**25.6**	**53.8**
InstructBLIP (Vicuna-7B)	**123.1**	82.4	49.2	54.3	43.1	50.1	45.2	**59.6**	**34.5**	60.5	41.8	22.1	52.2
InstructBLIP (Vicuna-13B)	121.9	82.8	**49.5**	52.1	44.8	**50.7**	45.4	57.5	33.4	63.1	41.2	24.8	51.0

Dai et al., InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, 2023

Qwen-VL



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

Pretraining Data

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%

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

Results on Image Captioning & General VQA

Model Type	Model	Image Caption		General VQA				
		Nocaps (0-shot)	Flickr30K (0-shot)	VQAv2	OKVQA	GQA	SciQA-Img (0-shot)	VizWiz (0-shot)
Generalist Models	Flamingo-9B	-	61.5	51.8	44.7	-	-	28.8
	Flamingo-80B	-	67.2	56.3	50.6	-	-	31.6
	Unified-IO-XL	100.0	-	77.9	54.0	-	-	-
	Kosmos-1	-	67.1	51.0	-	-	-	29.2
	Kosmos-2	-	80.5	51.1	-	-	-	-
	BLIP-2 (Vicuna-13B)	103.9	71.6	65.0	45.9	32.3	61.0	19.6
	InstructBLIP (Vicuna-13B)	**121.9**	82.8	-	-	49.5	63.1	33.4
	Shikra (Vicuna-13B)	-	73.9	77.36	47.16	-	-	-
	Qwen-VL (Qwen-7B)	121.4	**85.8**	**79.5**	**58.6**	**59.3**	67.1	35.2
	Qwen-VL-Chat	120.2	81.0	78.2	56.6	57.5	**68.2**	**38.9**
Specialist SOTAs	-	127.0 (PALI-17B)	84.5 (InstructBLIP -FlanT5-XL)	86.1 (PALI-X -55B)	66.1 (PALI-X -55B)	72.1 (CFR)	92.53 (LLaVa+ GPT-4)	70.9 (PALI-X -55B)

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

Text-to-Image Generation

Impressive conditional diffusion models

- Text-to-image generation

DALLE 2

"a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese"



IMAGEN

"A photo of a raccoon wearing an astronaut helmet, looking out of the window at night."



Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents", arXiv 2022.
Saharia et al., "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding", arXiv 2022.

Impressive conditional diffusion models

Text-to-Image Generation

DALL·E 2

"a teddy bear on a skateboard in times square"



"Hierarchical Text-Conditional Image Generation with CLIP Latents" Ramesh et al., 2022

Imagen

A group of teddy bears in suit in a corporate office celebrating the birthday of their friend. There is a pizza cake on the desk.



"Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding", Saharia et al., 2022

Impressive conditional diffusion modelsv

Text-to-Image Generation

Stable Diffusion



[Stable Diffusion Applications: Twitter Mega Thread]

["High-Resolution Image Synthesis with Latent Diffusion Models" Rombach et al., 2022]

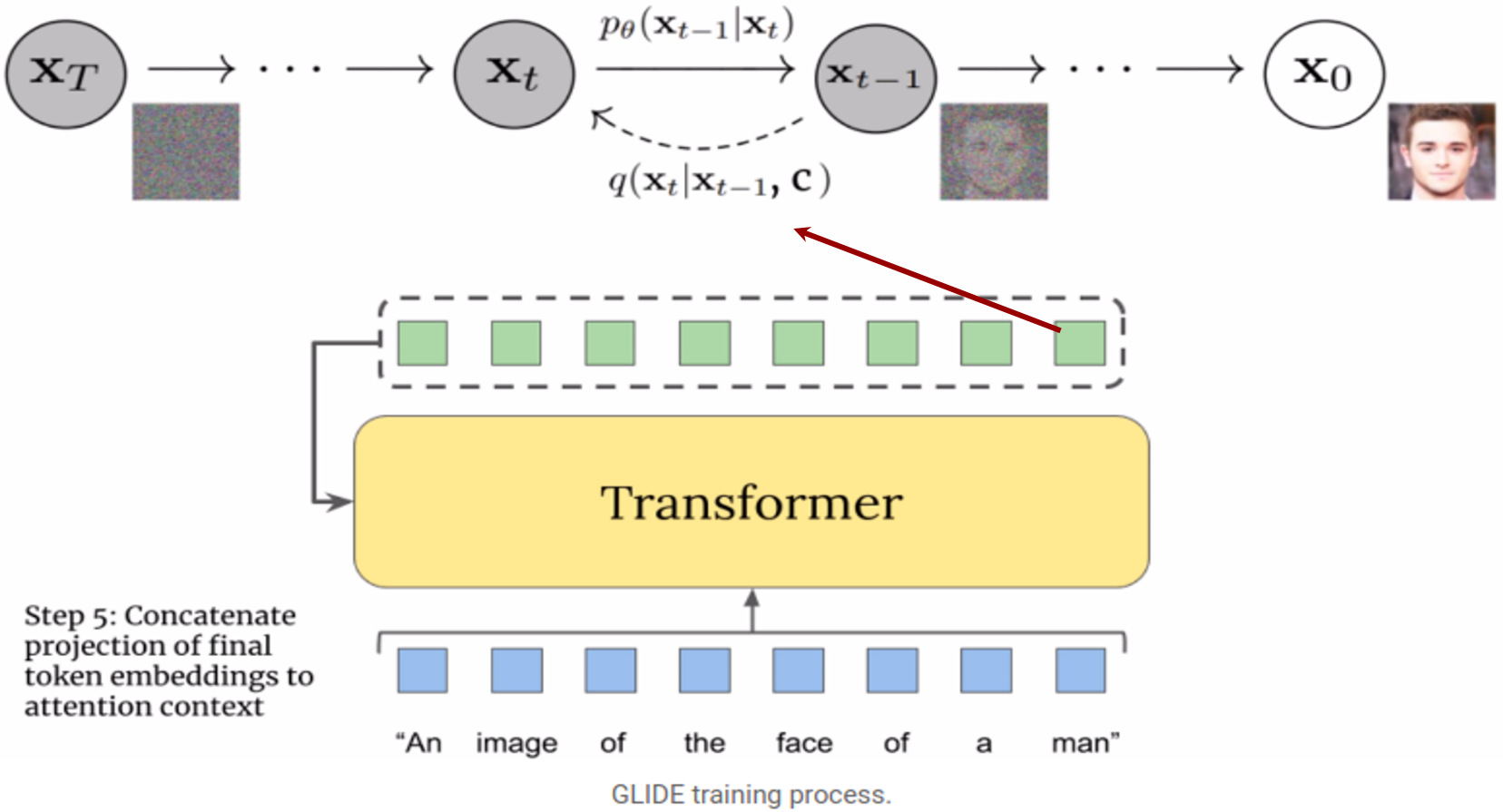
GLIDE

- **G**uided **L**anguage to **I**mage **D**iffusion for Generation and **E**diting
- Text-conditional image-generation using *diffusion models*
- Guidance strategies:
  - Classifier-free (superior in terms of photorealism and caption similarity)
  - CLIP
- Powerful tool for image inpainting: interactive image refinement!

Nichol et al., GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, 2021

GLIDE: Training process

- train a 3.5 billion parameter diffusion model that uses a text encoder to condition on natural language descriptions



GLIDE training process.

Nichol et al., GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, 2021



"a hedgehog using a calculator"



"a corgi wearing a red bowtie and a purple party hat"



"robots meditating in a vipassana retreat"



"a fall landscape with a small cottage next to a lake"



"a surrealist dream-like oil painting by salvador dalí of a cat playing checkers"



"a professional photo of a sunset behind the grand canyon"



"a high-quality oil painting of a psychedelic hamster dragon"



"an illustration of albert einstein wearing a superhero costume"

Nichol et al., GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, 2021

GLIDE: Classifier-free Guidance

- It allows a single model to leverage its knowledge during guidance
  - Instead of relying on a separate classifier
- It simplifies guidance for other conditions like text

$$\hat{\epsilon}_\theta(x_t|c) = \epsilon_\theta(x_t|\emptyset) + s \cdot (\epsilon_\theta(x_t|c) - \epsilon_\theta(x_t|\emptyset))$$

Nichol et al., GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, 2021

GLIDE: CLIP Guidance

- Instead of $\nabla_{x_t} \log p(y|x_t)$, it uses $\nabla_{x_t} f(x_t).\, g(c)$ where f is the image encoder and g is the text encoder

- Guiding using the public CLIP model adversely impacts sample quality
  - the noised intermediate images encountered during sampling are out-of-distribution for the model

- Noised CLIP is proposed

Nichol et al., GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, 2021

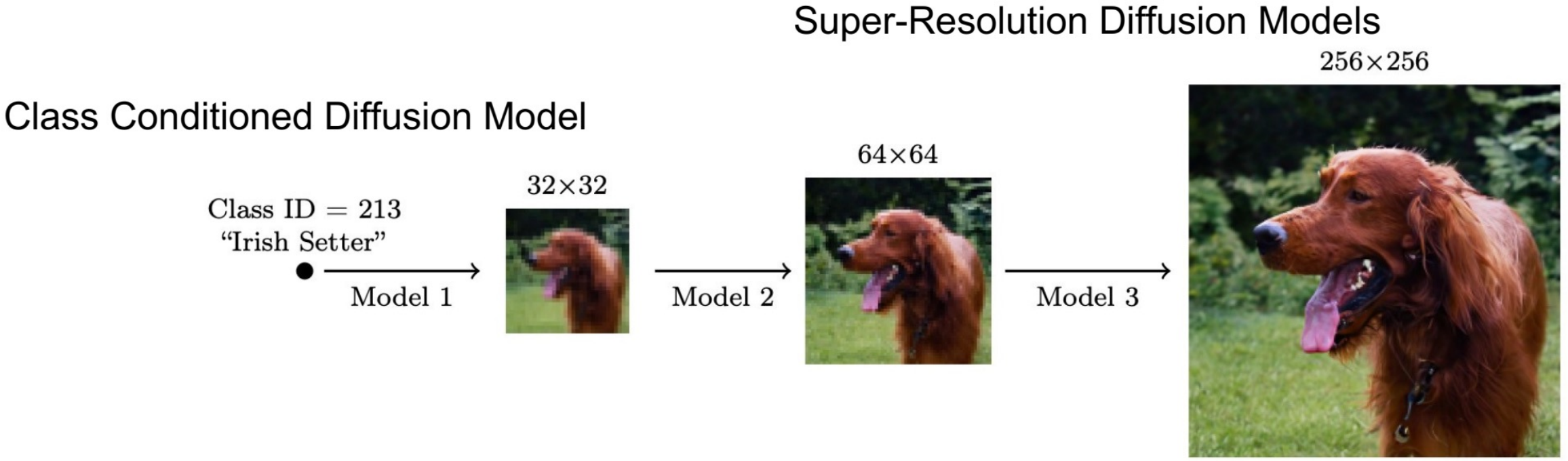
GLIDE: Architecture

- Text encoding transformer
- Base model: a diffusion model for low resolution image generation conditioned on the text encoding
- Supper-resolution model: a diffusion model for high resolution image generation conditioned on the low-resolution image and the text encoding

Nichol et al., GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, 2021

Cascaded generation

- Pipeline

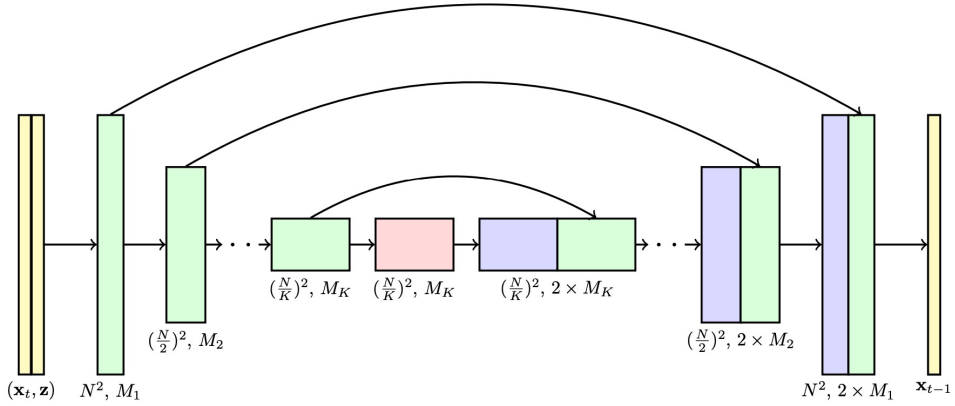


$p_\theta(x_{t-1}|x_t, \tilde{x})$ additionally conditions on the downsampled input $\tilde{x}$ (bicubic upsampled version of $\tilde{x}$ is concatenated in the channel dimension)

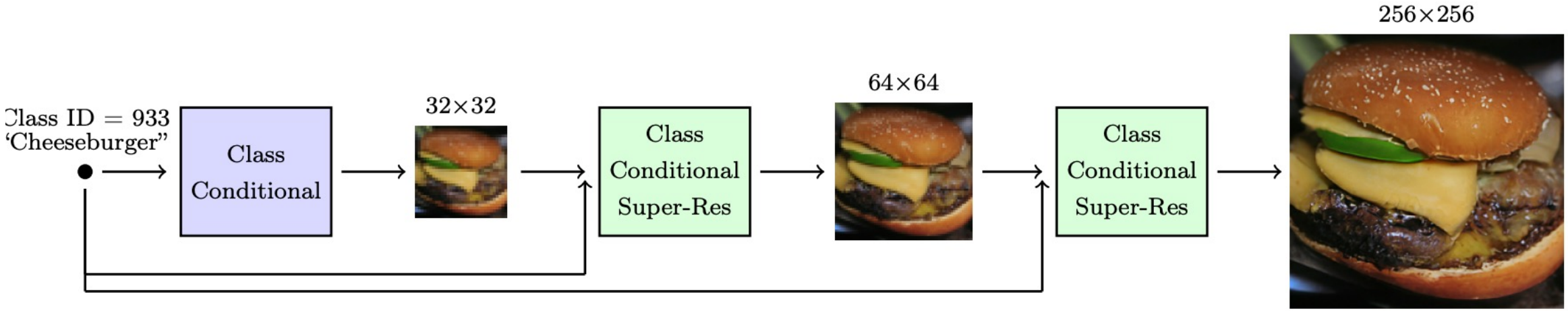
Similar cascaded / multi-resolution image generation also exist in GAN (Big-GAN & StyleGAN)

Ho et al., Cascaded Diffusion Models for High Fidelity Image Generation, 2021

Cascaded generation: Super-resolution



- Conditioning Augmentation: training each super-resolution model using data augmentation on its low resolution input.



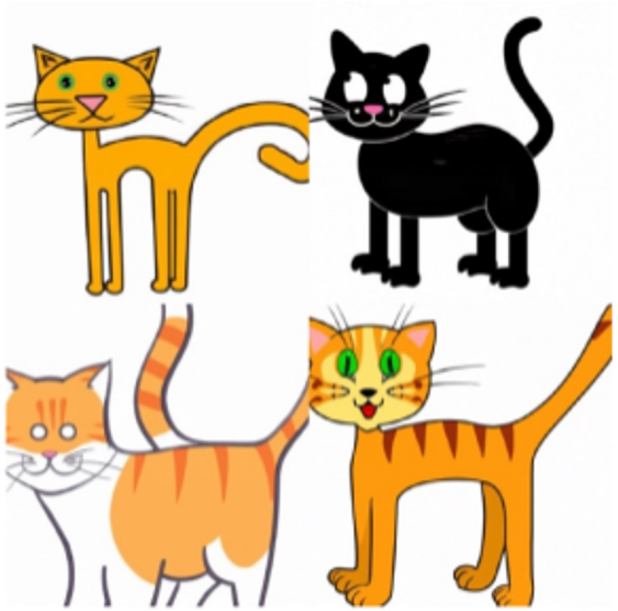
Ho et al., Cascaded Diffusion Models for High Fidelity Image Generation, 2021



Nichol et al., GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, 2021

GLIDE: Failure Cases

- When prompted with unusual objects or scenarios



"an illustration of a cat that has eight legs"



"a bicycle that has continuous tracks instead of wheels"



"a mouse hunting a lion"



"a car with triangular wheels"

Nichol et al., GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, 2021

Image inpainting

- Fine-tune model to perform inpainting
- 4 additional channels (including RGB of remaining portions along with the mask) are fed as additional conditioning information
- For upsampling, a full low-resolution is fed but unmasked regions of high-resolution image



Nichol et al., GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, 2021



"zebras roaming in the field"



"a girl hugging a corgi on a pedestal"



"a man with red hair"



"an old car in a snowy forest"

Nichol et al., GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, 2021

OpenAI DALL-E 2



an espresso machine that makes coffee from human souls, artstation



panda mad scientist mixing sparkling chemicals, artstation



a corgi's head depicted as an explosion of a nebula

A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

OpenAI DALL-E 2



a dolphin in an astronaut suit on saturn, artstation



a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese



a teddy bear on a skateboard in times square

A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

OpenAI DALL-E 2



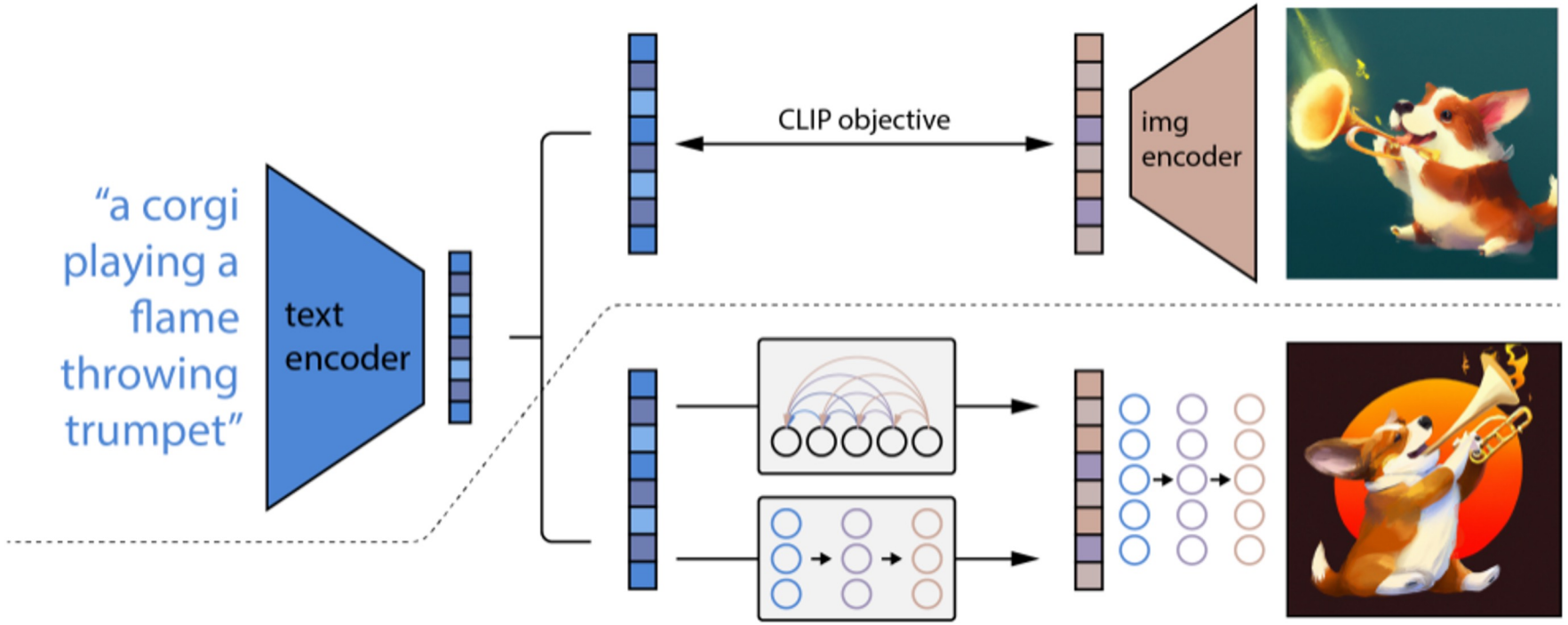
Teddy bears shopping for groceries in ancient egypt



What does the OpenAI office look like?

A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

OpenAI DALL-E 2



A prior to generate possible CLIP image embeddings from a given text caption

a diffusion decoder to invert the CLIP image encoder

- CLIP text embedding is first prepared (from a frozen CLIP)
- It is fed to a diffusion (or autoregressive) prior to produce an image embedding
- This embedding is used to condition a diffusion decoder which produces image

A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

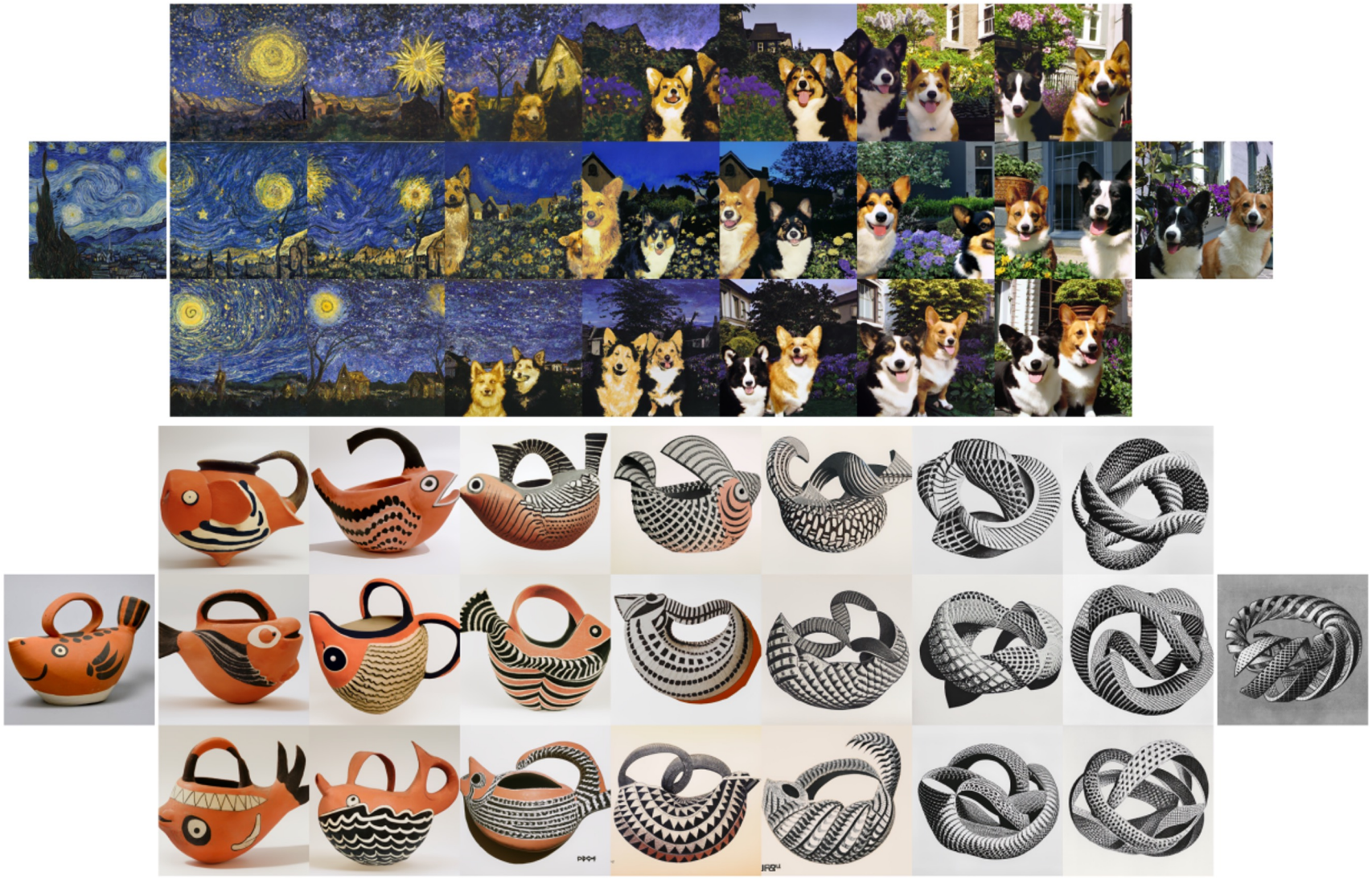
OpenAI DALL-E 2

- A decoder (of CLIP image embedding) is combined with a prior model
  - The produced image embedding is used to condition the diffusion decoder
- Decoder is non-deterministic
  - can produce multiple images corresponding to a given image embedding

- The presence of an encoder and its approximate inverse (the decoder) allows for capabilities beyond text-to-image translation
  - E.g., using the CLIP latent space provides the ability to semantically modify images by moving in the direction of any encoded text vector

A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

Image manipulations: Interpolations



A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

Encoding with clip + decoding with a diffusion model



A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

Probing the CLIP Latent Space



Figure 7: Visualization of reconstructions of CLIP latents from progressively more PCA dimensions (20, 30, 40, 80, 120, 160, 200, 320 dimensions), with the original source image on the far right. The lower dimensions preserve coarse-grained semantic information, whereas the higher dimensions encode finer-grained details about the exact form of the objects in the scene.

A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

Image manipulations: Text diffs

$$z_d = \text{norm}(z_t - z_{t_0}) \qquad z_\theta = \text{slerp}(z_i, z_d, \theta)$$



a photo of a cat → an anime drawing of a super saiyan cat, artstation



a photo of a victorian house → a photo of a modern house

A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

“A green vase filled with red roses sitting on top of table.”



A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

Quantitative analysis

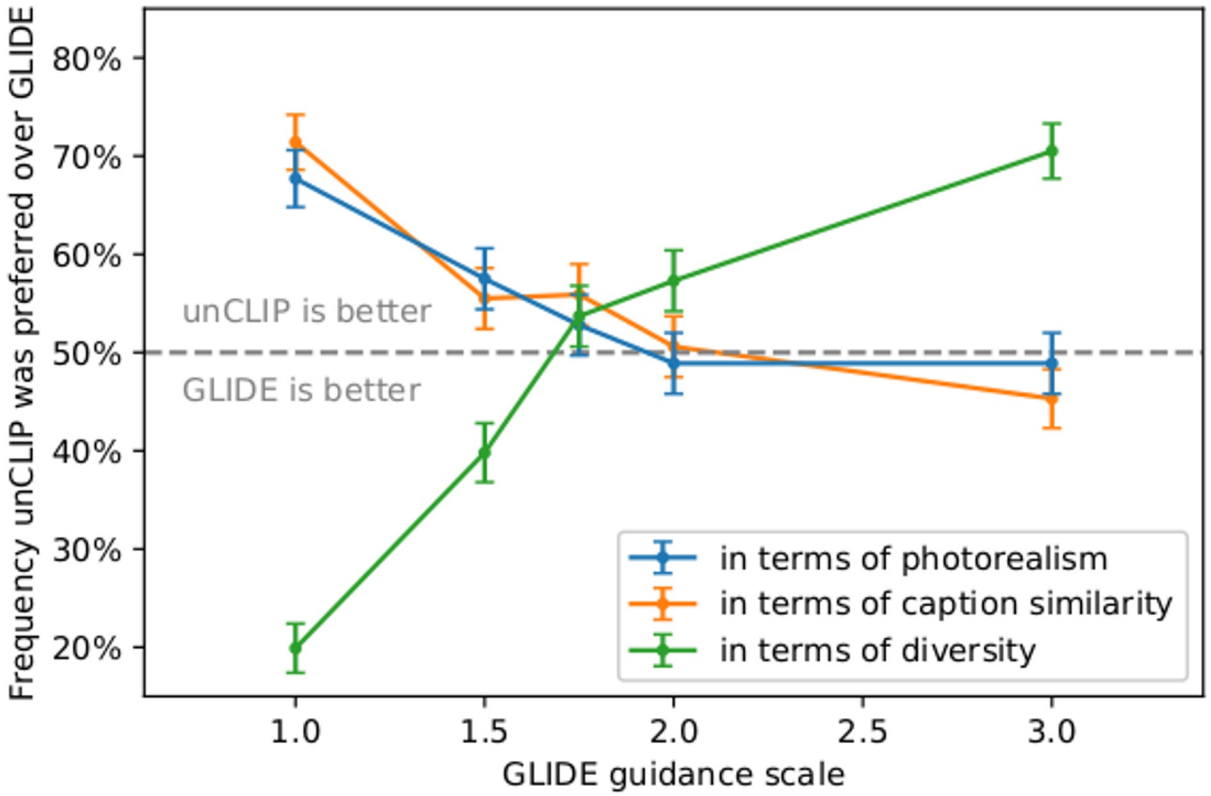


Figure 10: When comparing unCLIP (with our best sampling settings) to various settings of guidance scale for GLIDE, unCLIP was preferred by human evaluators on at least one axis among photorealism, caption similarity, and diversity for each comparison. At the higher guidance scales used to generate photorealistic images, unCLIP yields greater diversity for comparable photorealism and caption similarity.

A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

Quantitative analysis

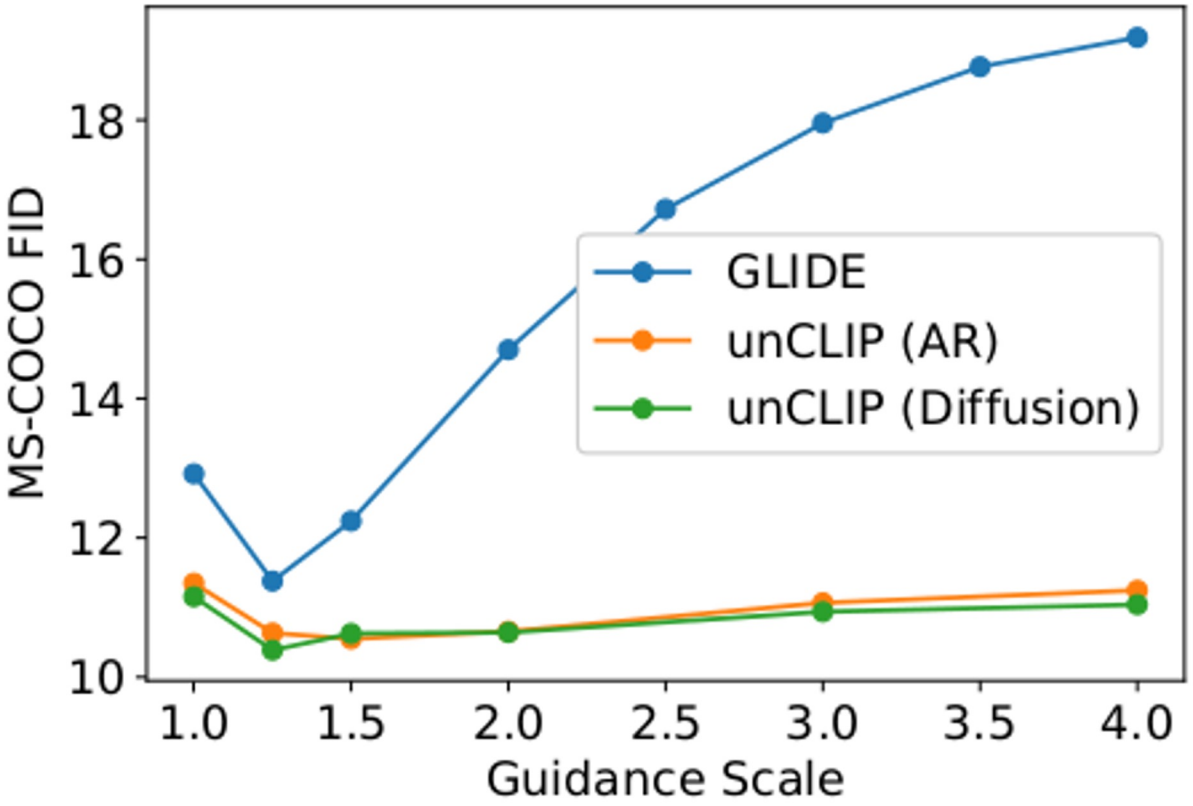
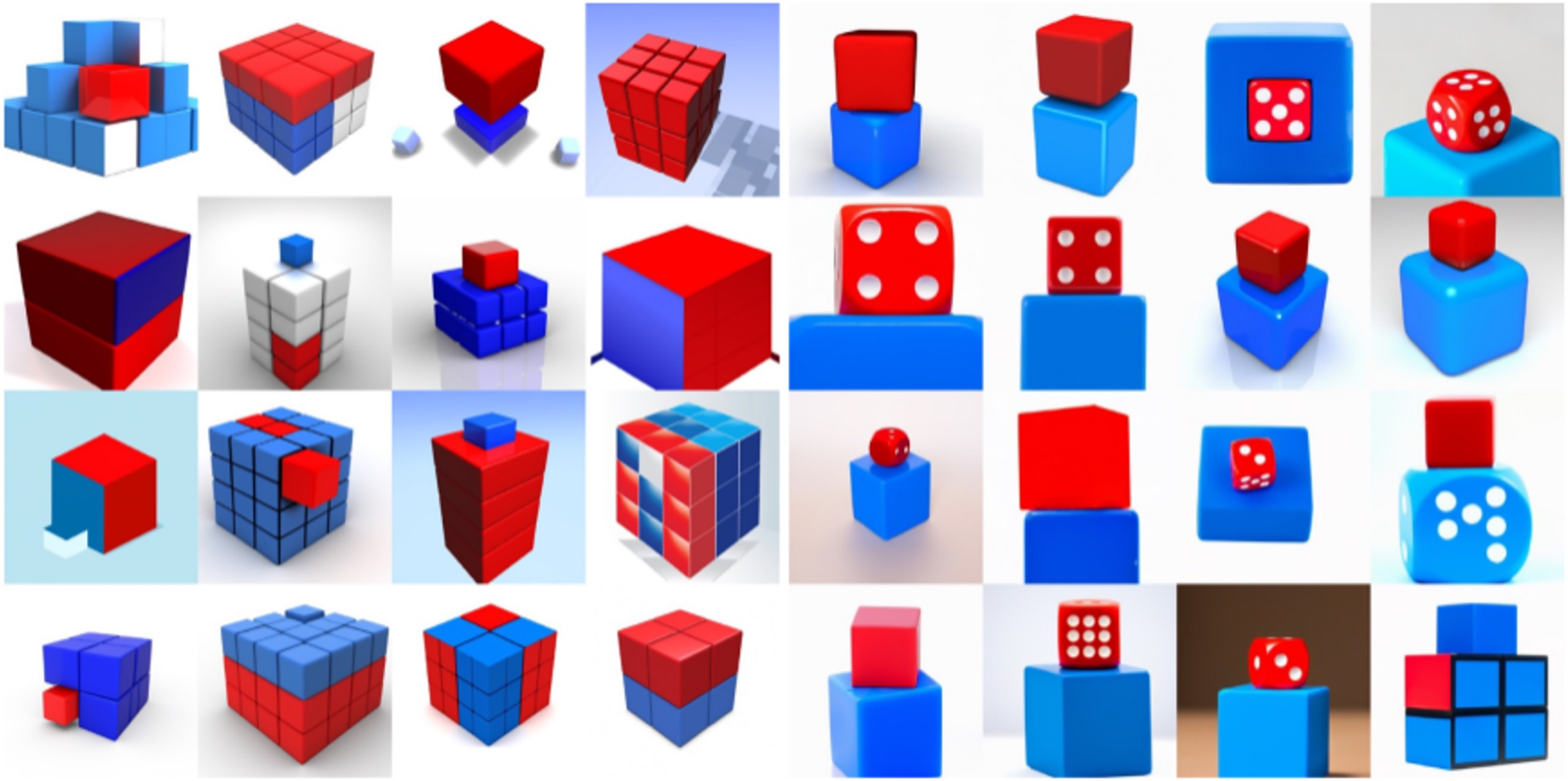


Figure 11: FID versus guidance scale for unCLIP and GLIDE. For the unCLIP priors, we swept over sampling hyperparameters and fixed to the settings with the best minimum FID.

A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

Failure cases

“A red cube on the top of a blue cube.”



(a) unCLIP (b) GLIDE

A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022



Figure 15: Reconstructions from the decoder for difficult binding problems. We find that the reconstructions mix up objects and attributes. In the first two examples, the model mixes up the color of two objects. In the rightmost example, the model does not reliably reconstruct the relative size of two objects.



Figure 16: Samples from unCLIP for the prompt, "A sign that says deep learning."

A. Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

Architecture of DALL-E 2 & Imagen

- Pixel-based Diffusion (No encoder-decoder)
- pre-trained text encoder (CLIP, T5)
- Diffusion model + classifier-free guidance
- Cascaded models: 64->128->512





https://cdn.openai.com/papers/dall-e-2.pdf

https://arxiv.org/abs/2205.11487

DALL-E 3

- Using a Large Language Model to encode the prompt, and **directly** condition the image generation process

- DALL-E 3 represents a leap forward in the ability to generate images that exactly adhere to the text
  - avoid the tendency to ignore words or descriptions

Stable Diffusion

- A diffusion-based model for generating high-quality, diverse images from text prompts.
- Built on the Latent Diffusion Model (LDM) framework.
- Performs the diffusion process in a compressed latent space rather than pixel space, significantly reducing memory and computation.
- Widely adopted for tasks beyond generation, including inpainting and image editing.

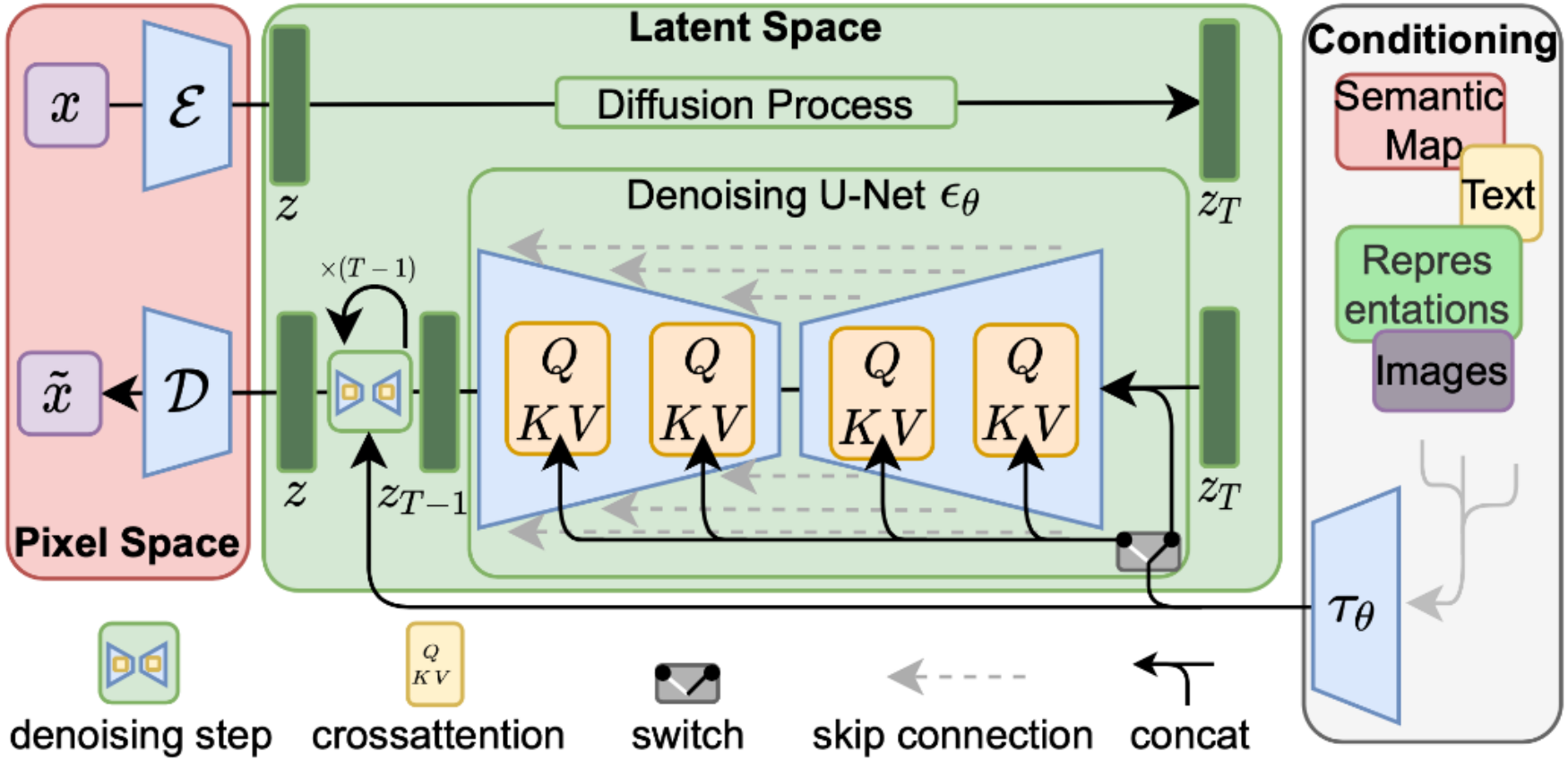
Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, CVPR 2022

Latent Space Diffusion Models

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \cdot \mathbf{V}$$

$$\text{where } \mathbf{Q} = \mathbf{W}_Q^{(i)} \cdot \varphi_i(\mathbf{z}_i),\ \mathbf{K} = \mathbf{W}_K^{(i)} \cdot \tau_\theta(y),\ \mathbf{V} = \mathbf{W}_V^{(i)} \cdot \tau_\theta(y)$$

$$\text{and } \mathbf{W}_Q^{(i)} \in \mathbb{R}^{d \times d_\epsilon^i},\ \mathbf{W}_K^{(i)}, \mathbf{W}_V^{(i)} \in \mathbb{R}^{d \times d_\tau},\ \varphi_i(\mathbf{z}_i) \in \mathbb{R}^{N \times d_\epsilon^i},\ \tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$$



Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", CVPR 2022

Stable Diffusion: Main Components

- VAE (Encoder + Decoder)
- Text Encoder
- Denoiser Network
- Scheduler

Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, CVPR 2022

Stable Diffusion: VAE

- The VAE encoder maps images to low-dimensional latent space, enabling efficient forward process with reduced memory/compute.
- The VAE decoder reconstructs images from noise-free latent representations.
- Both encoder and decoder are frozen during training.
- During inference, only the decoder is used to convert denoised latents back to images.

Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, CVPR 2022

Stable Diffusion: Text Encoder

- Maps input tokens to dense text embeddings using a frozen transformer-based encoder (typically CLIP).
- Remains frozen during training to leverage pre-trained language understanding.
- Text embeddings are fed into the U-Net denoiser via cross-attention, guiding the image generation process.
  - Enables prompt-conditioned generation, aligning outputs with user instructions.

Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, CVPR 2022

Stable Diffusion: Denoiser Network

- Operates in latent space to progressively remove noise.
- Conditioned on text embeddings via cross-attention to align generation with prompts.
- Uses a U-Net architecture with residual blocks and attention mechanisms.
  - Receives as input: noisy latent, timestep, and text embeddings.
  - Outputs a denoised latent, which is refined over multiple steps.

Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, CVPR 2022

Stable Diffusion: Scheduler

- Determines the noise schedule across timesteps (e.g., linear, cosine, or learnable schedules).
- Provides the timestep embedding input to the U-Net during training and inference.
- Used in both training (forward + reverse) and inference (reverse only).
- A training-free component

Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, CVPR 2022

Stable Diffusion: Training Configurations

- **Dataset**: LAION-5B, a large-scale text-image pair dataset.
- **Image Resolution**: 512×512 or 1024×1024
- Latent Dimension: Usually 64×64 or 32×32
- **Text Encoder:** Frozen CLIP (ViT-L/14)
- **Loss Function**: Mean Squared Error (MSE) between predicted and true noise.

Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, CVPR 2022

Stable Diffusion: Examples

a man that is on a surfboard with a dog



a person in red chopping green apples under an umbrella



a splash dog chasing a frisbee into a pool during a performance



A person painting colorful images on a vase



Text-to-Image Models: Challenges

- **Compositional Generation:** Struggles to accurately render prompts with multiple distinct objects, complex scenes, or spatial relationships
  - e.g., “a red cube on a blue sphere next to a cat”
- **Bias:** May reflect biases from training data and text encoder
- **Prompt Sensitivity:** Small changes in phrasing can lead to drastically different results

Text to Image Generation: Summary

- Diffusion-based models (Classifier-guided)
  - GLIDE
  - DALL-E2
- Latent space diffusion models
  - Stable Diffusion
    - Cross-attention on embedding of text tokens