

Some Recent Image-Text Models

M. Soleymani

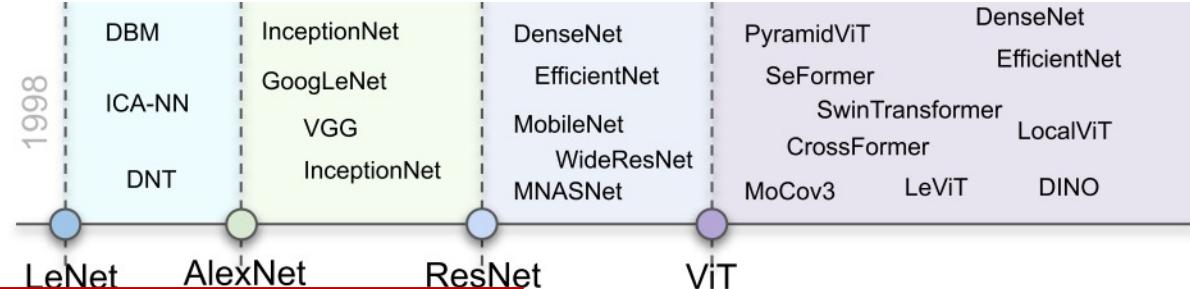
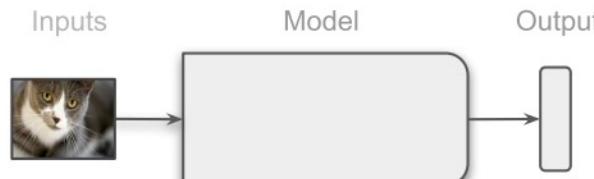
Sharif University of Technology

Spring 2024

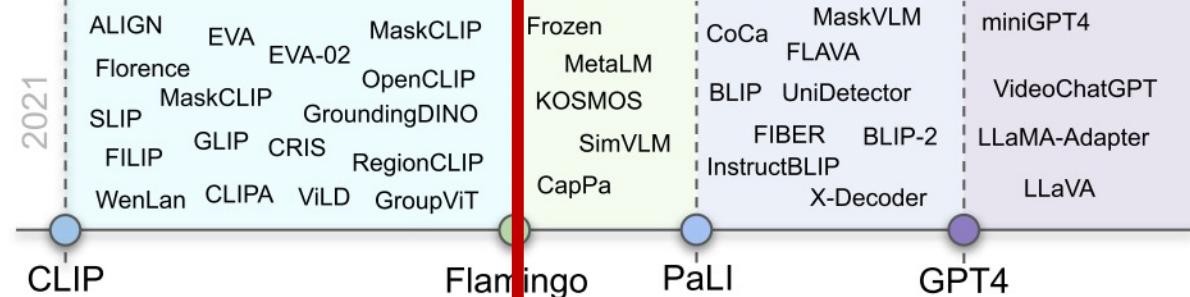
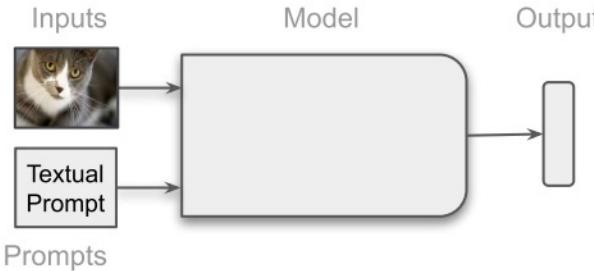
Multi-modal data

- 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)

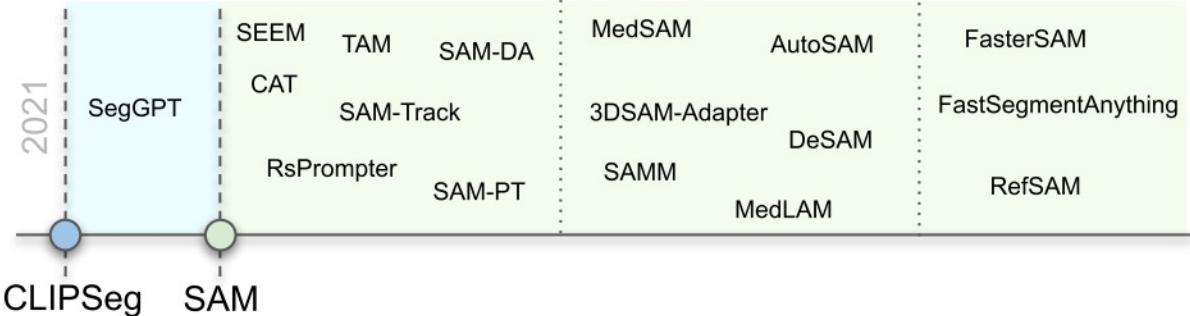
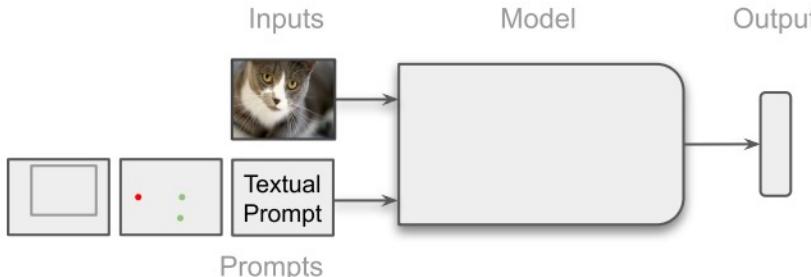
Traditional Models



Textually Prompted Models



Visually Prompted Models

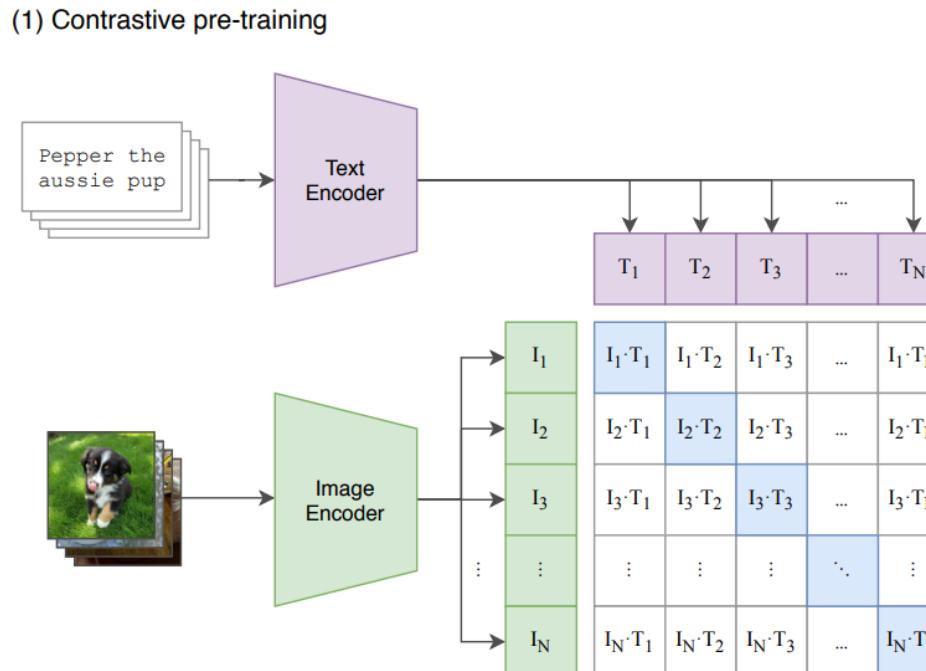


Vision-language models: Contrastive learning

- 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

CLIP (OpenAI)

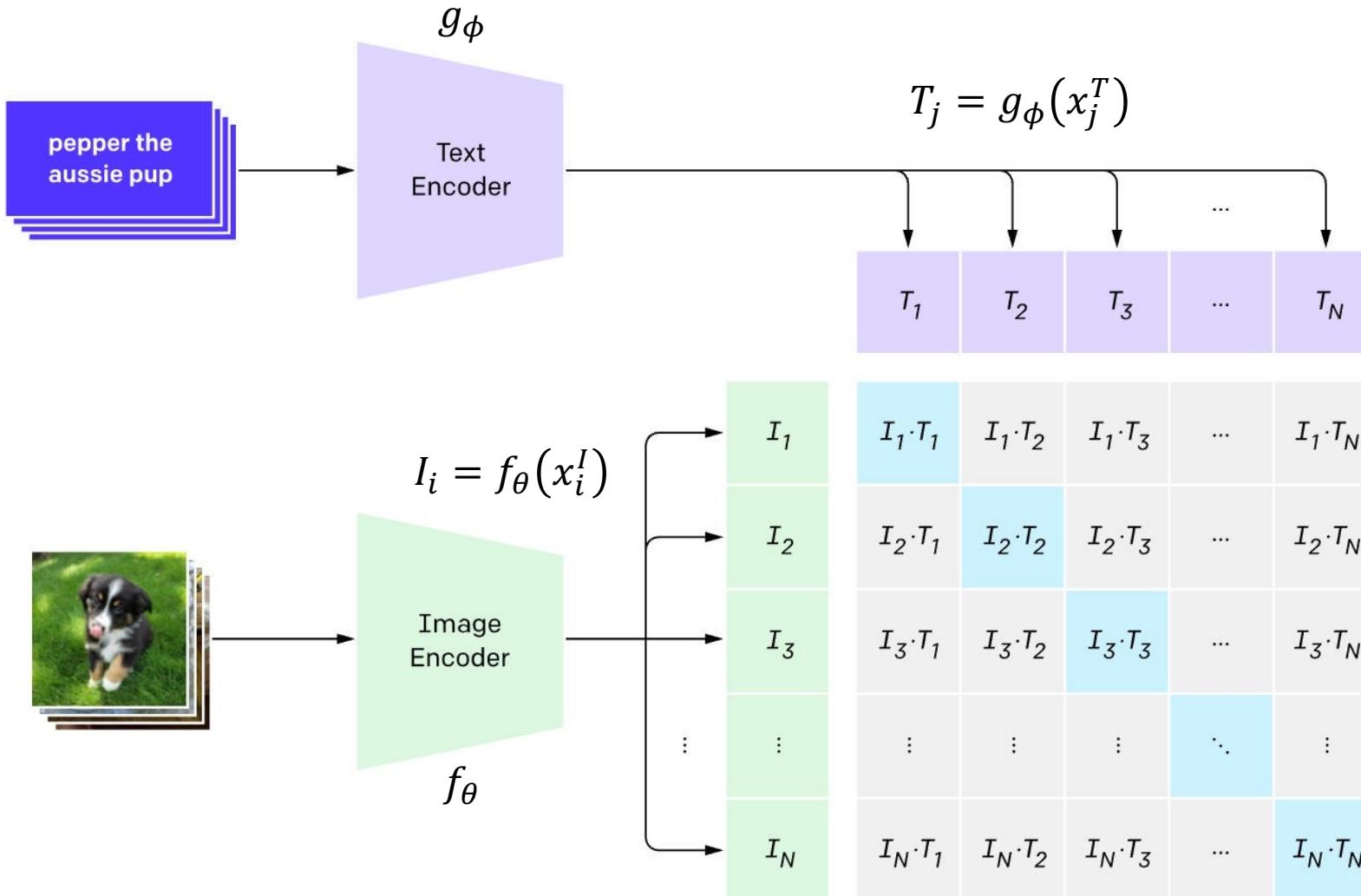
- Caption as a weak supervision
- Learns a multi-modal embedding space by jointly training an image encoder and text encoder
 - maximize the cosine similarity of the image and text embeddings of the N real pairs in the batch while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect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



Positive pairs:
Images and corresponding texts

Negative pairs:
Random pairs

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

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

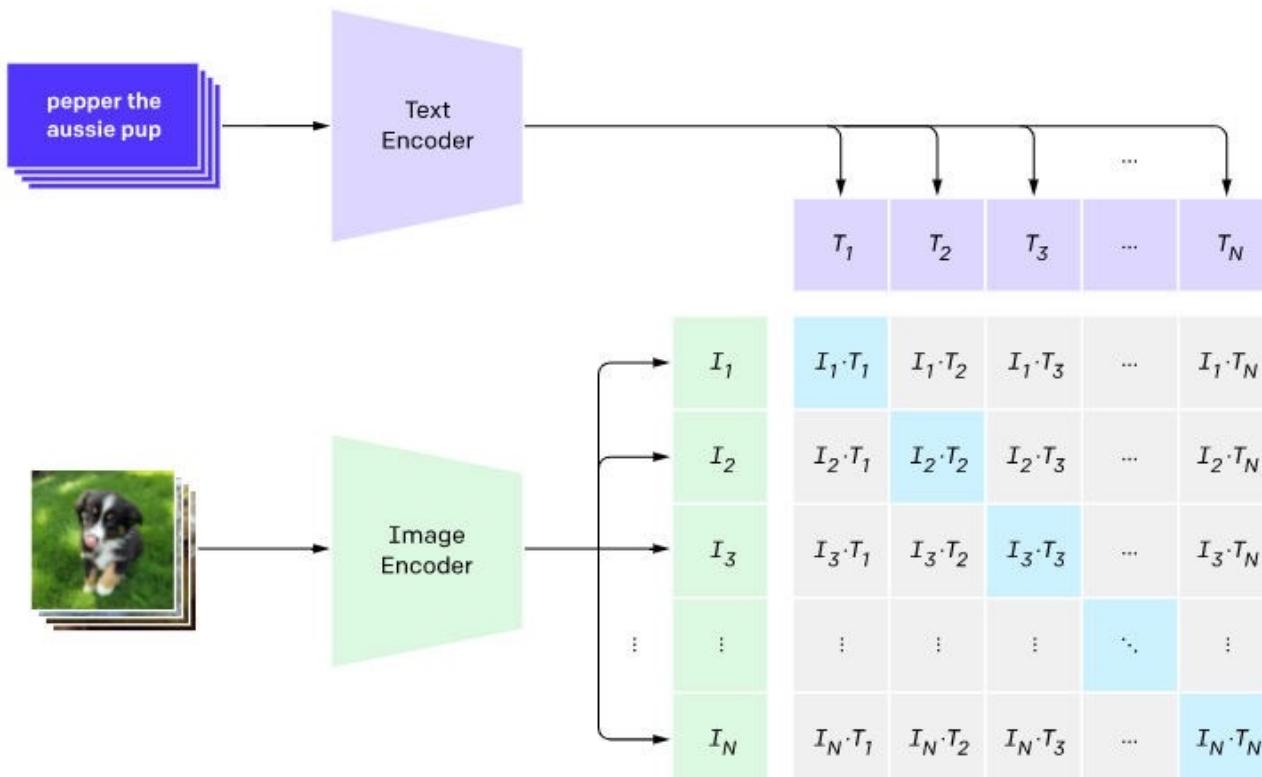
# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# 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)

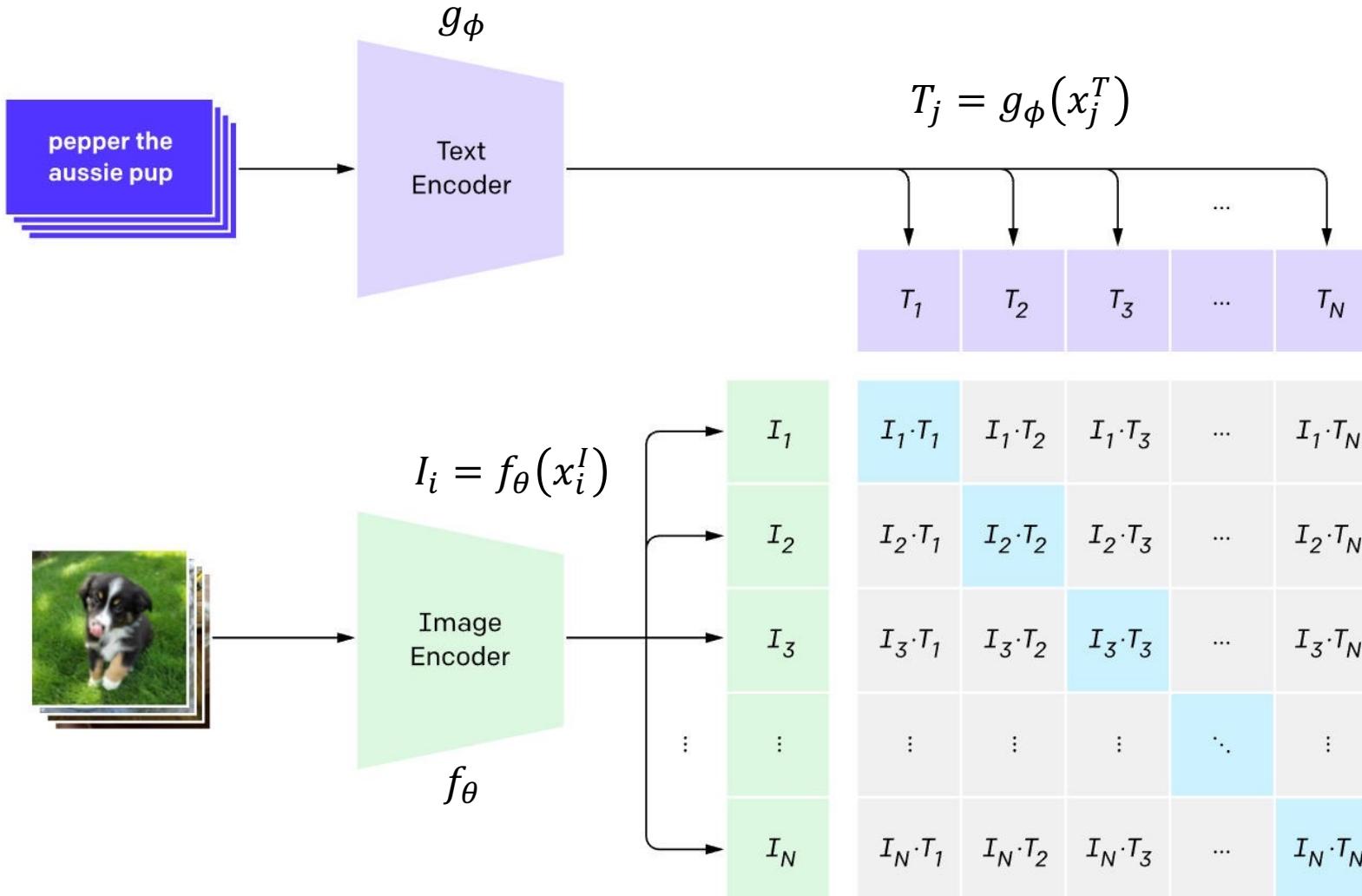
# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

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

1. Contrastive pre-training



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

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

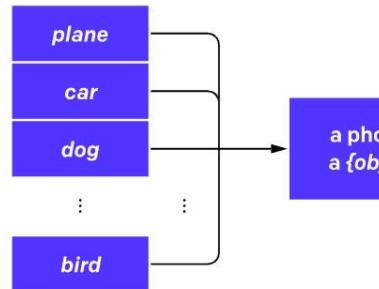
$$\mathcal{L}_j^T = -\log \frac{e^{s_{i,i}^T}}{\sum_{i=1}^N e^{s_{i,j}^T}}$$

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

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

CLIP for zero-shot learning

2. Create dataset classifier from label text



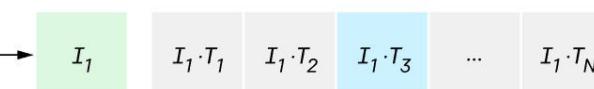
Text Encoder

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

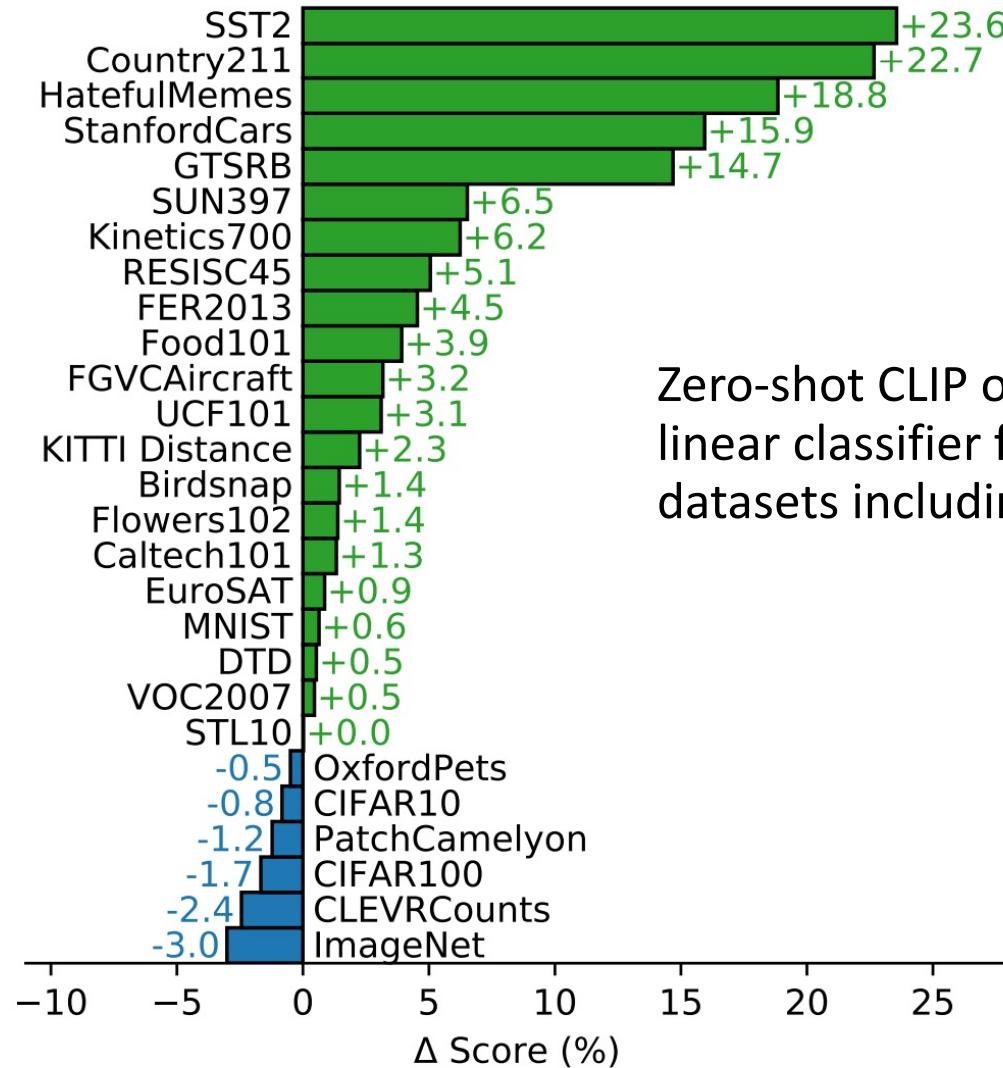
3. Use for zero-shot prediction



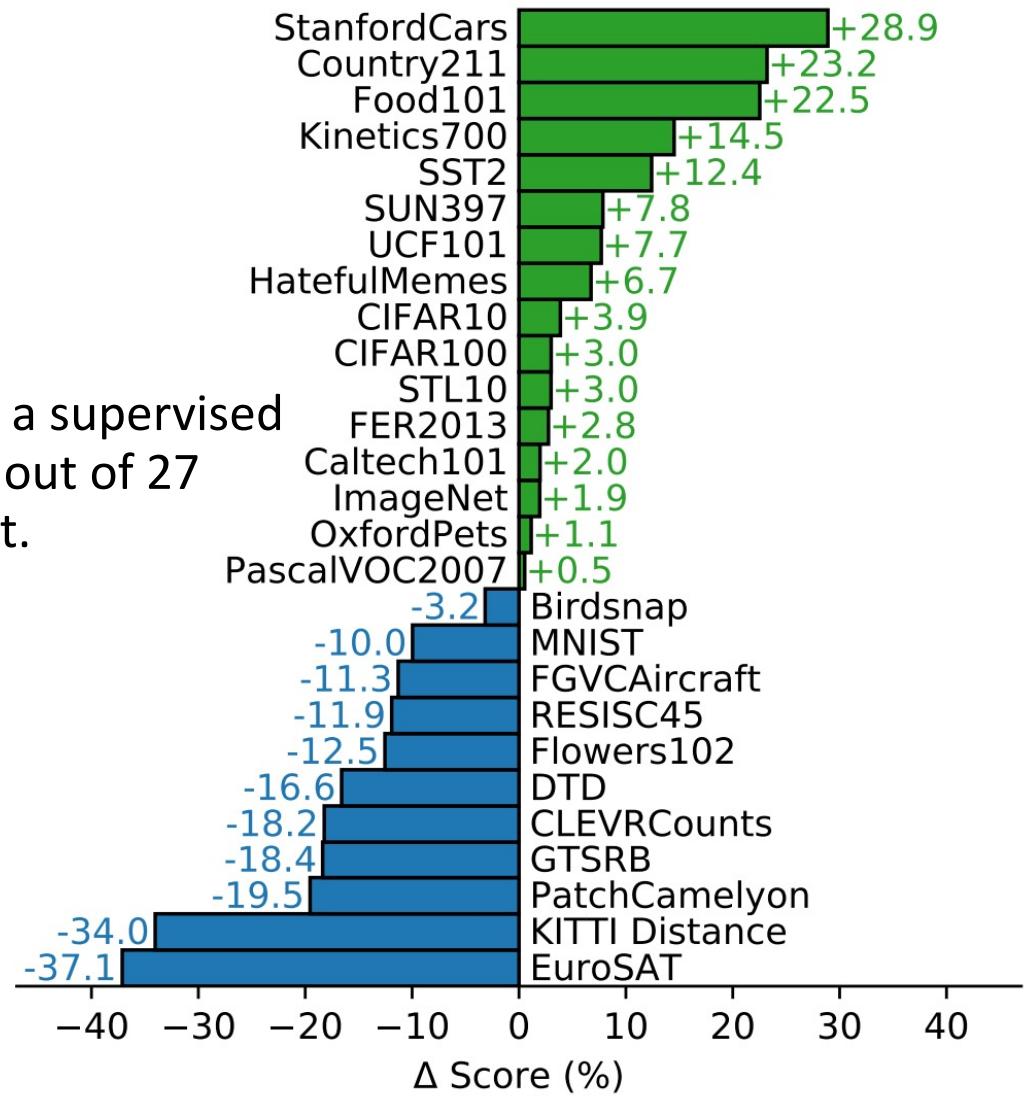
Image Encoder



a photo of a dog.



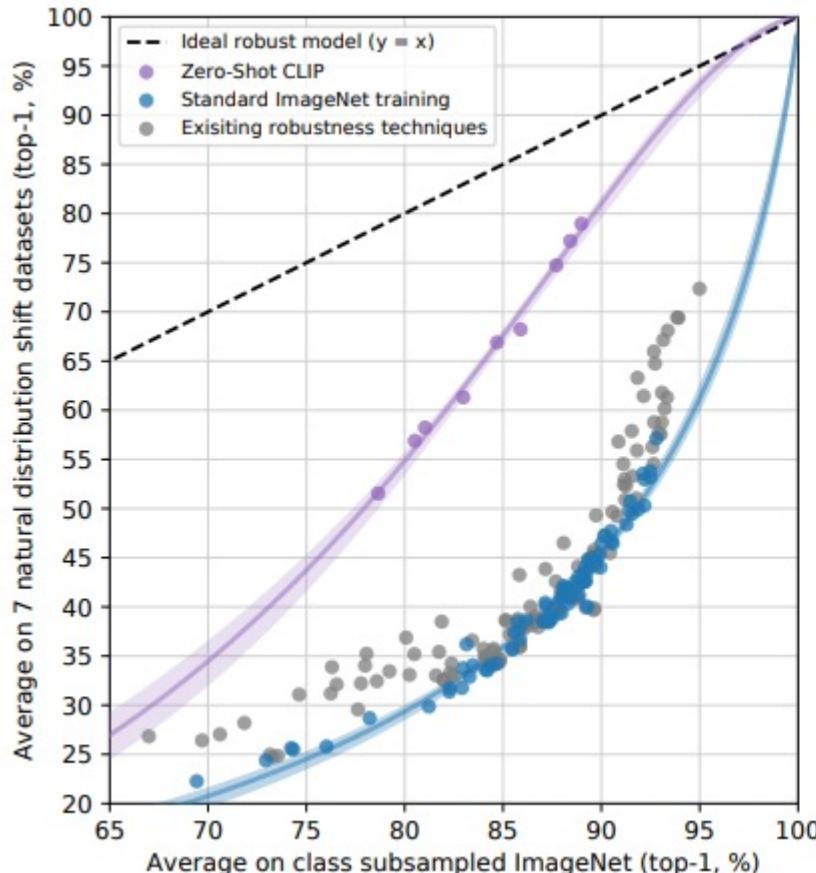
Zero-shot CLIP outperforms a supervised linear classifier fitted on 16 out of 27 datasets including ImageNet.



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

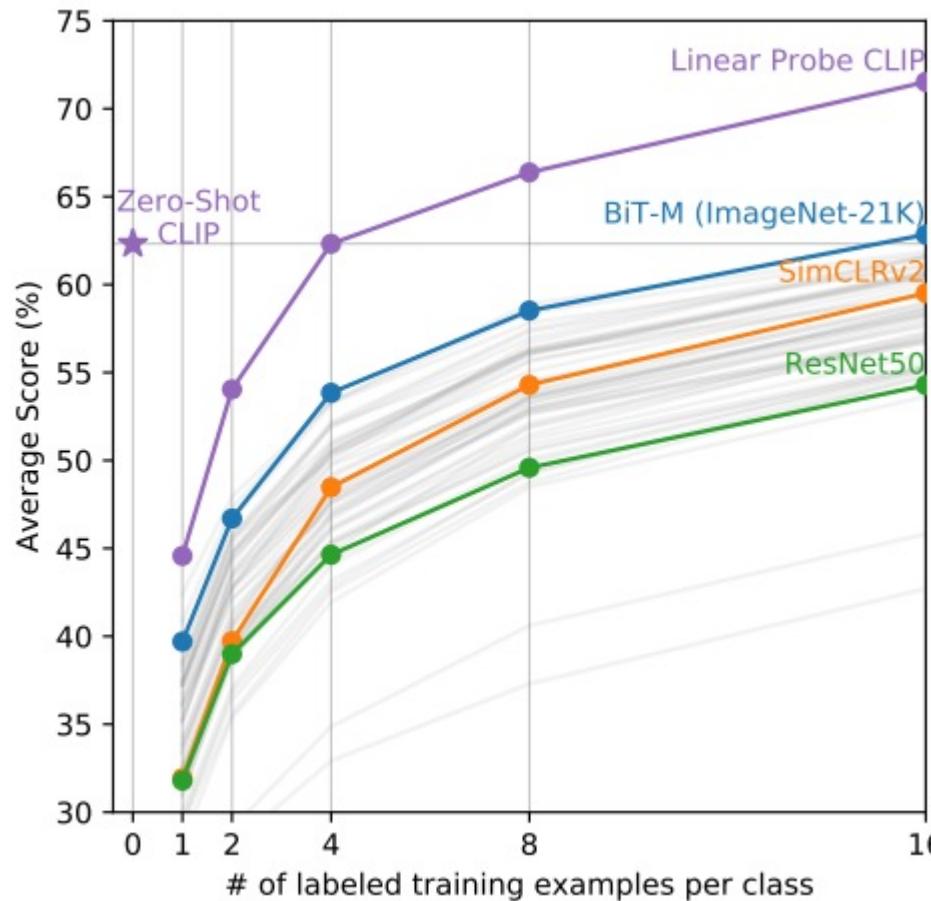
Robustness to distribution shift

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

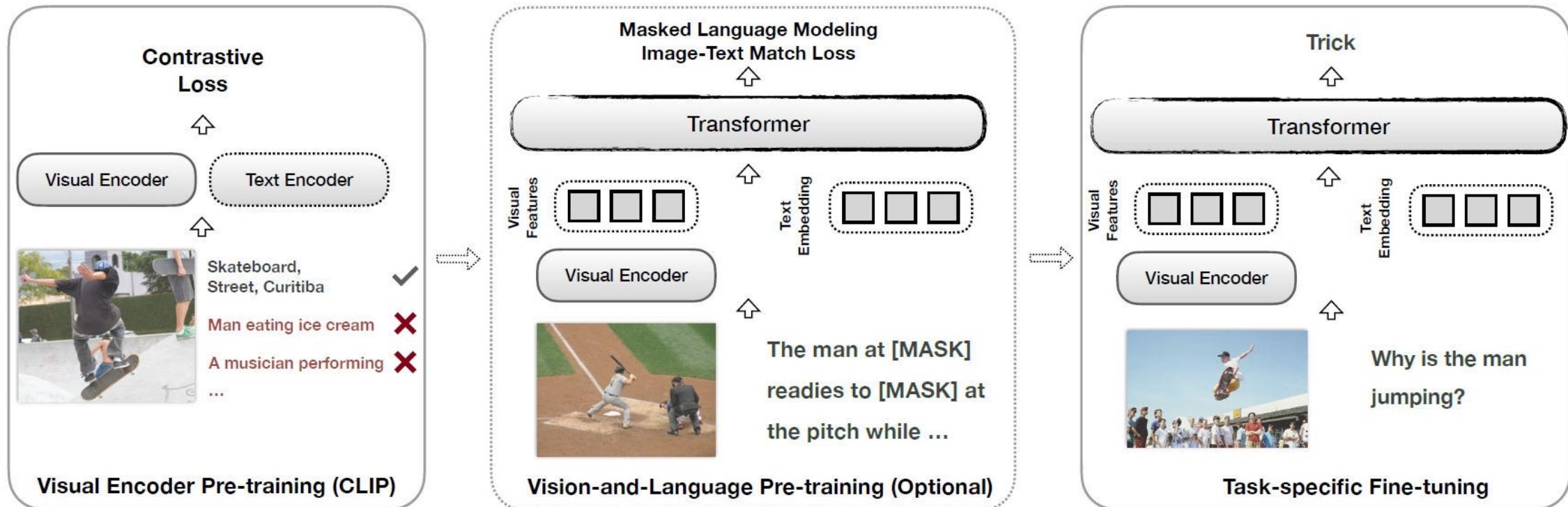


	ImageNet	Zero-Shot	ResNet101	CLIP	Δ Score
ImageNet			76.2	76.2	0%
ImageNetV2			64.3	70.1	+5.8%
ImageNet-R			37.7	88.9	+51.2%
ObjectNet			32.6	72.3	+39.7%
ImageNet Sketch			25.2	60.2	+35.0%
ImageNet-A			2.7	77.1	+74.4%

CLIP: Results



CLIP for Vision-and-Language Tasks?



CLIP for Vision-and-Language Tasks?

Model	VisualEncoder	V&L Pretrain		VQA	
		Data	Epoch	Test-Dev	Test-Std
PixelBERT	ImageNet-Res50	5.5M	40	71.35	71.42
PixelBERT	ImageNet-ResX152	5.5M	40	74.45	74.55
LXMERT	BUTD-Res101	9.2M	20	72.42	72.54
UNITER	BUTD-Res101	6.5M	-	72.70	72.91
Oscar	BUTD-Res101	6.5M	118	73.16	73.44
VinVL	VinVL-ResX152	8.9M	116	75.95	76.12
CLIP-ViL_p	CLIP-Res50	9.2M	20	73.92	74.09
	CLIP-Res50x4	9.2M	20	76.48	76.70

Model	B@4	M	C	S
BUTD (Anderson et al., 2018a)	36.3	27.7	120.1	21.4
VLP (Zhou et al., 2020)	39.5	29.3	129.8	22.4
AoANet (Huang et al., 2019b)	38.9	29.2	129.8	22.4
Oscar _{base} (Li et al., 2020)	40.5	29.7	137.6	22.8
VinVL _{base} (Zhang et al., 2021)	40.9	30.9	140.4	25.1
BUTD-Transformer* (Luo, 2020)	-	-	127.7	22.5
ImageNet-Res50Transformer	36.2	27.6	118.8	21.2
ImageNet-Res101Transformer	36.8	27.8	121.1	21.5
CLIP-Res50Transformer	38.6	28.8	127.9	22.7
CLIP-Res101Transformer	39.2	29.1	130.3	23.0
CLIP-Res50x4Transformer	40.2	29.7	134.2	23.8
CLIP-ViT-BTransformer	21.1	19.4	58.0	12.2

CLIP-ViL achieves competitive performance on VQA v2 (Left) and COCO image captioning (Right).

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. A pool of images	Query:  A pool of texts	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

CLIP: Summary

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

X CLIP doesn't go directly from image to text or vice versa. It just connects 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

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)

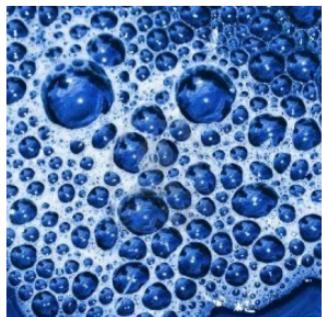
Flowers102



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)

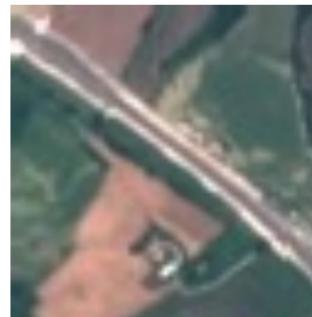
Describable Textures (DTD)



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)

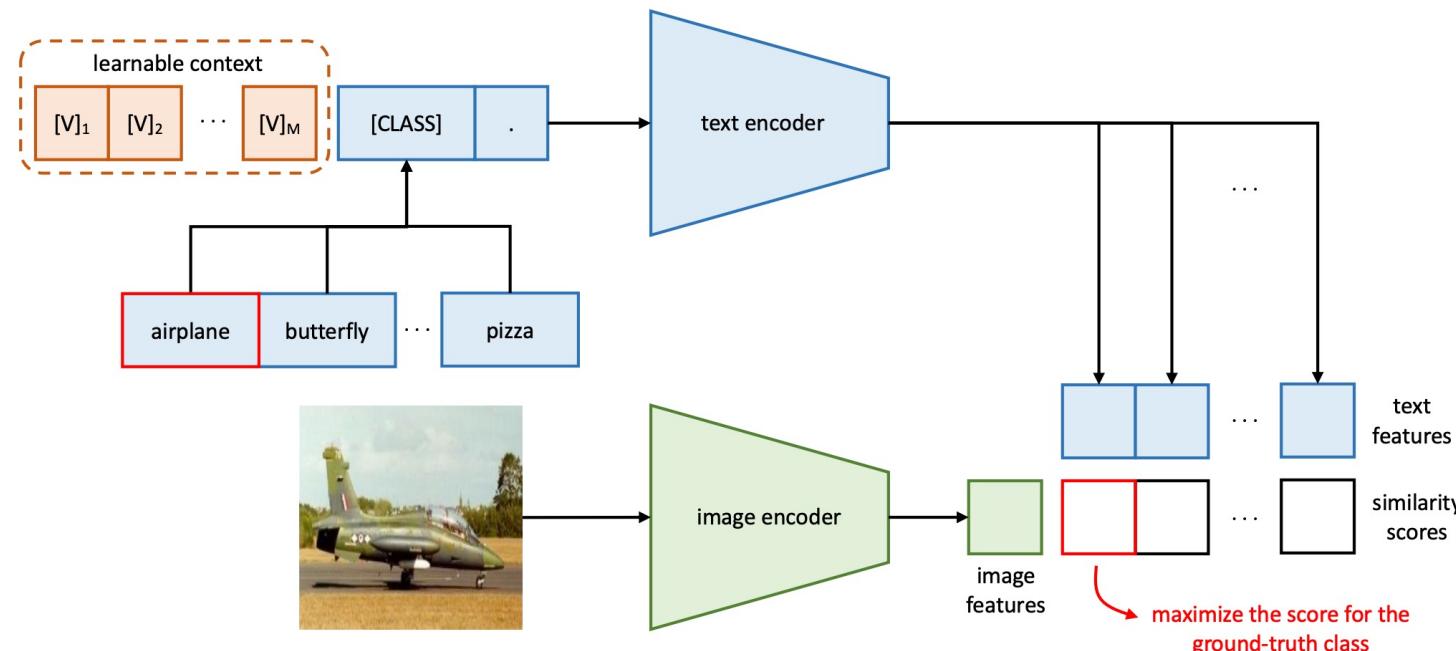
EuroSAT



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)

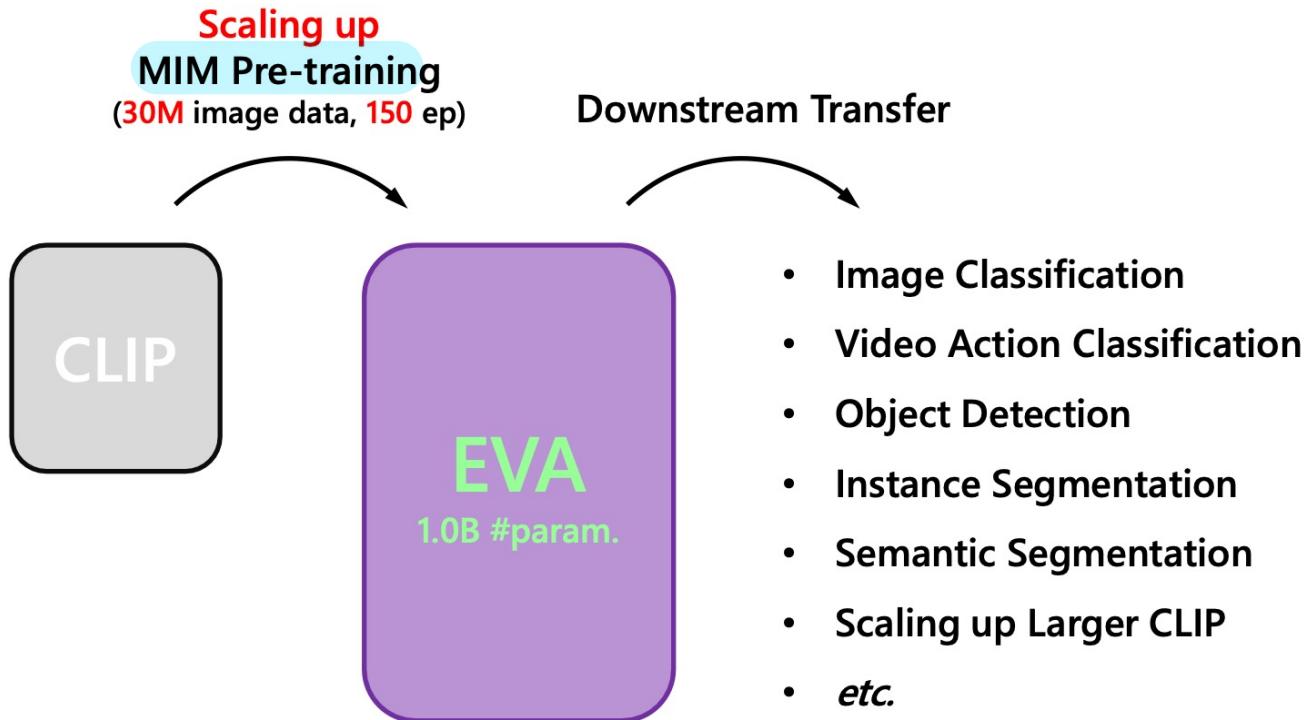
Learning to Prompt for VLMs



Method	Source ImageNet	Target			
		-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

EVA

- 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.



EVA

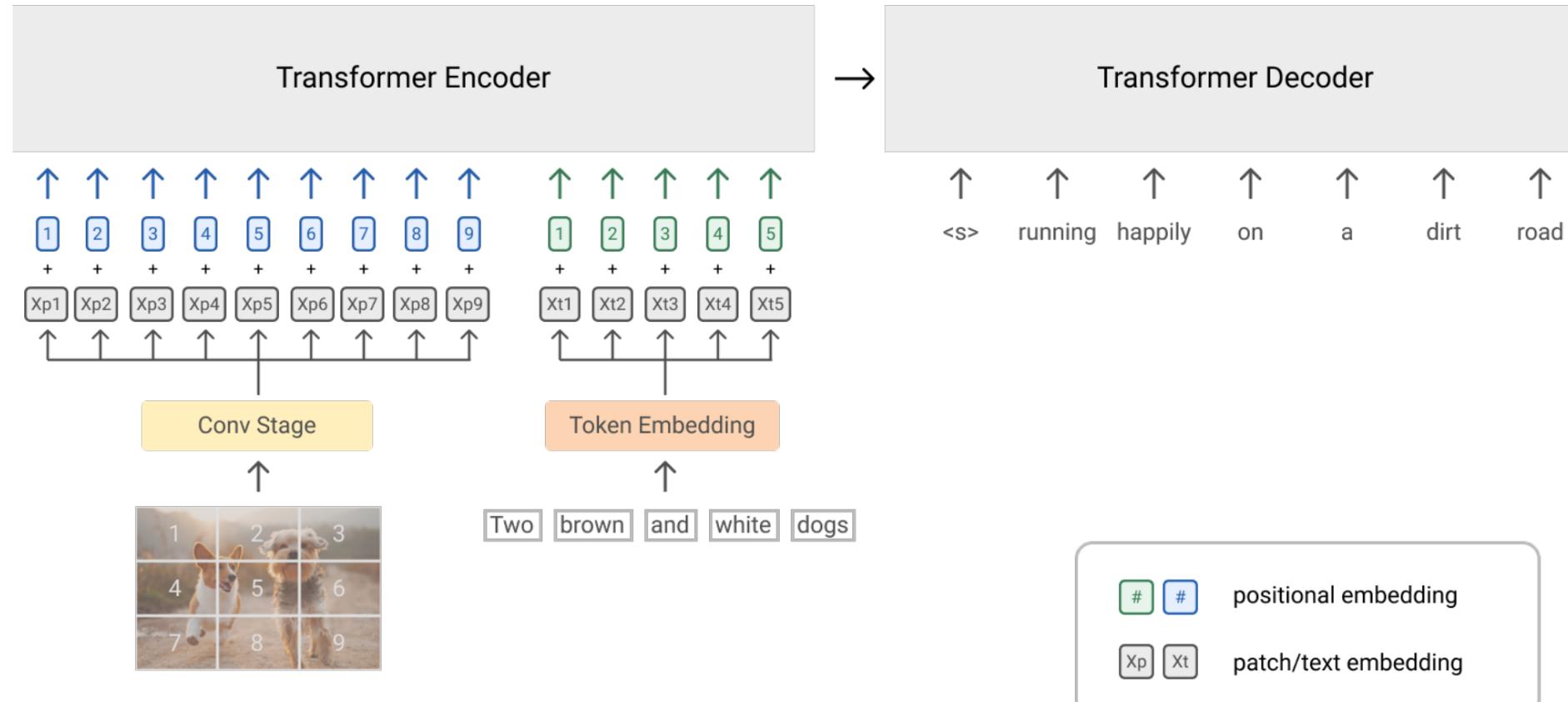
- The merged dataset for pre-training has 29.6 million images in total
- The output feature of EVA is first normalized and then projected to the same dimension as the CLIP feature via a linear layer.
 - negative cosine similarity as the loss function
 - The CLIP features used as MIM prediction targets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

SimVLM

- PrefixLM



running happily on a dirt road </s>

Transformer Decoder

<s> running happily on a dirt road

Chunyuan Li's CVPR 2023
Tutorial: Large Multimodal Models

Transformer Encoder

→

Transformer Decoder

Conv Stage

Token Embedding



Two brown and white dogs

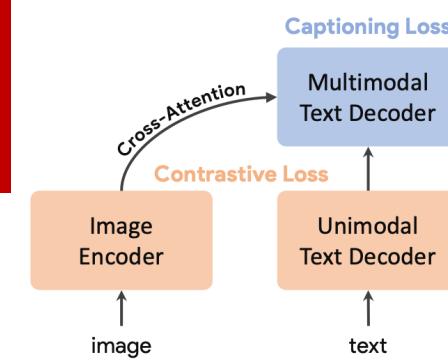
positional embedding

Xp Xt patch/text embedding

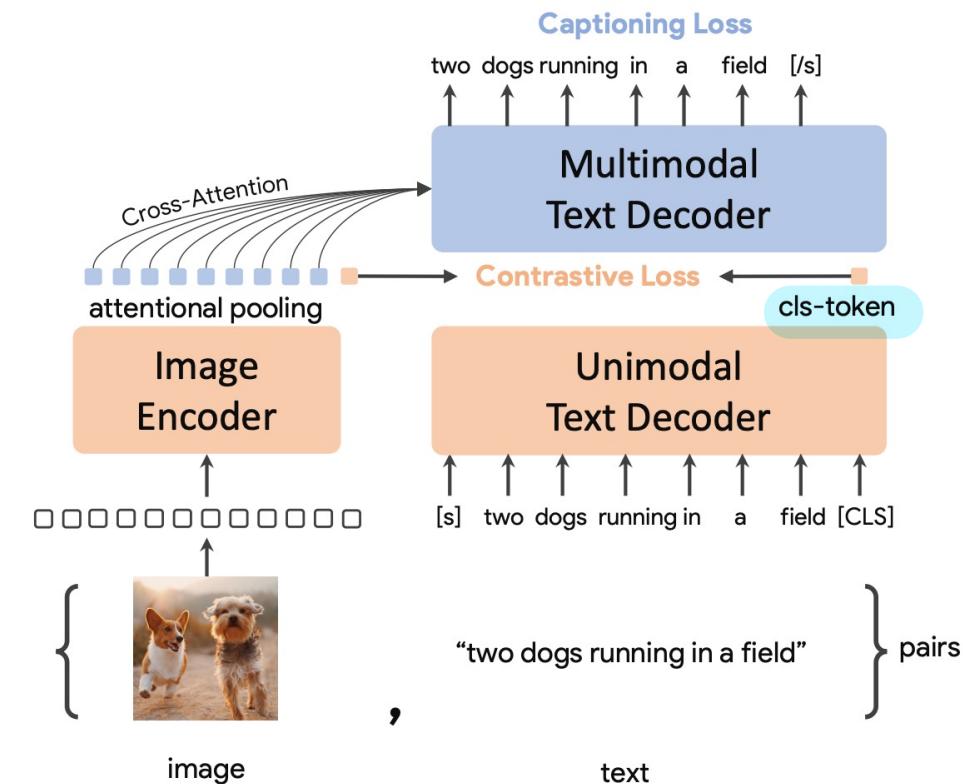
CoCa: Contrastive Captioner

- 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

Multi-modal fusion

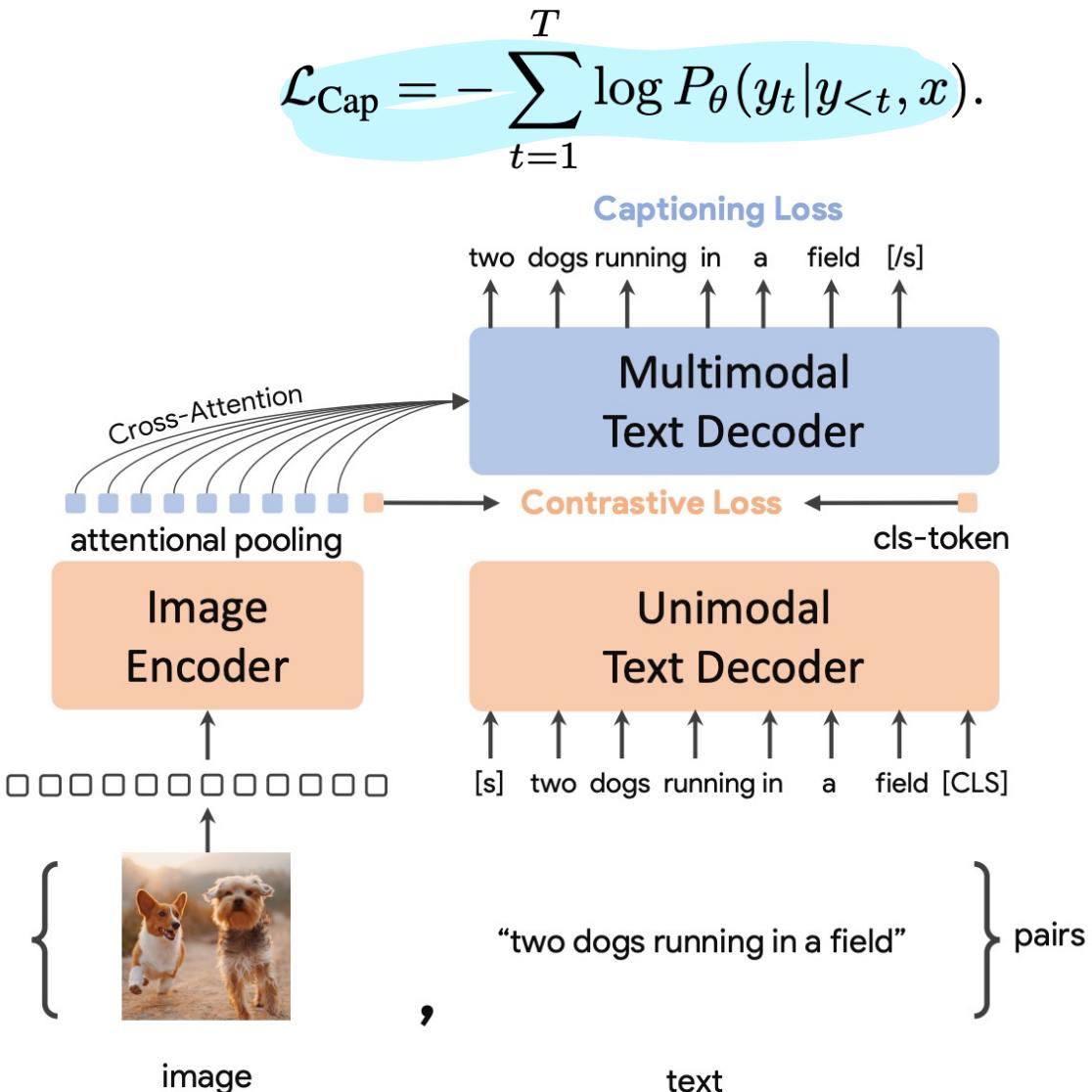


CoCa

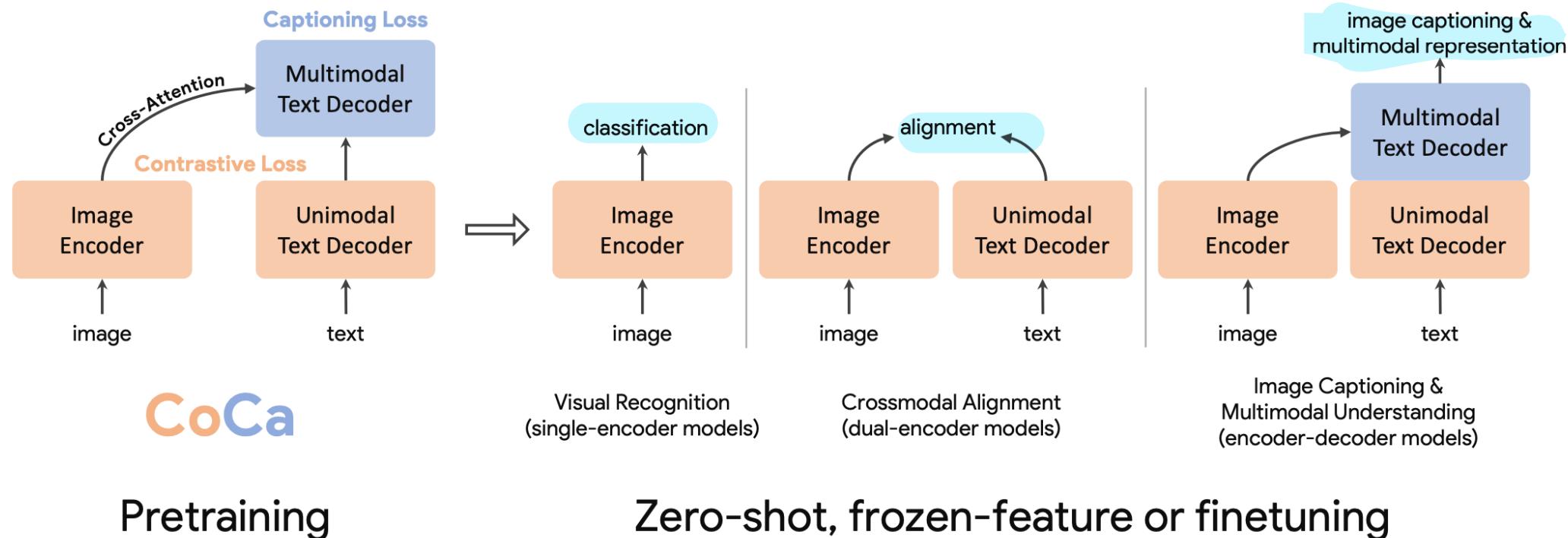


CoCa: Contrastive Captioner

- 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

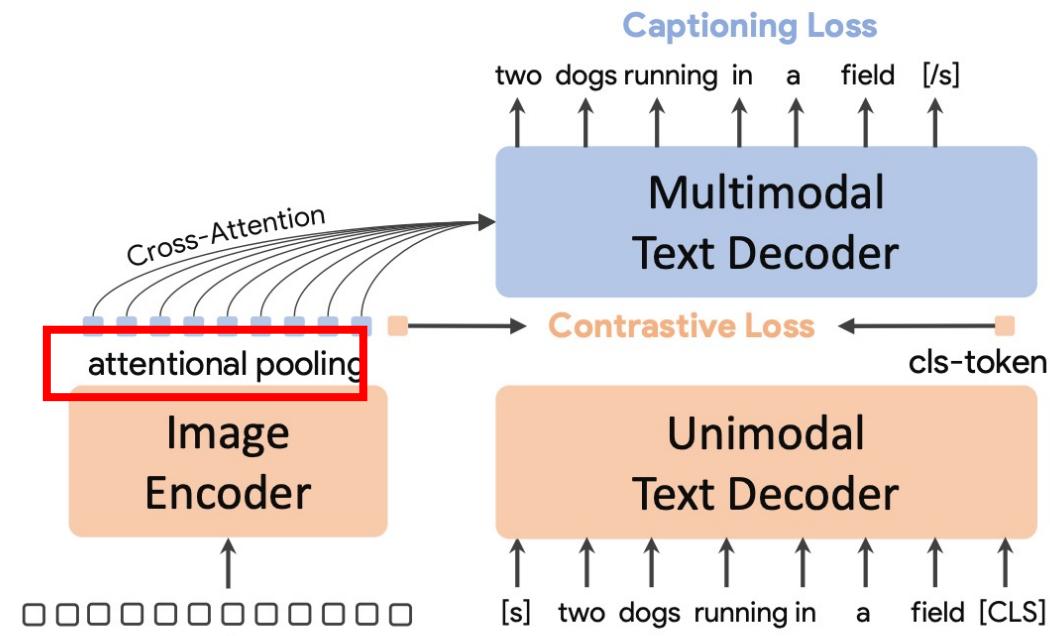


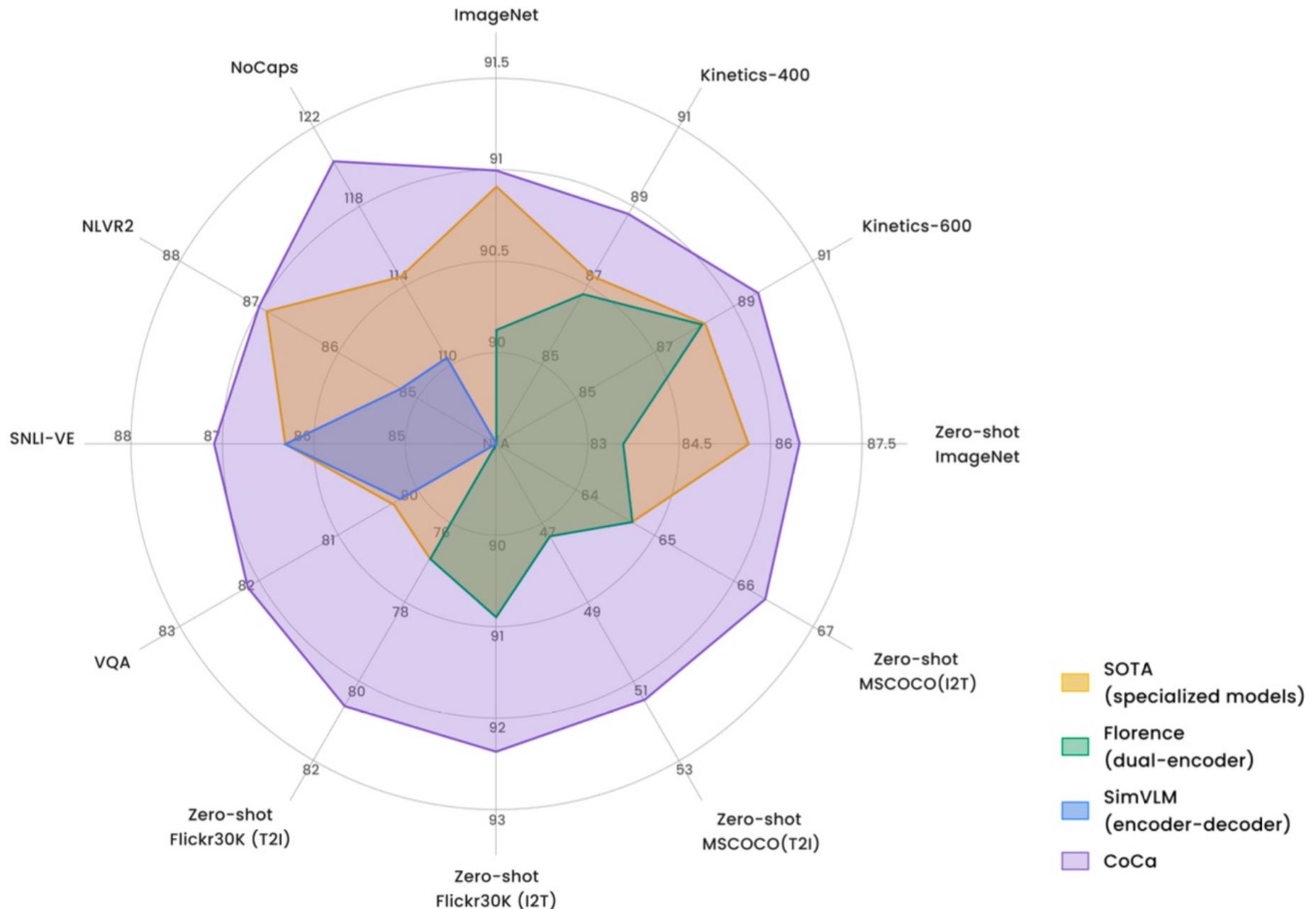
CoCa: Contrastive Captioner



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.





Architecture of Multimodal Models

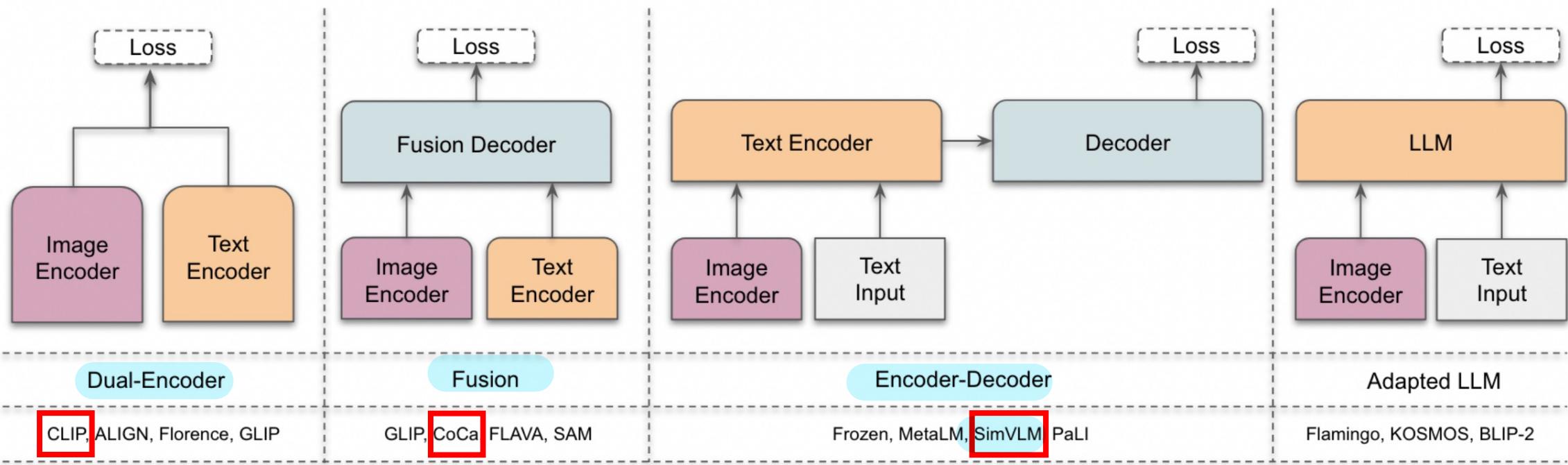


Image-Text Models: Summary

- 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

Text-to-Image Generation

Impressive conditional diffusion models

- Text-to-image generation

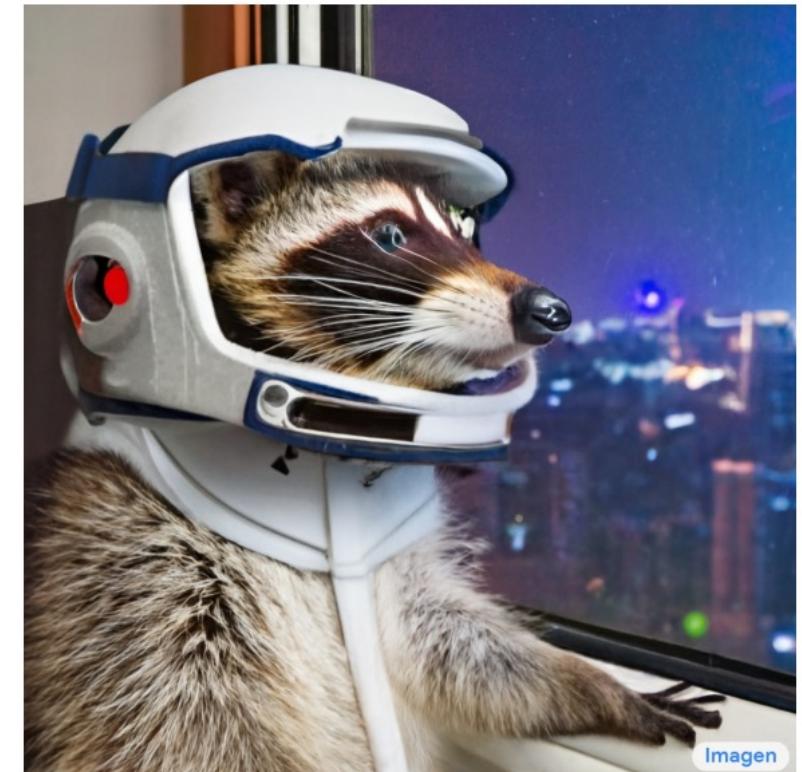
DALLE 2

“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.”



Successful Applications

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 suits 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](#)

Successful Applications

Text-to-Image Generation

Stable Diffusion



[Stable Diffusion Applications: Twitter Mega Thread](#)

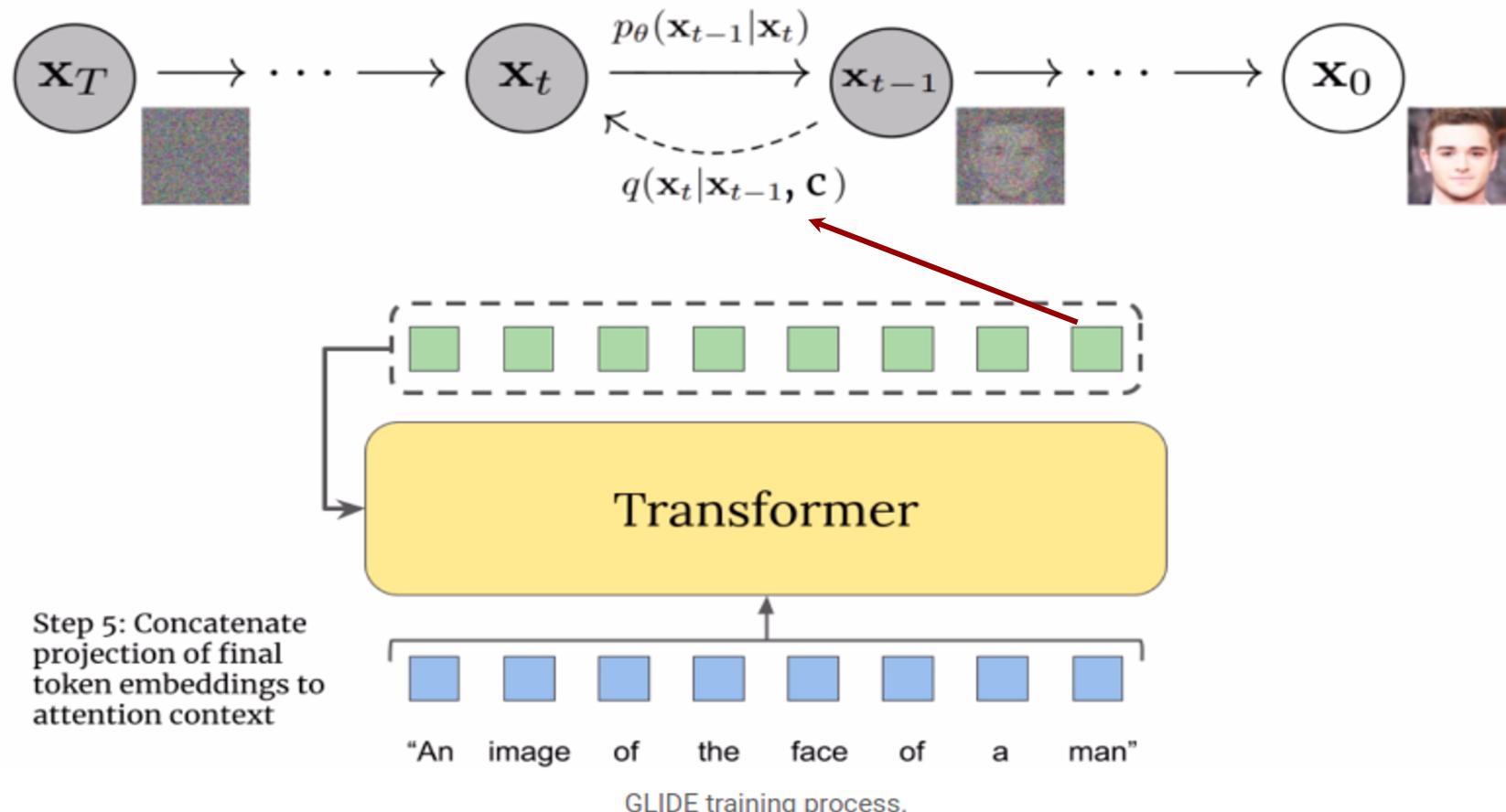
["High-Resolution Image Synthesis with Latent Diffusion Models" Rombach et al., 2022](#)

GLIDE

- Guided Language to Image Diffusion for Generation and Editing
- 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!

GLIDE: Training process

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





“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 dali 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”

[A. Nichol et al., GLIDE, 2021](#)



“an illustration of albert einstein wearing a superhero costume”

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

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

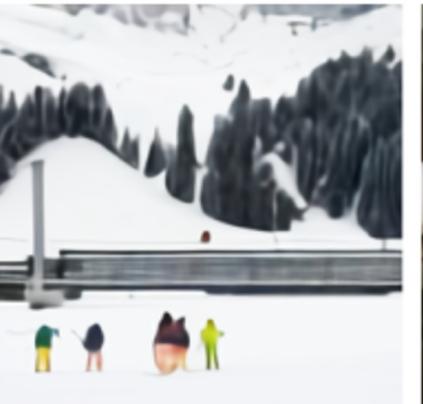
GLIDE: CLIP Guidance

- Instead of $\nabla_{x_t} \log p(y|x_t)$, it uses $\nabla_{x_t} f(x_t) \cdot g(c)$ 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

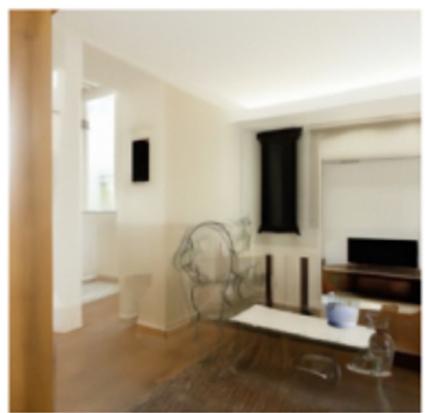
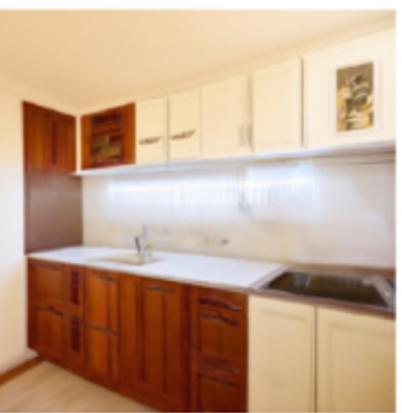
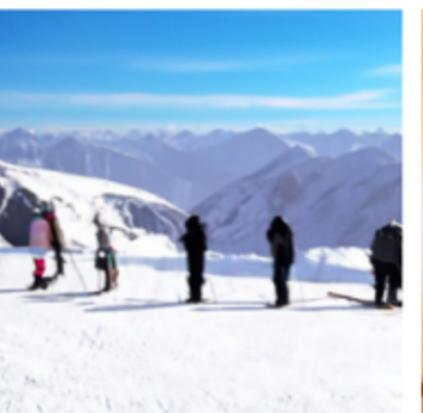
GLIDE: Architecture

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

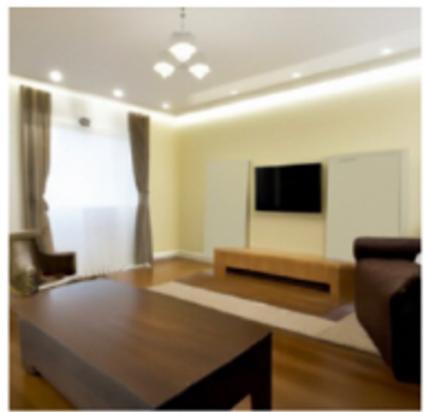
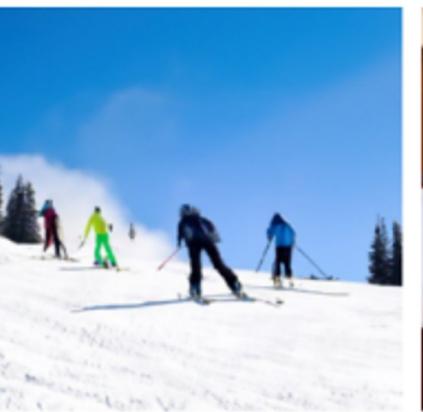
DALL-E



GLIDE (CLIP Guid.)



GLIDE (CF Guid.)



"a green train is coming down the tracks"

"a group of skiers are preparing to ski down a mountain."

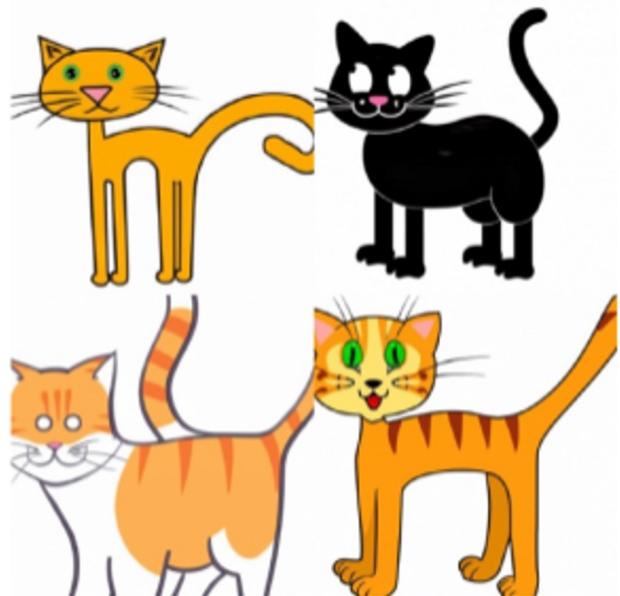
"a small kitchen with a low ceiling"

"a group of elephants walking in muddy water."

"a living area with a television and a table"

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”



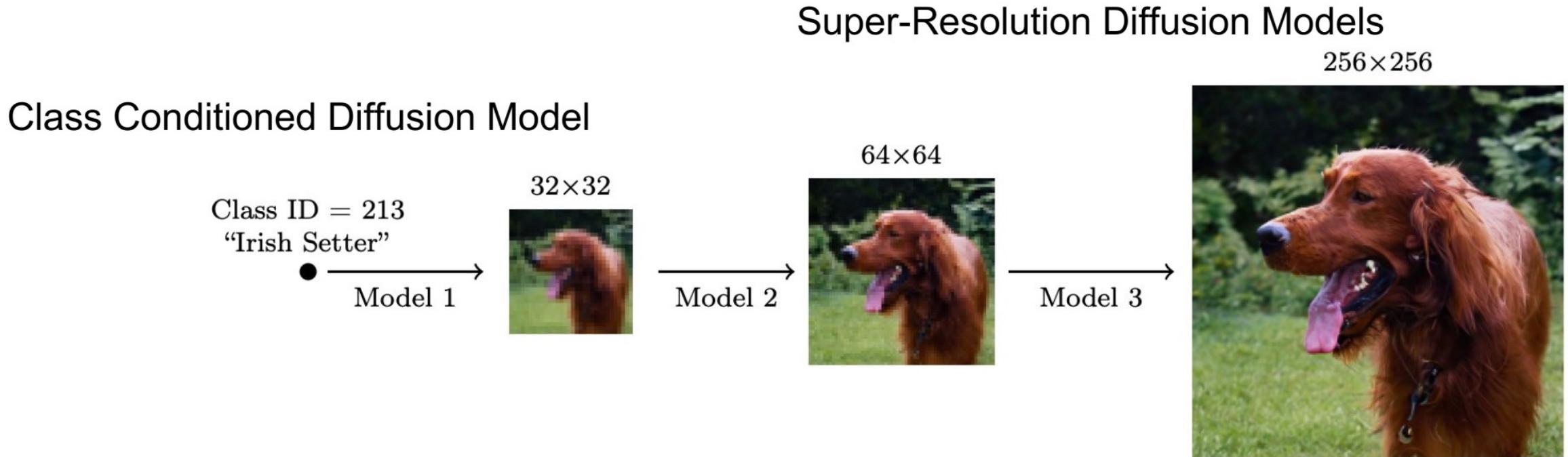
“a mouse hunting a lion”



“a car with triangular wheels”

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)



“zebras roaming in the field”



“a girl hugging a corgi on a pedestal”



“a man with red hair”



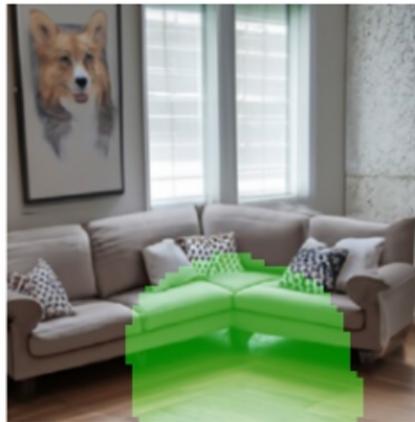
“an old car in a snowy forest”

Image inpainting

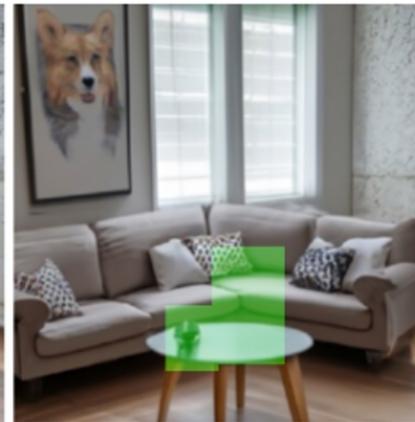
- Fine-tune model to perform inpainting
- 4 additional channels (including RGB of remaining portions along with the mask) are fed as additional conditioning information
- For upsampling, a full low-resolution is fed but unmasked regions of high-resolution image



“a cozy living room”



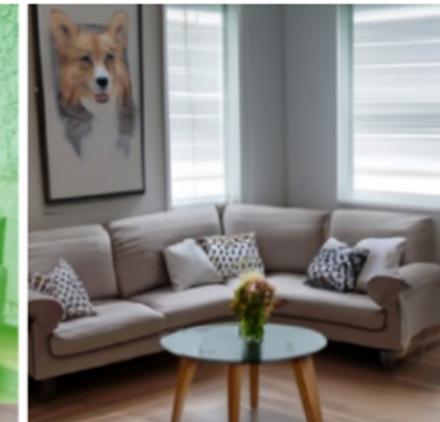
“a painting of a corgi
on the wall above
a couch”



“a round coffee table
in front of a couch”



“a vase of flowers on a
coffee table”



“a couch in the corner
of a room”

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

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

OpenAI DALL-E 2

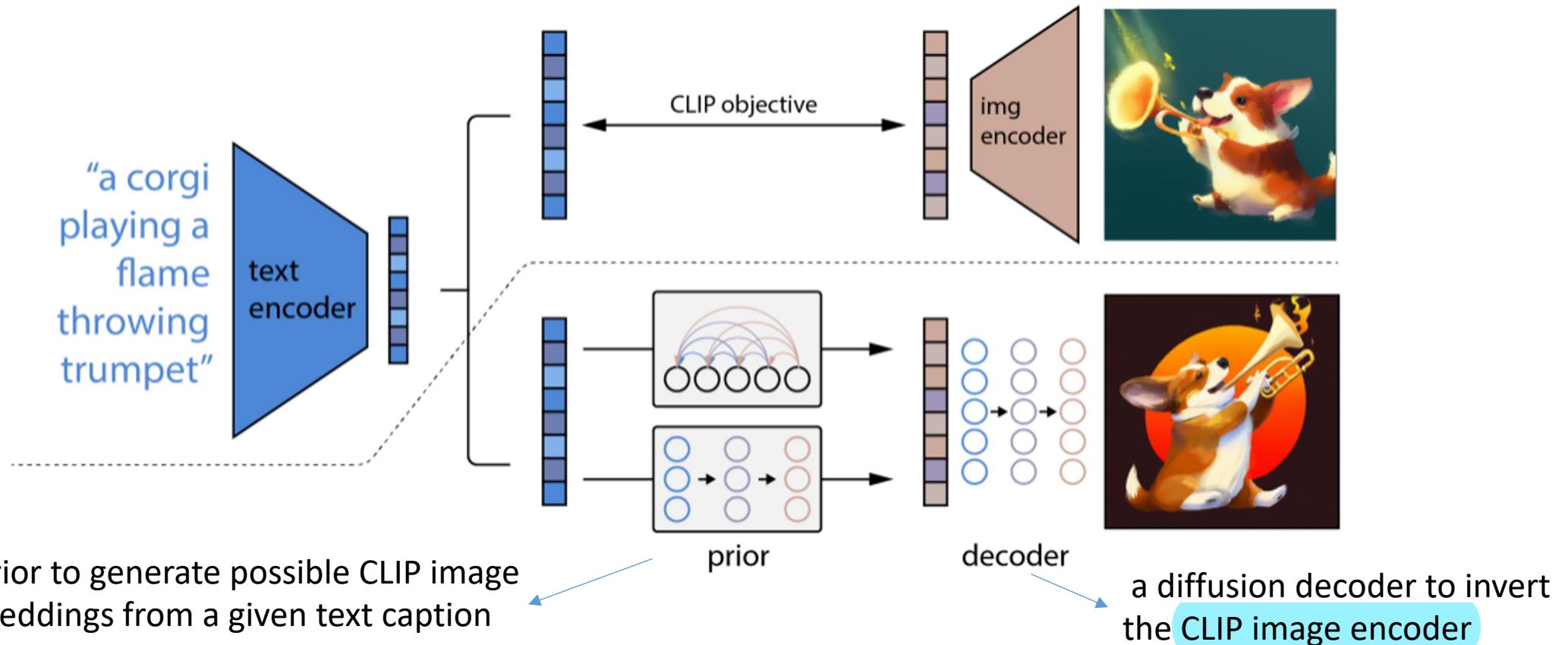


Teddy bears shopping for groceries in ancient egypt



What does the OpenAI office look like?

OpenAI DALL-E 2



- 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

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

Image manipulations: Interpolations



Encoding with clip + decoding with a diffusion model

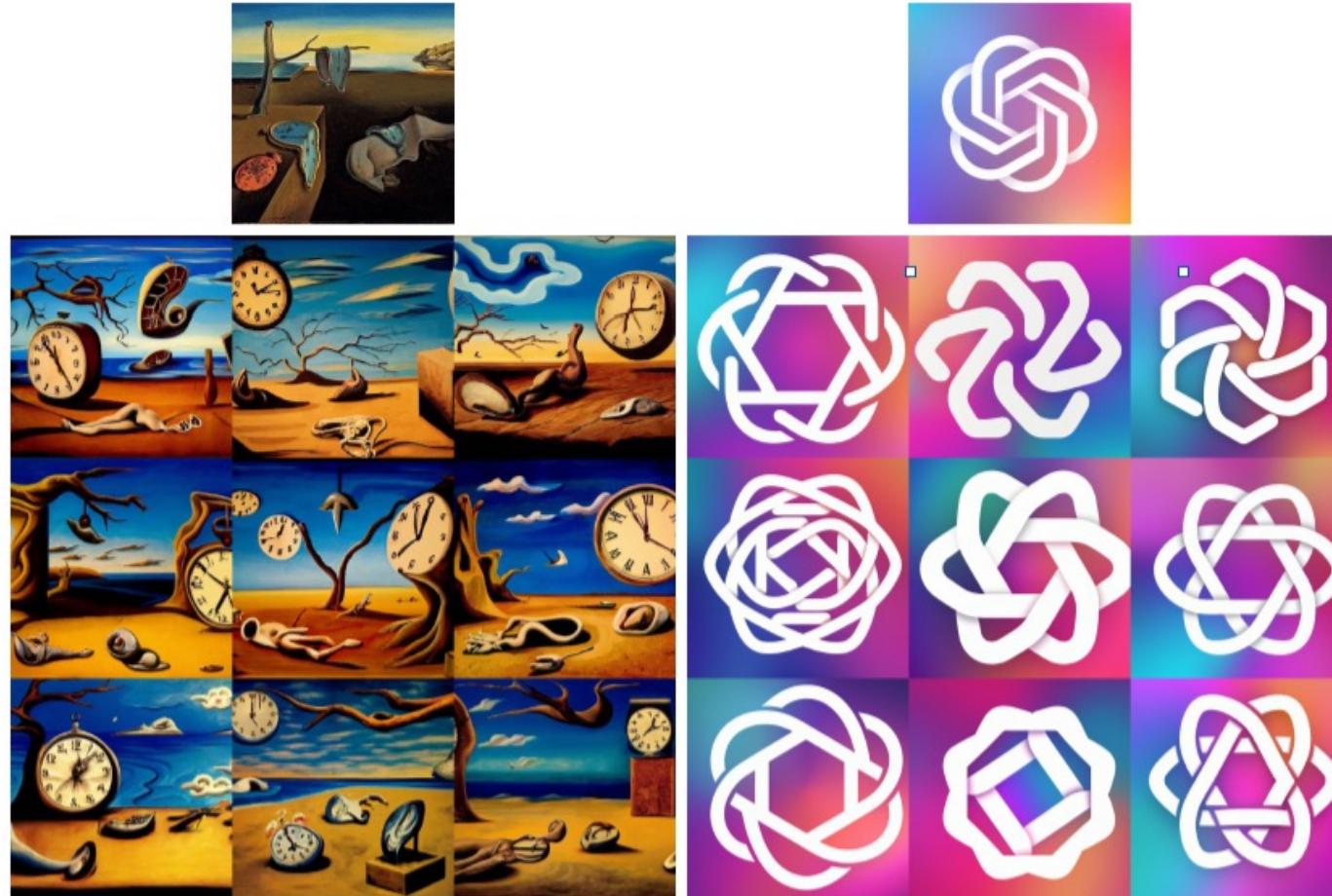


Image manipulations: PCA reconstructions



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.

Image manipulations: Text diffs

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



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

1.0
2.0
3.0
4.0



unCLIP

GLIDE

Quantitative analysis

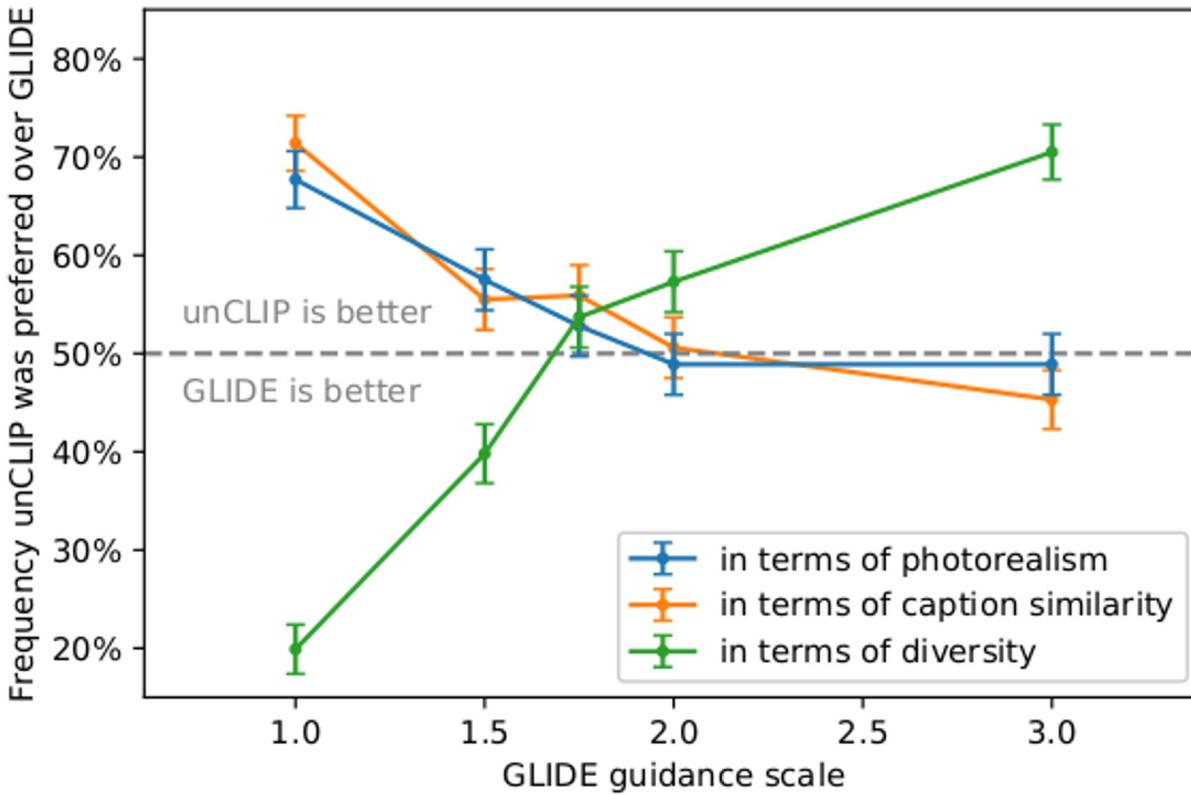


Figure 10: When comparing unCLIP (with our best sampling settings) to various settings of guidance scale for GLIDE, unCLIP was preferred by human evaluators on at least one axis among photorealism, caption similarity, and diversity for each comparison. At the higher guidance scales used to generate photorealistic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

Failure cases

"A red cube on the top of a blue cube."

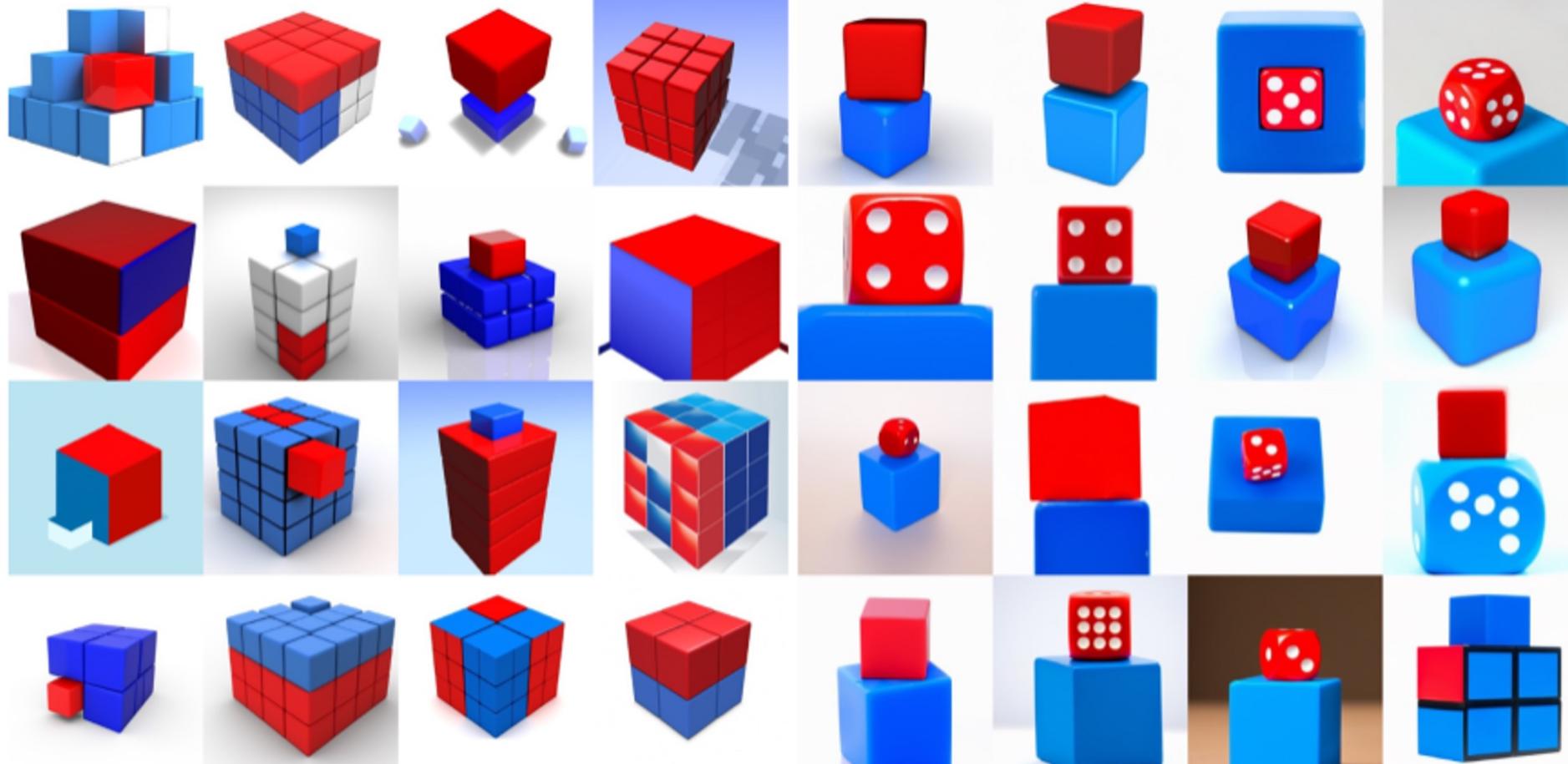




Figure 15: Reconstructions from the decoder for difficult binding problems. We find that the reconstructions mix up objects and attributes. In the first two examples, the model mixes up the color of two objects. In the rightmost example, the model does not reliably reconstruct the relative size of two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



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

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>

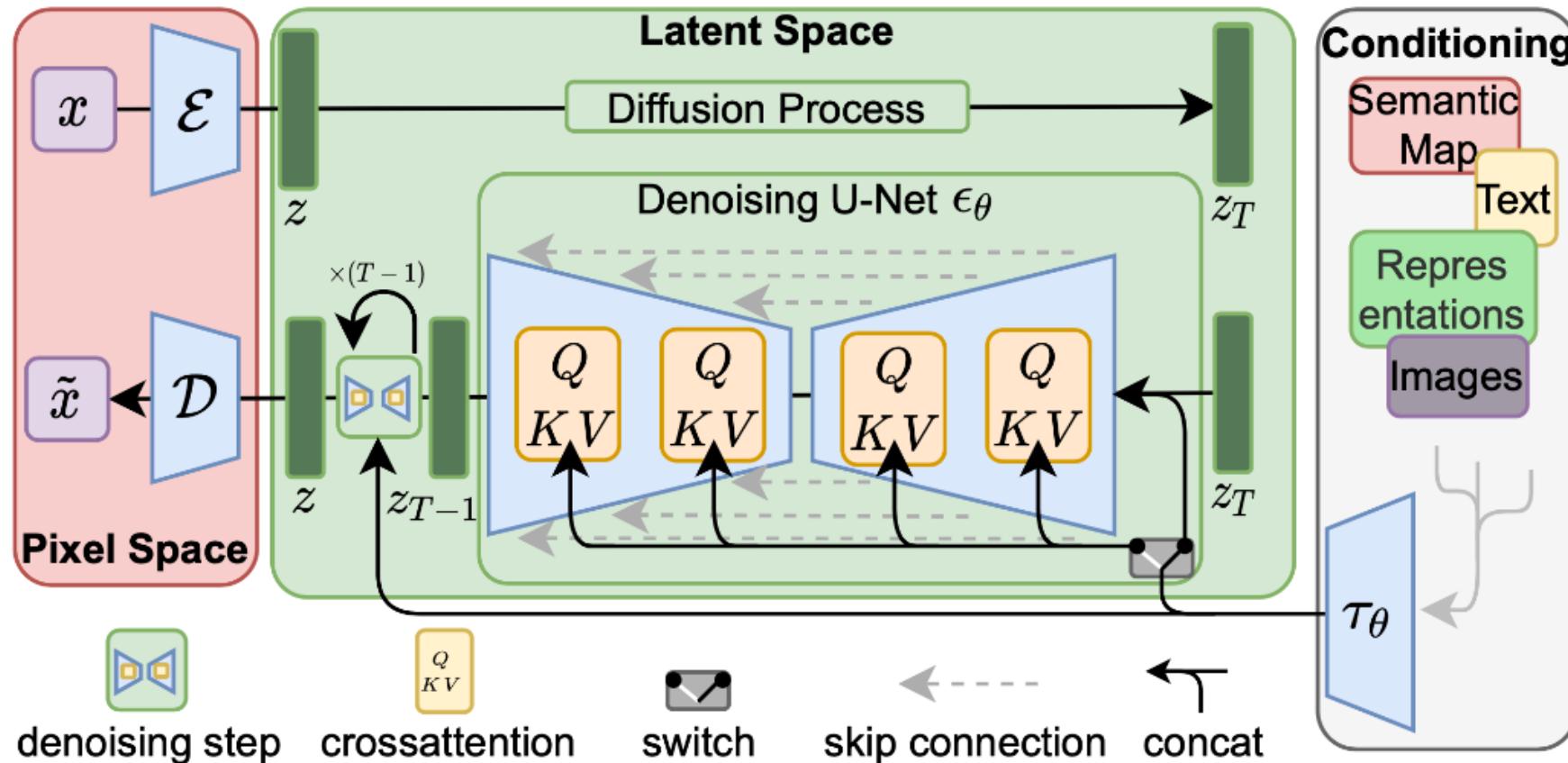
<https://arxiv.org/abs/2205.11487>

Latent Space Diffusion Models

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

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

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



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

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