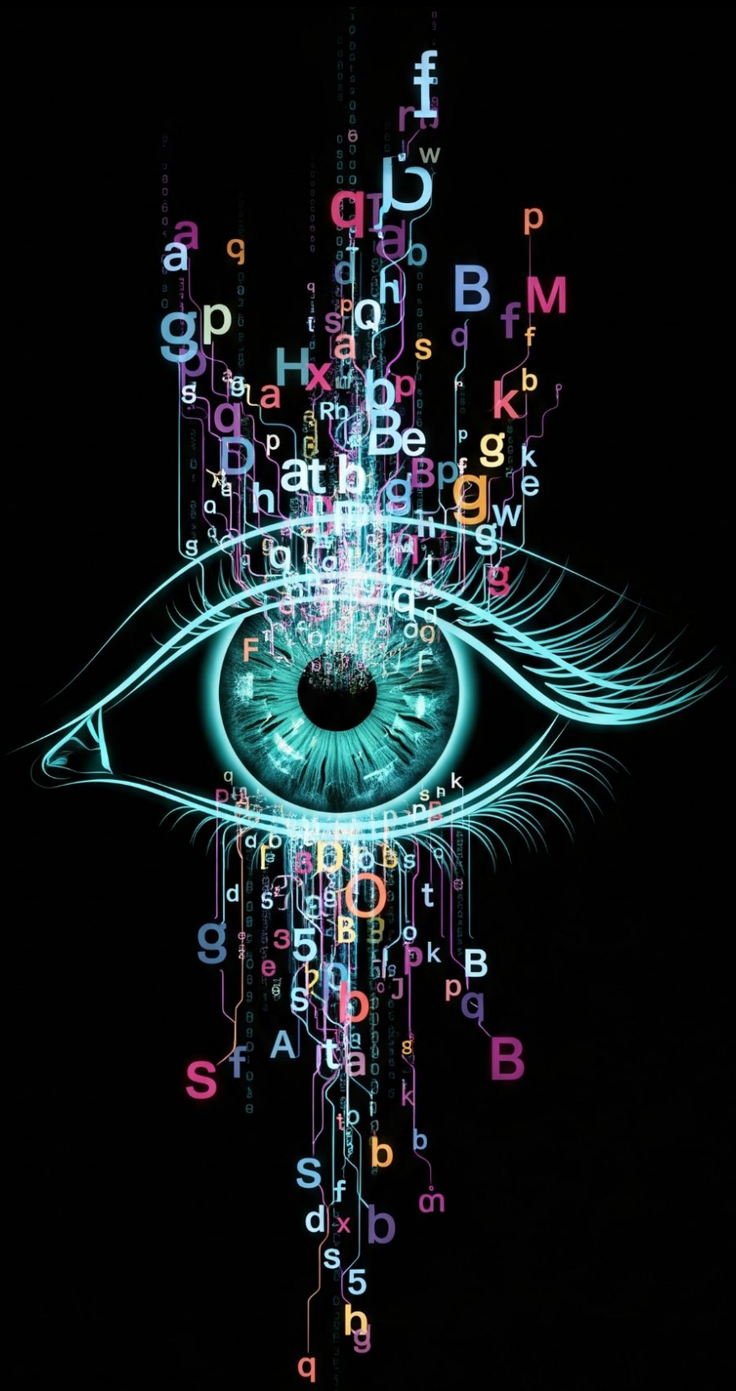
Vision-Language Models
EEML 2025

Otilia Stretcu
Research Scientist, Google Research
otiliastr@google.com



Vision Tasks

Image Classification

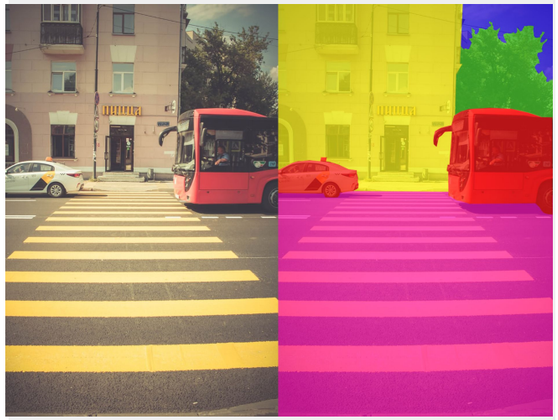
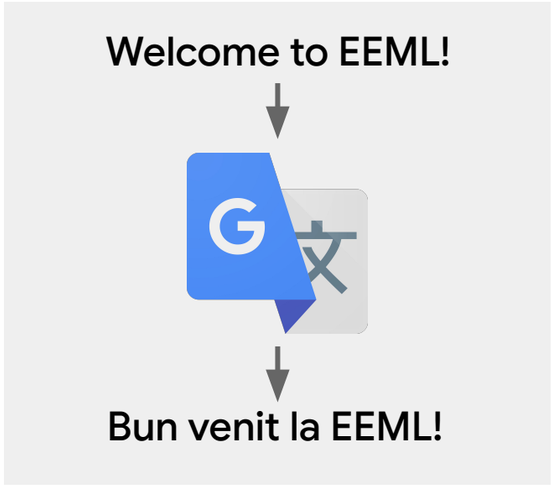
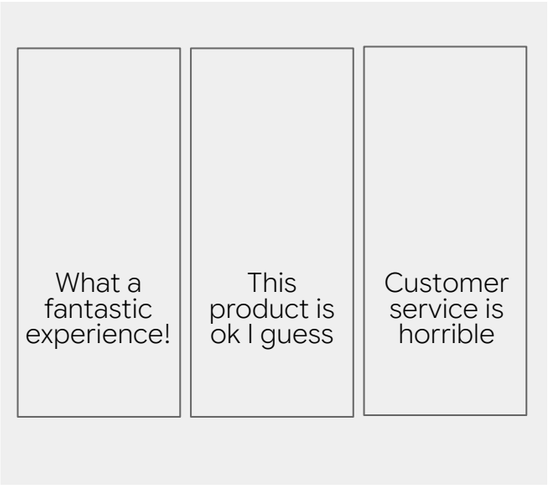


Image Segmentation

Language Tasks



Machine Translation



Sentiment Analysis

Vision-Language Tasks

Image Captioning

[Xu et. al, 2015]

A surfer riding on a wave

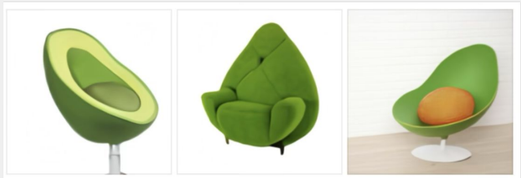
Visual Question Answering (VQA)

[Agrawal et al, 2016]

How many slices of pizza are there?

Text-to-Image Generation

An armchair in the shape of an avocado



[Ramesh et. al, 2021]

Multimodal Search

red dress with dots


Garfield Striped

Vision-Language Models Timeline

What are VLMs?

Definition

Vision-language models (VLMs) are multimodal models that simultaneously process and understand both visual data (image, video) and text, to solve different tasks such as visual question-answering, image / video captioning, retrieval, classification etc.

VLM Key Components

Model Architecture

- How is each modality is encoded?
- How do we fuse the different modalities?

Training Strategy

- Loss function
- Initialization (pretrained weights or from scratch?)
- Training stages (e.g., pretraining, fine-tuning)
- Data
  - Types of tasks (e.g., VQA, captioning)
  - Supervision (where do the target outputs come from?)

Structure the talk based on model architectures...

... and for each discuss training strategies.

Vision-Language Model Architectures

Encoder-Decoder

Dual-Encoder

Cross-Modal

Natively Multimodal

Vision-Language Model Architectures

Encoder-Decoder

Dual-Encoder

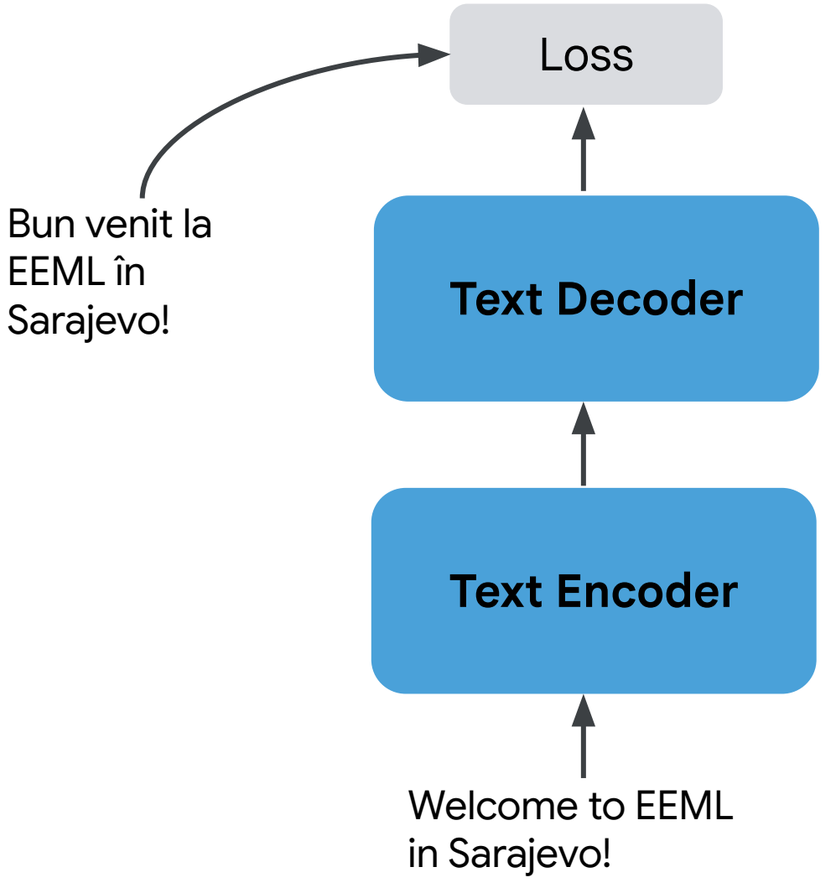
Cross-Modal

Natively Multimodal

Encoder-Decoder Models

Key Idea

- Inspired by encoder-decoder models in NLP (e.g., machine translation)

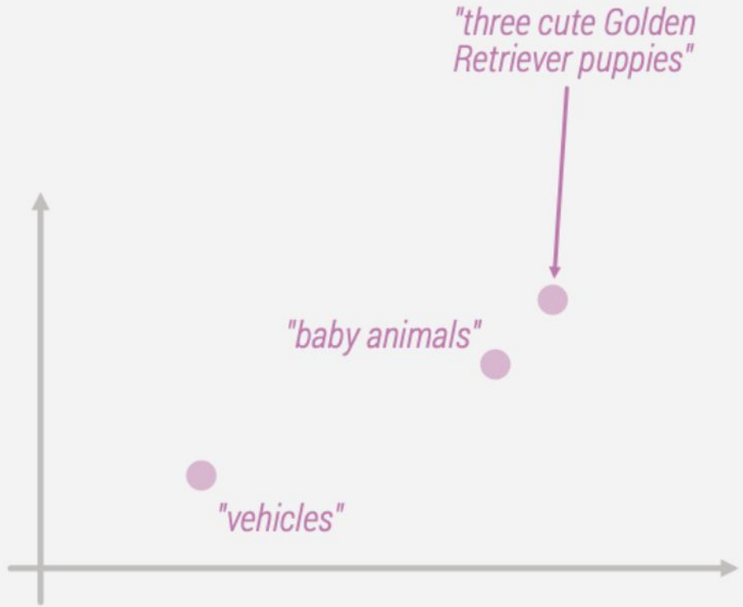


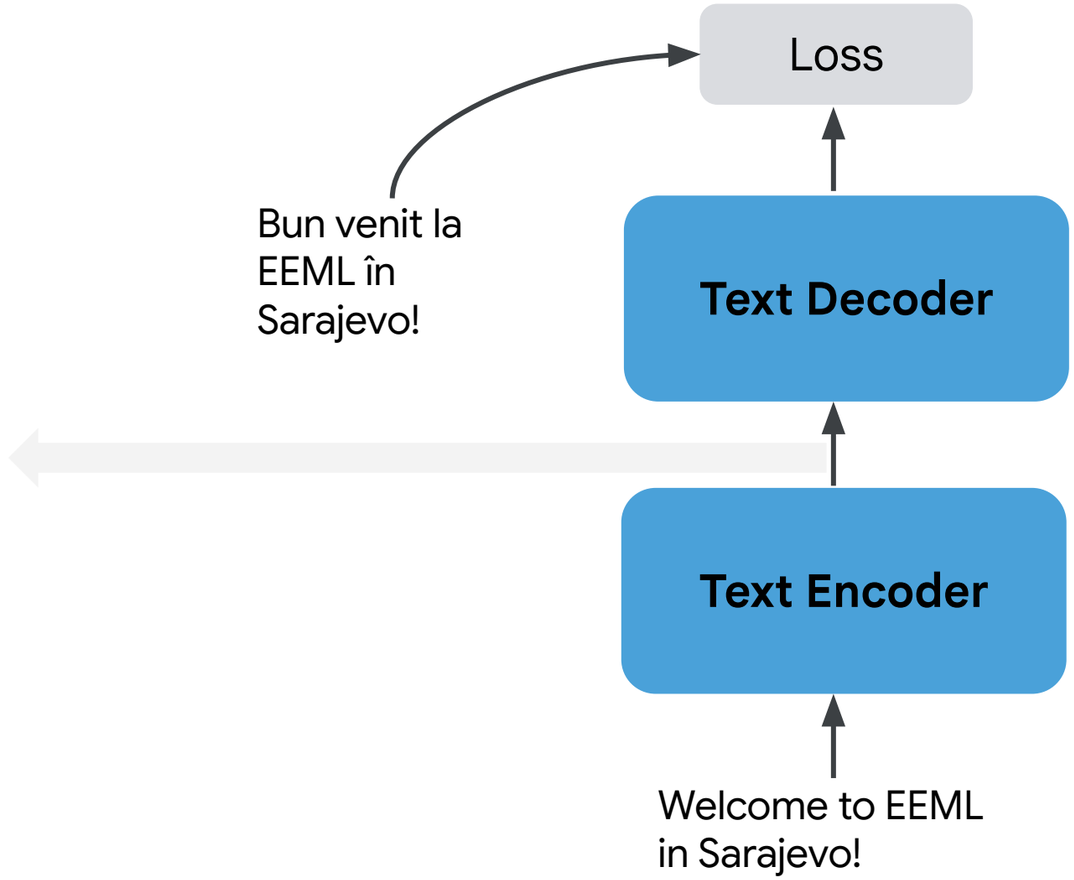
Machine Translation in NLP

Encoder-Decoder Models

What is an embedding?

embedding = a vector representation of some data, that encodes its *meaning*



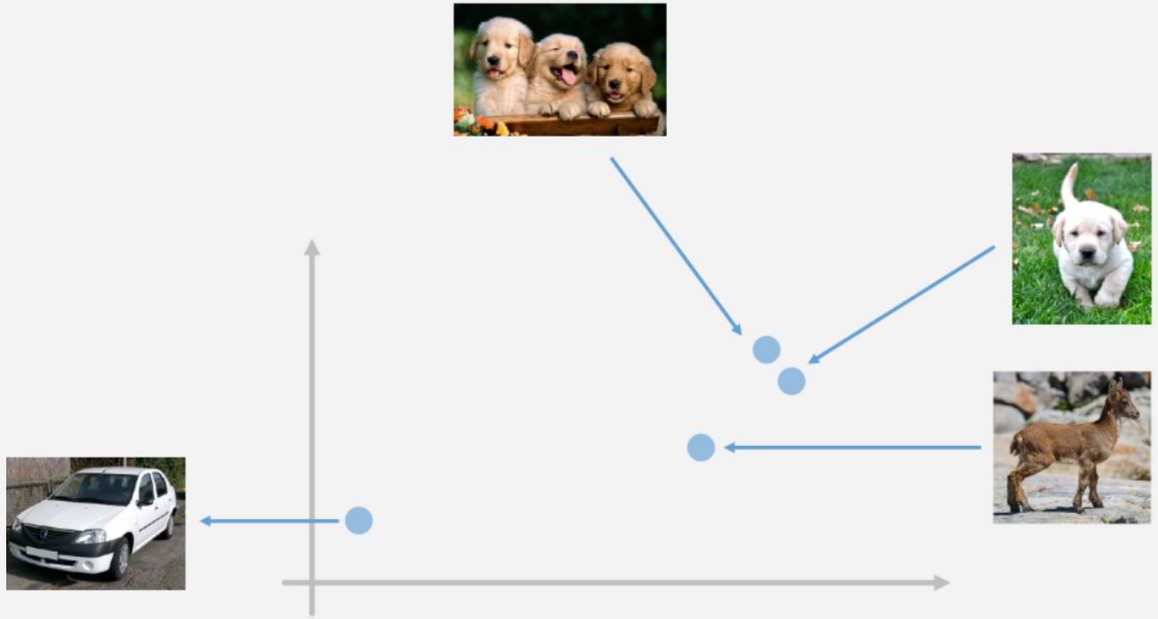


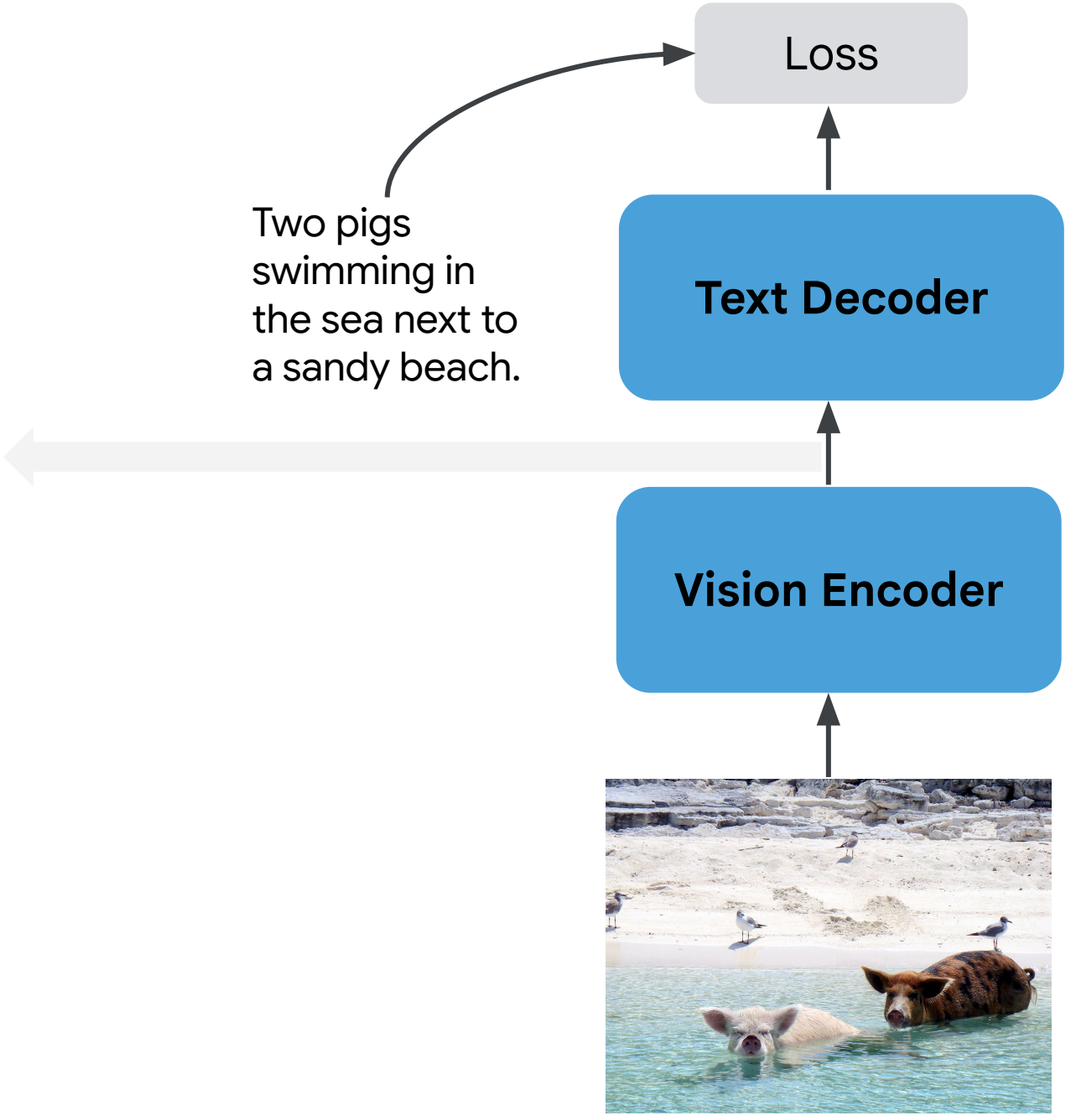
Machine Translation in NLP

Encoder-Decoder Models

What is an embedding?

embedding = a vector representation of some data, that encodes its *meaning*



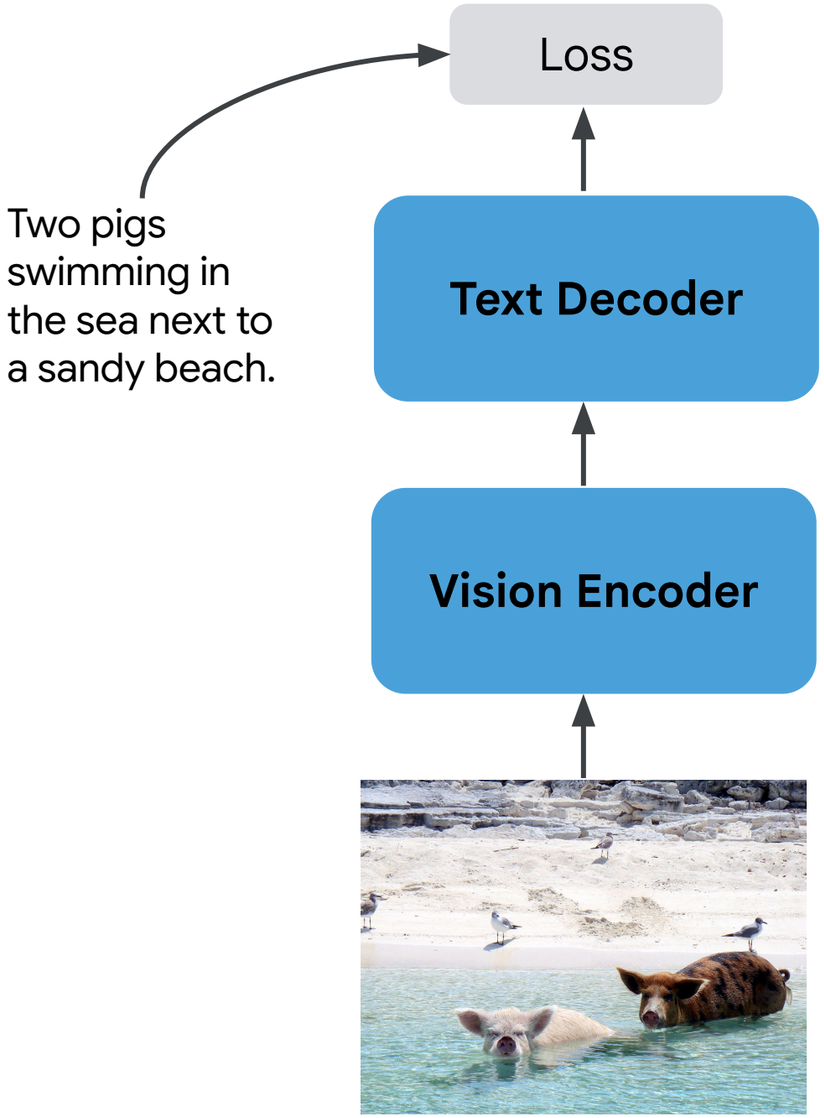


Encoder-Decoder Model

Encoder-Decoder Models

Key Idea

- Inspired by encoder-decoder models in NLP (e.g., machine translation)
- Can be used to generate natural sentences describing an image
- Trained to maximize the likelihood of the target description sentence given the training image



Encoder-Decoder Model

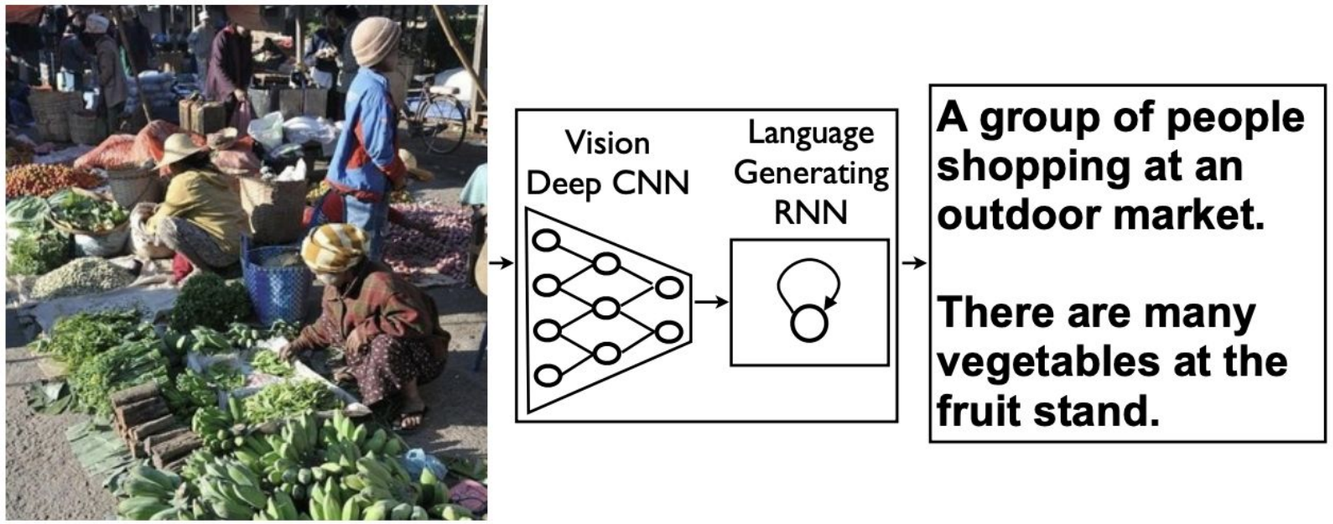
Show and Tell: A Neural Image Caption Generator

Architecture

- Encoder: CNN
- Decoder: RNN

Training Strategy

- Pretrain the CNN for image classification.
- Drop the last CNN layer, and pass encoded image to RNN.
- Train to “translate” the image into text, by maximizing the probability of the correct caption given the image.



$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta)$$

model parameters

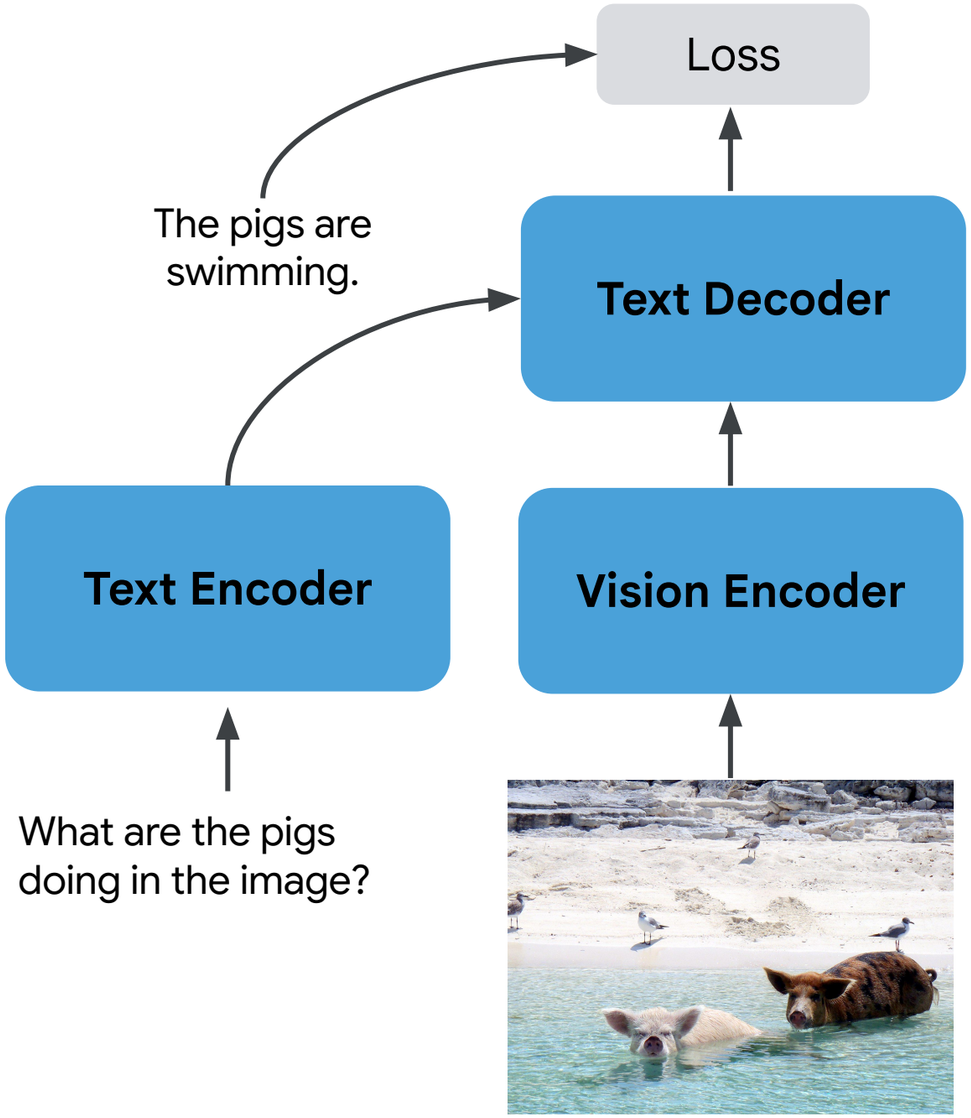
caption

image

Vinyals et. al, *Show and Tell: A Neural Image Caption Generator,* CVPR 2015

VQA: Visual Question Answering

To solve tasks such as VQA we can also include a **text encoder**.

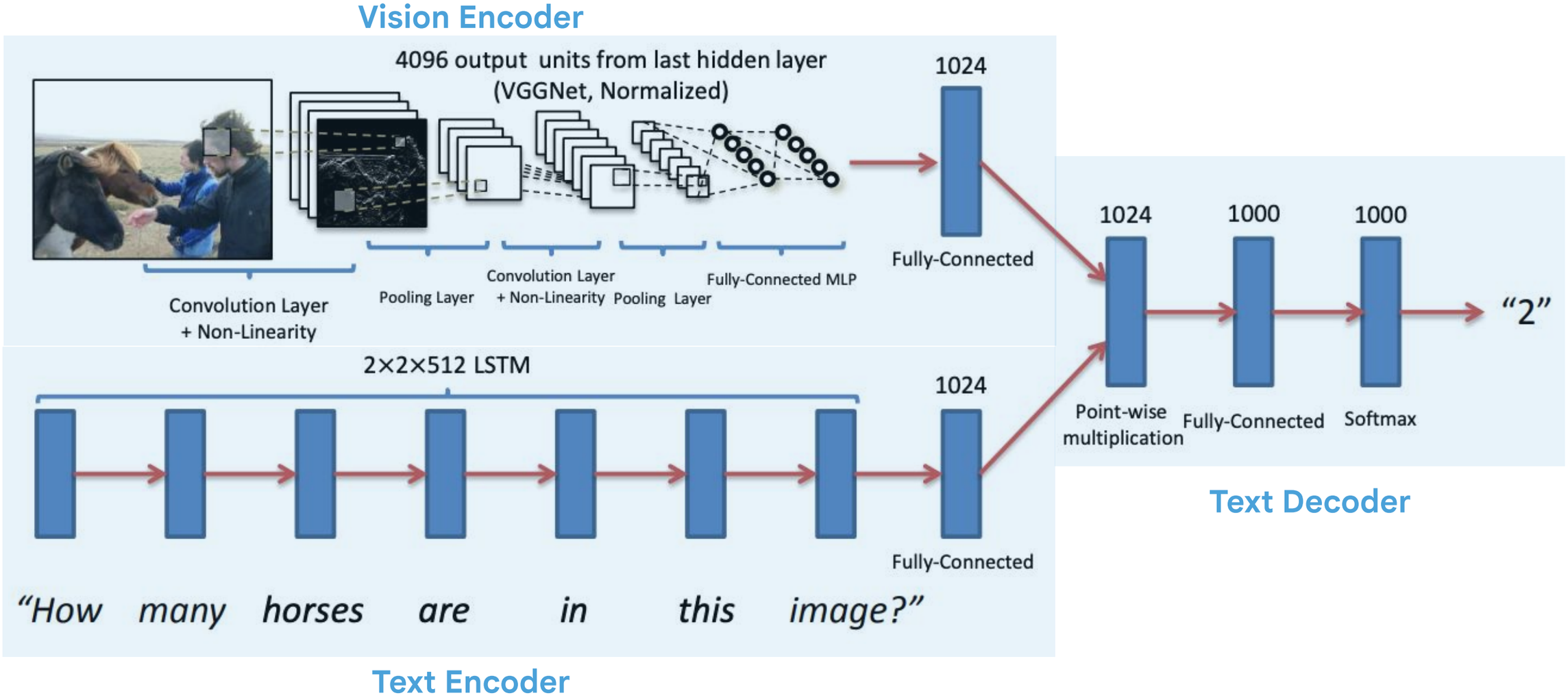


Encoder-Decoder Model

Antol et. al, *VQA: Visual Question Answering,* ICCV 2015

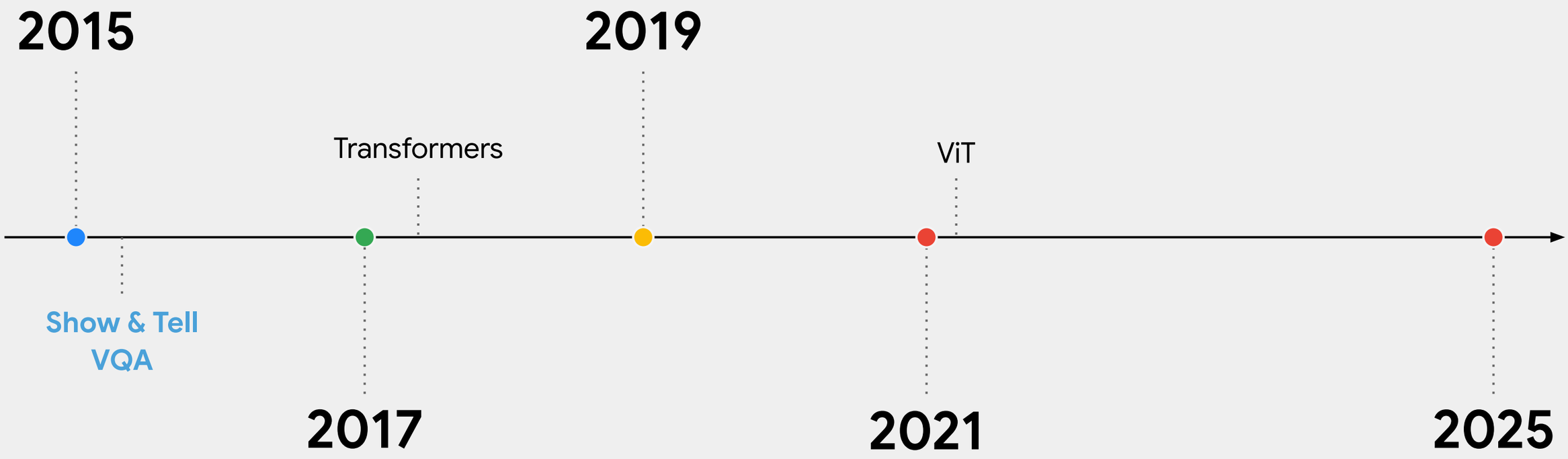
VQA: Visual Question Answering

Best performing model on VQA dataset introduced by Antol et. al, 2015:



Antol et. al, *VQA: Visual Question Answering,* ICCV 2015

Vision-Language Models Timeline



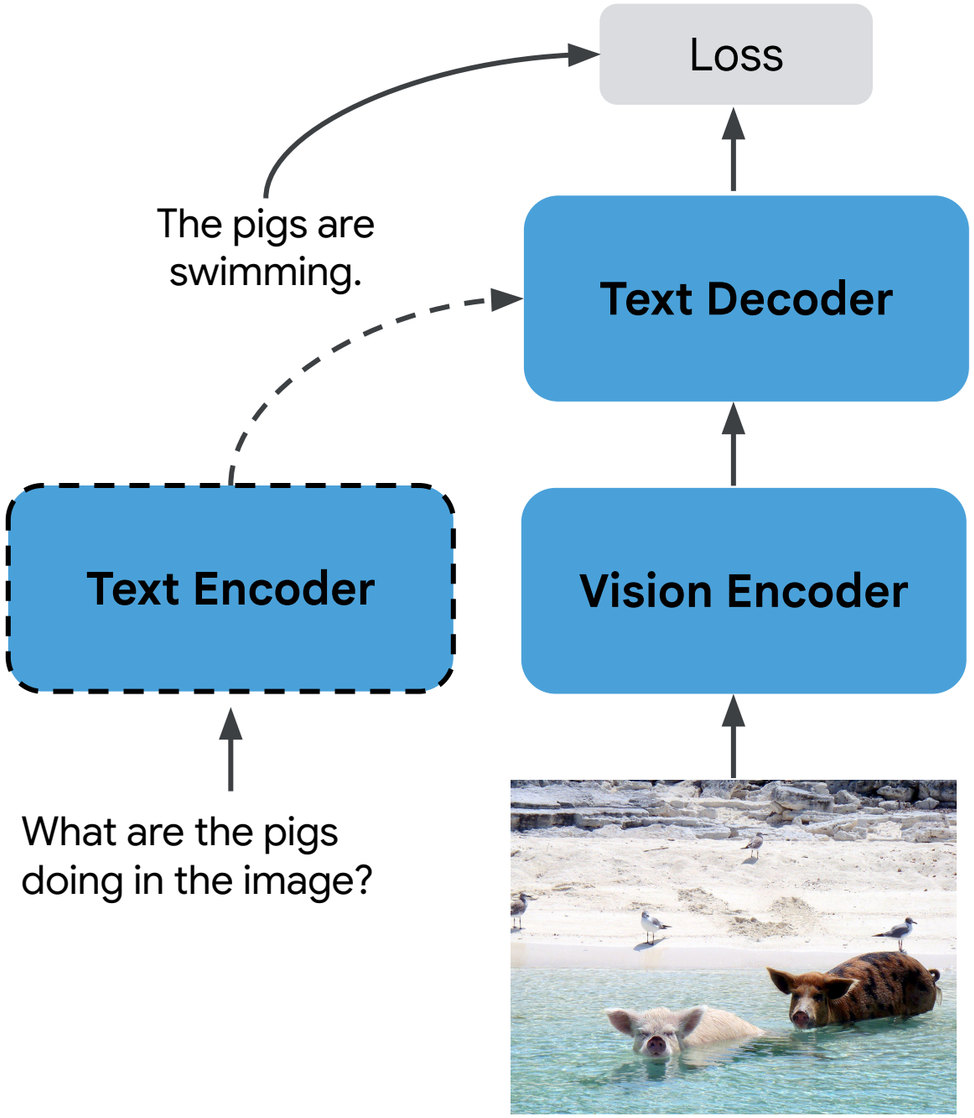
Encoder-Decoder Models

Strengths

- Simple: end-to-end training
- Enabled a new category of vision-language tasks
- Modular: can use various architectures for the encoders and decoder

Weaknesses

- Fixed-length vector encoding causes information bottleneck



Encoder-Decoder Model

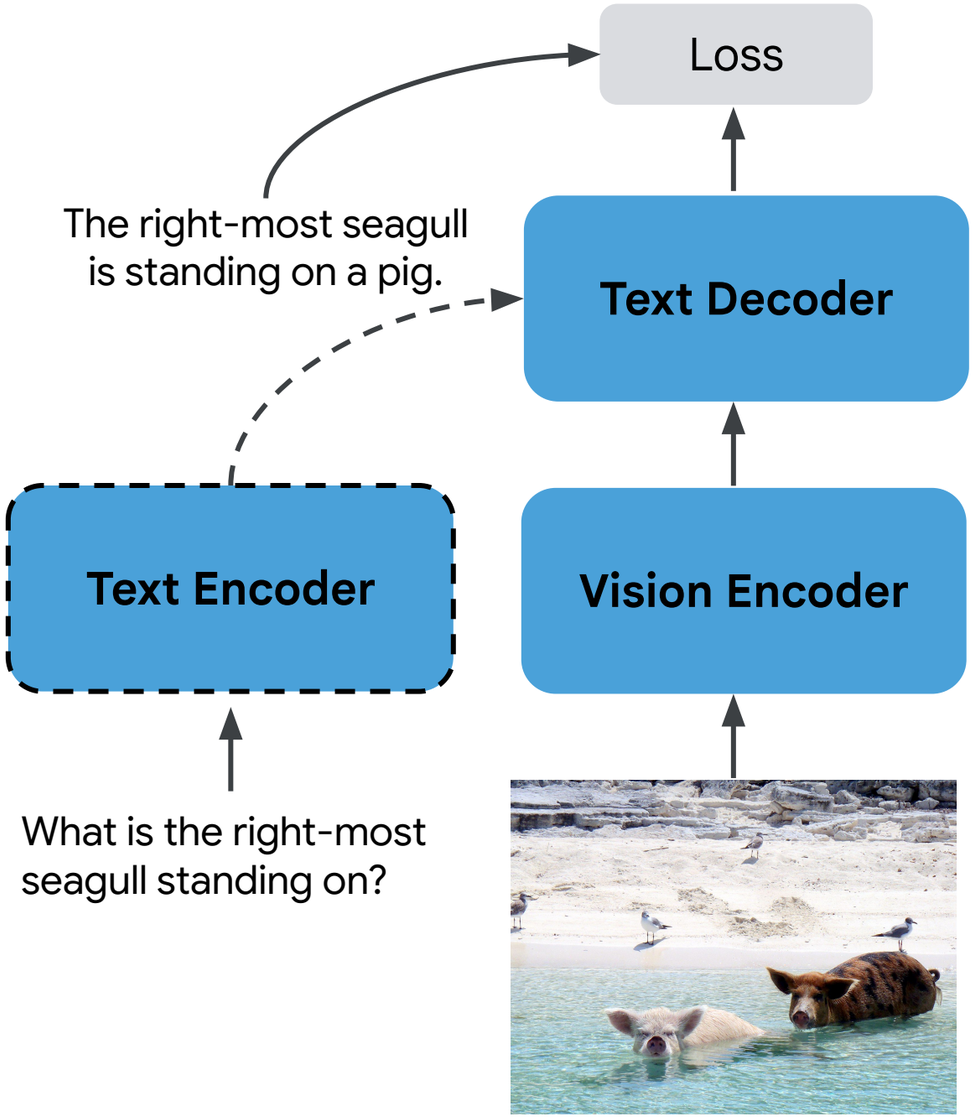
Encoder-Decoder Models

Strengths

- Simple: end-to-end training
- Enabled a new category of vision-language tasks
- Modular: can use various architectures for the encoders and decoder

Weaknesses

- Fixed-length vector encoding causes information bottleneck
- No cross-attention between encoders
- Lacks explicit grounding and alignment



Encoder-Decoder Model

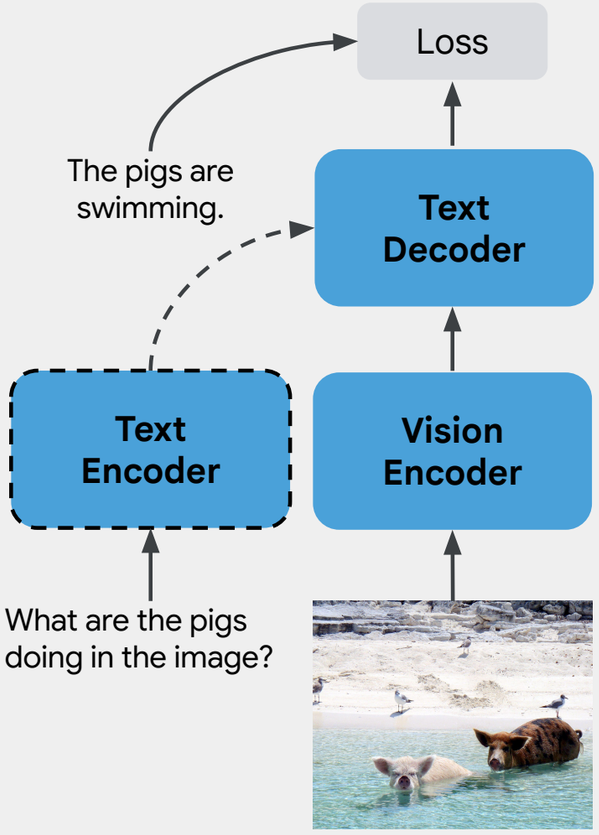
Vision-Language Model Architectures

Encoder-Decoder

Dual-Encoder

Cross-Modal

Natively Multimodal



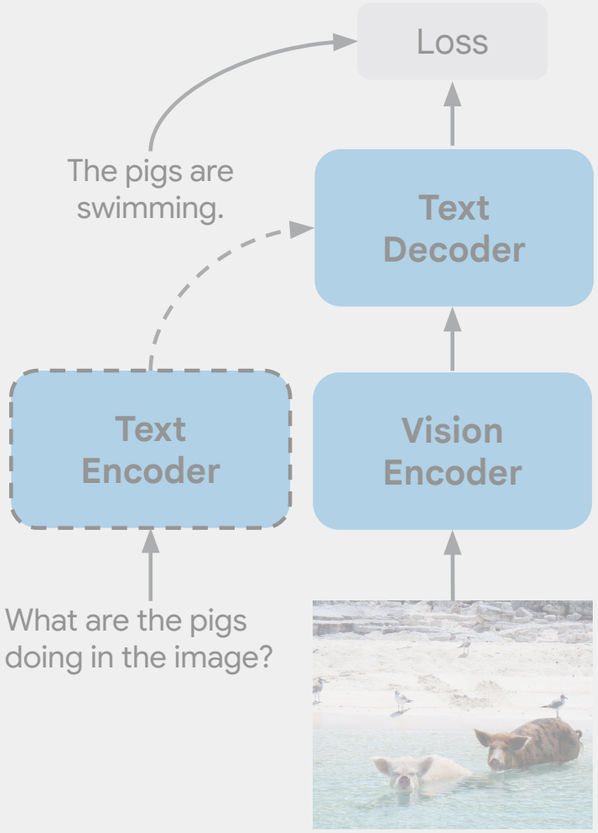
Vision-Language Model Architectures

Encoder-Decoder

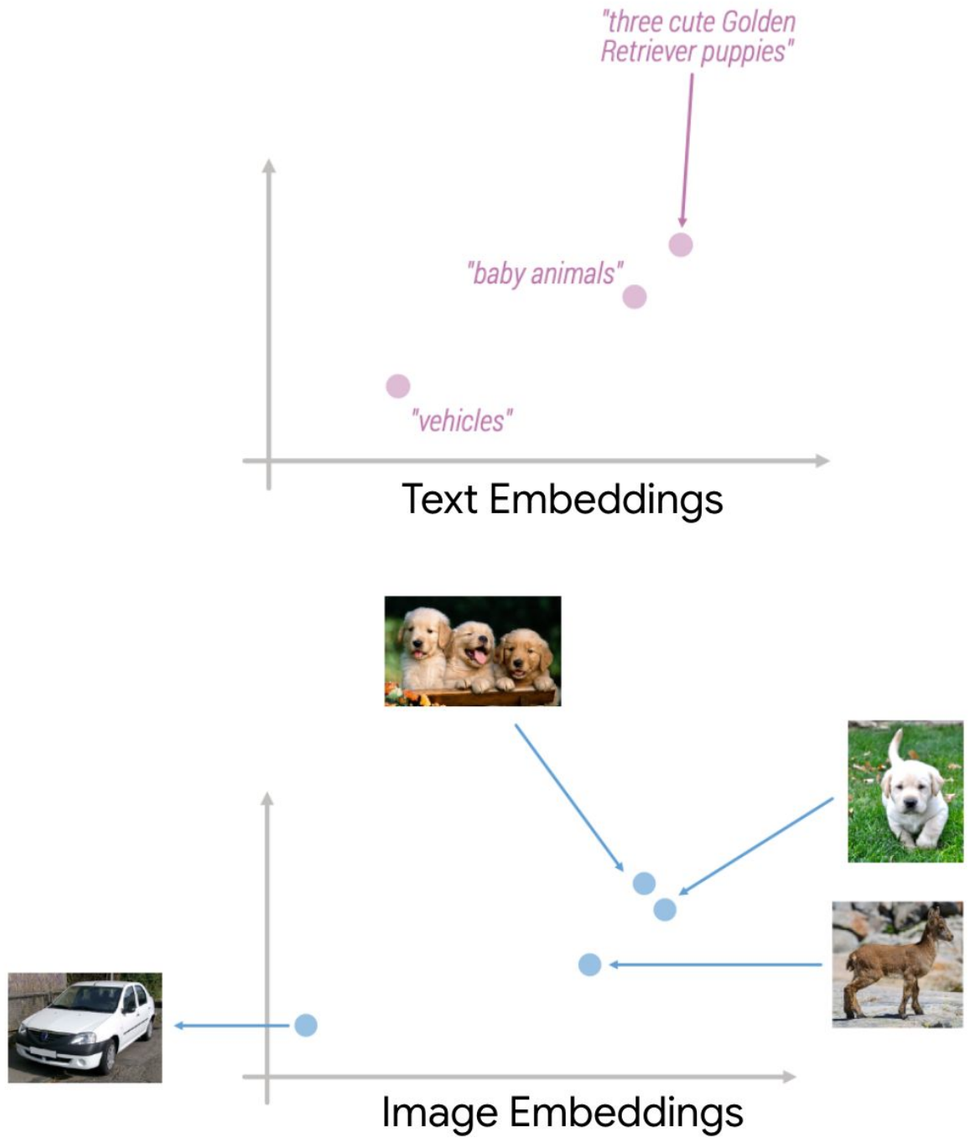
Dual-Encoder

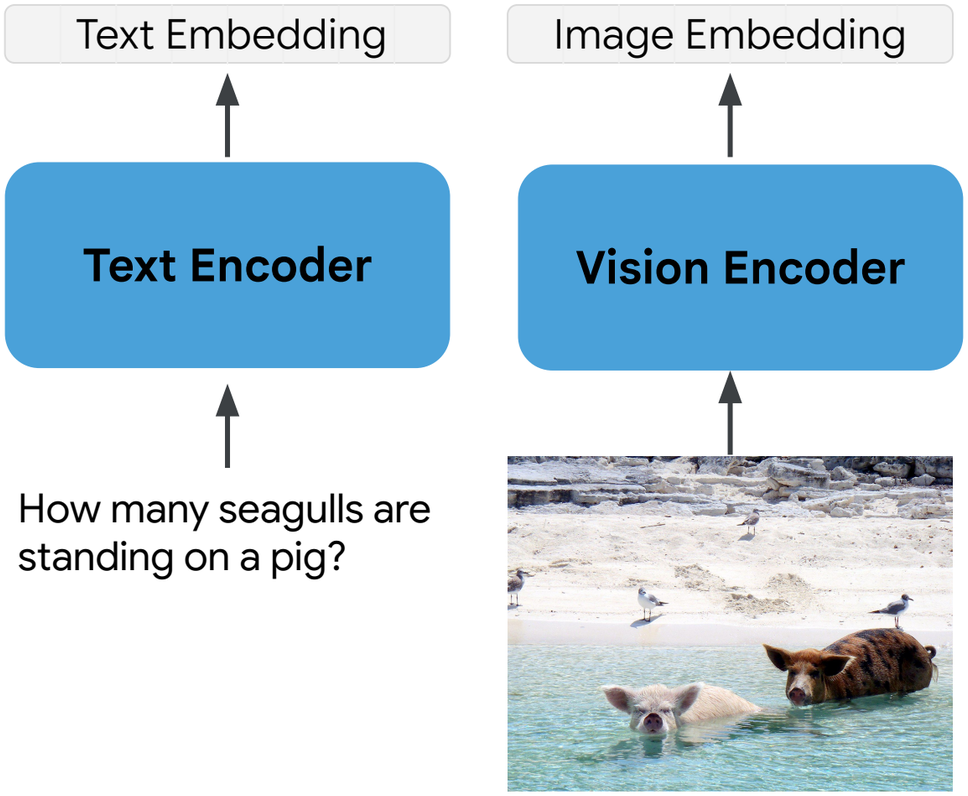
Cross-Modal

Natively Multimodal



Intuition: A shared embedding space





Intuition: A shared embedding space

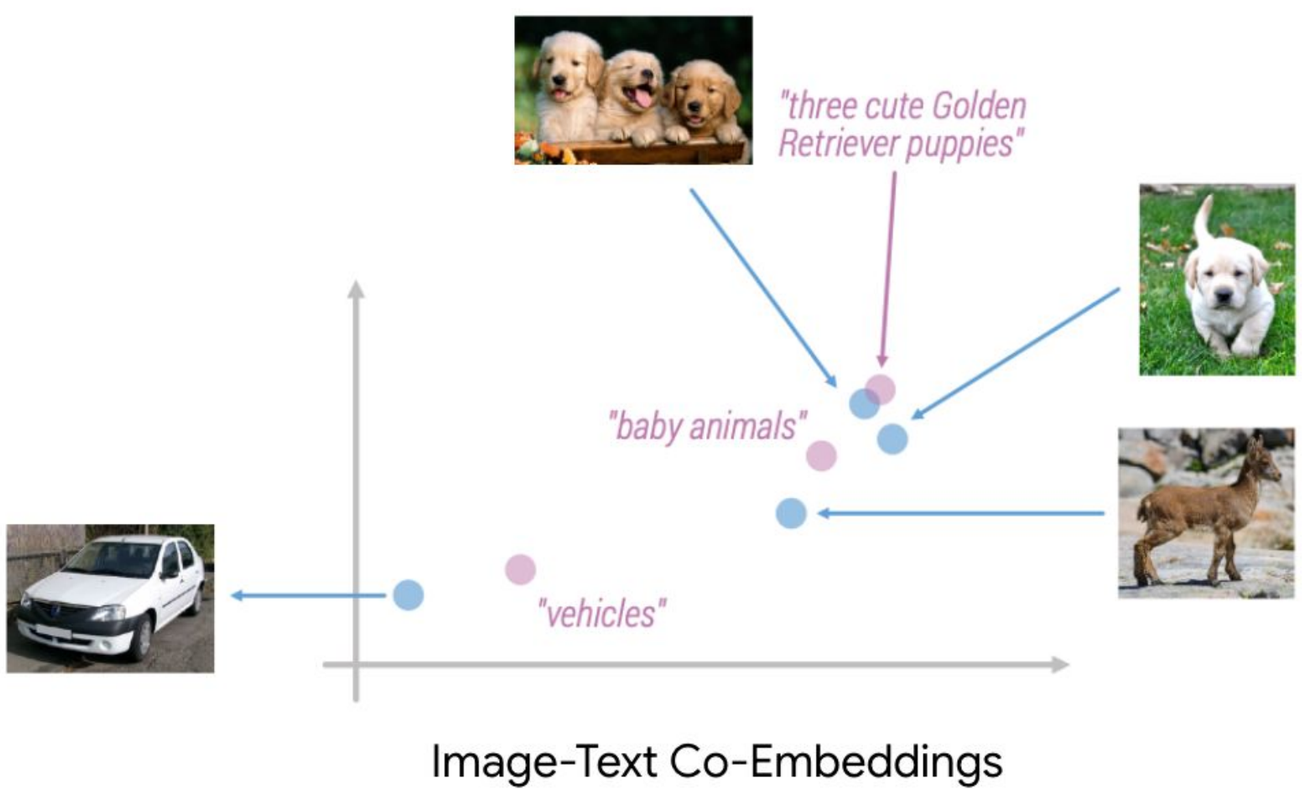
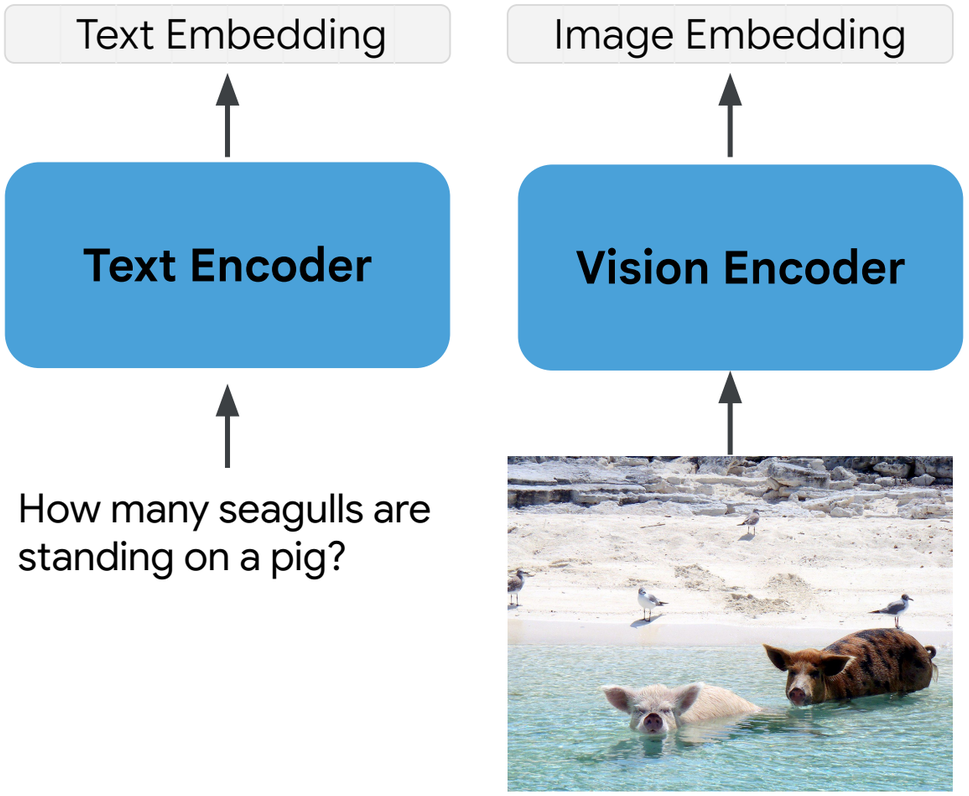


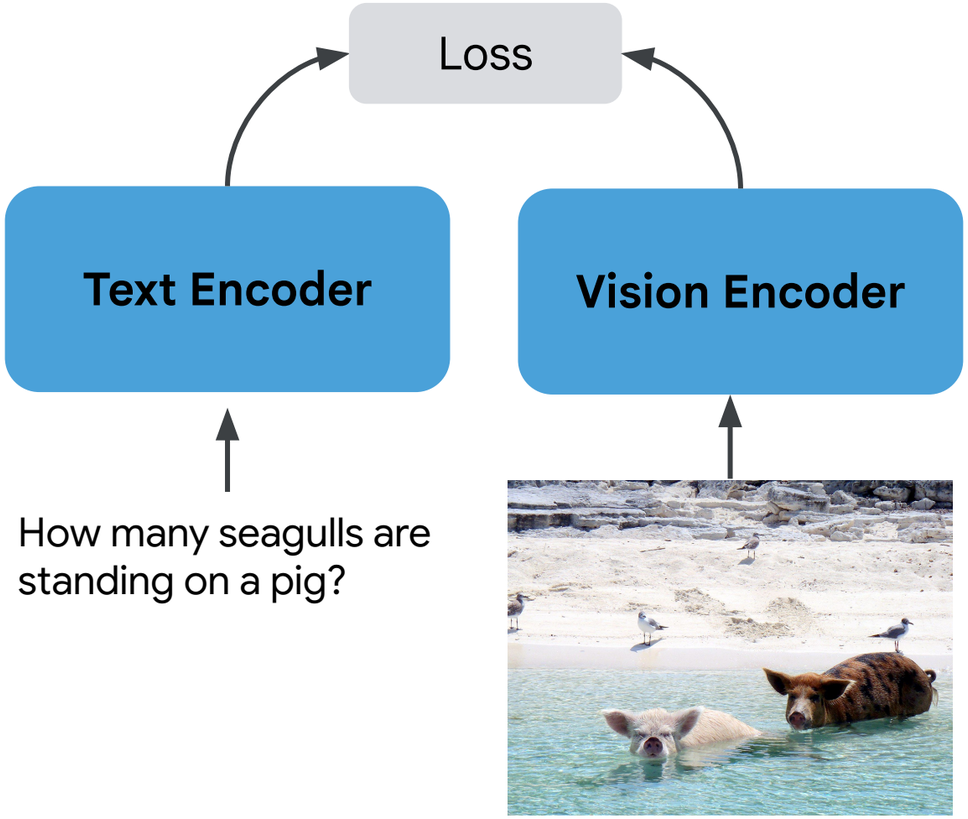
Image-Text Co-Embeddings



Dual-Encoder Models

Architecture

- Each modality is encoded separately.
- Modality fusion is done in the final stages of the encoder.



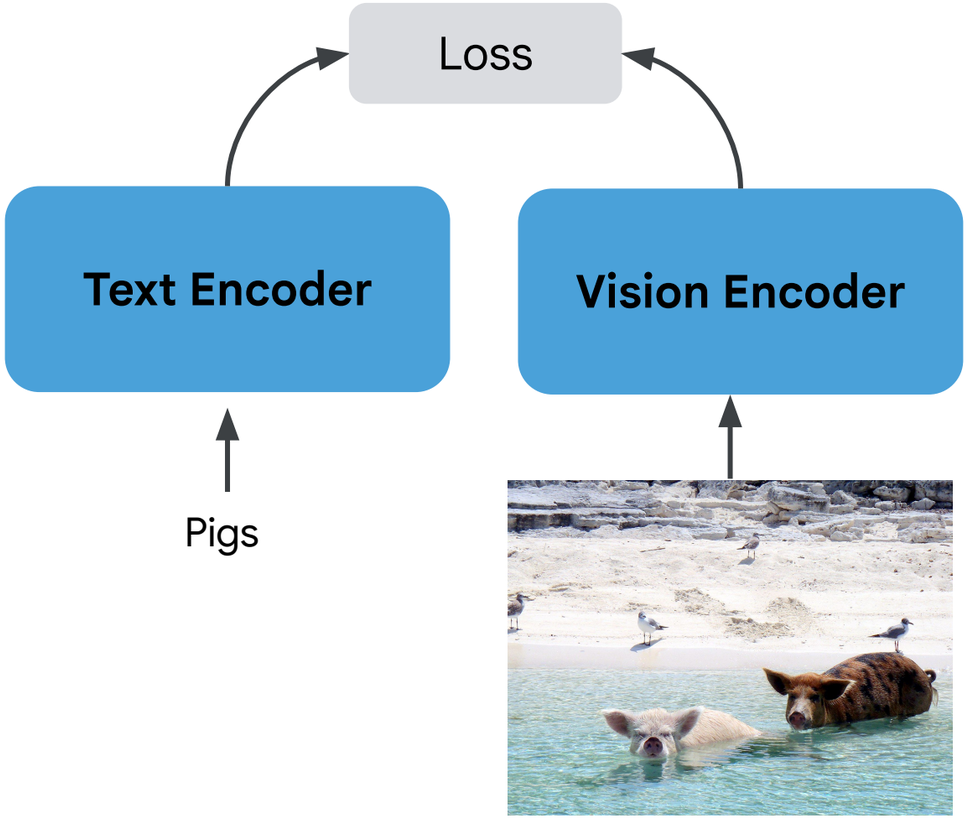
Dual-Encoder Models

Architecture

- Each modality is encoded separately.
- Modality fusion is done in the final stages of the encoder.

Training Strategy

- **Early approaches:** First learn to embed images and text separately, then align their embeddings on labeled image-text corpora
  [Frome et. al 2013; Socher et. al 2013]



Dual-Encoder Models

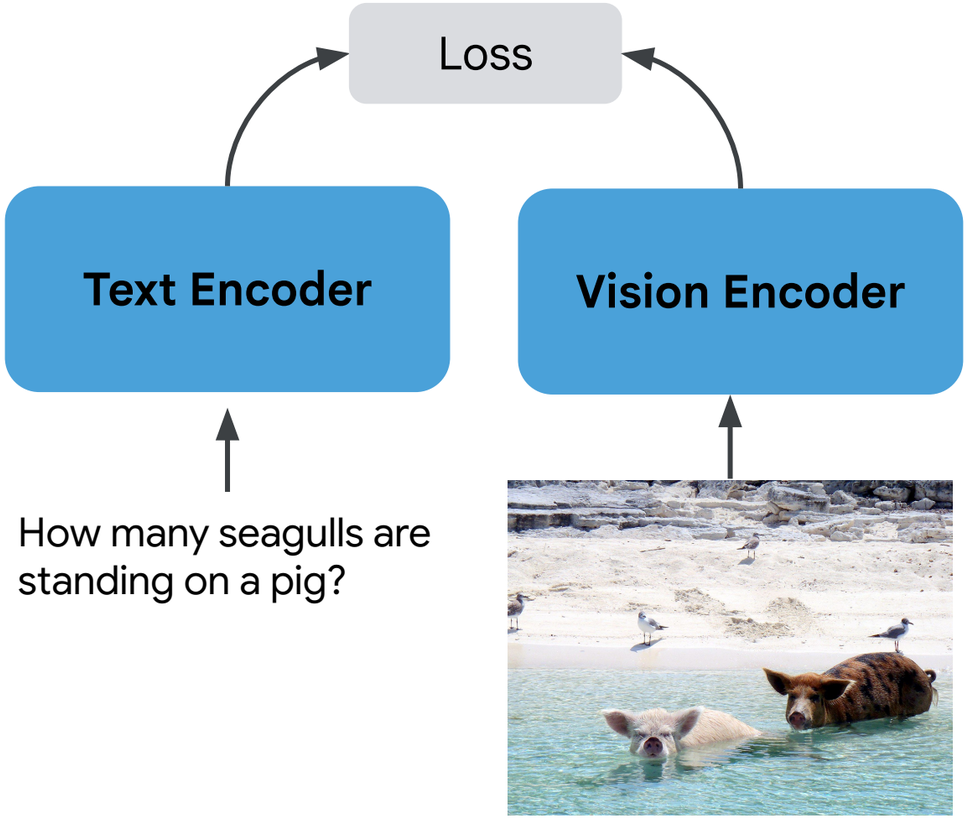
Architecture

- Each modality is encoded separately.
- Modality fusion is done in the final stages of the encoder.

Training Strategy

- **Early approaches:** First learn to embed images and text separately, then align their embeddings on labeled image-text corpora
  [Frome et. al 2013; Socher et. al 2013]
- **A revolution in the field**: Image-text co-embeddings are *jointly* learned on *large* datasets of paired image-text data scraped from the web
  [Radford et. al, 2021; Jia et. al, 2021]

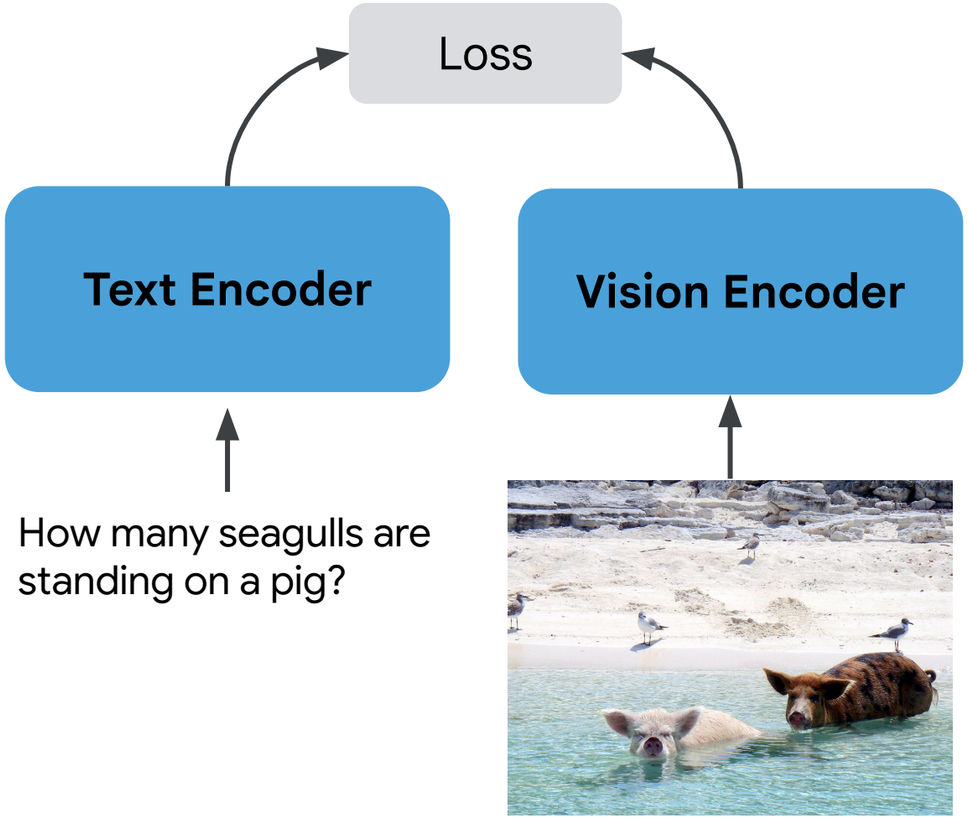
CLIP ALIGN



CLIP: Contrastive Language–Image Pre-training

Key idea

- **Data:** Trade off quality for scale → scrape a large dataset (400M) of image-text pairs (image and corresponding alt-text) from the web.

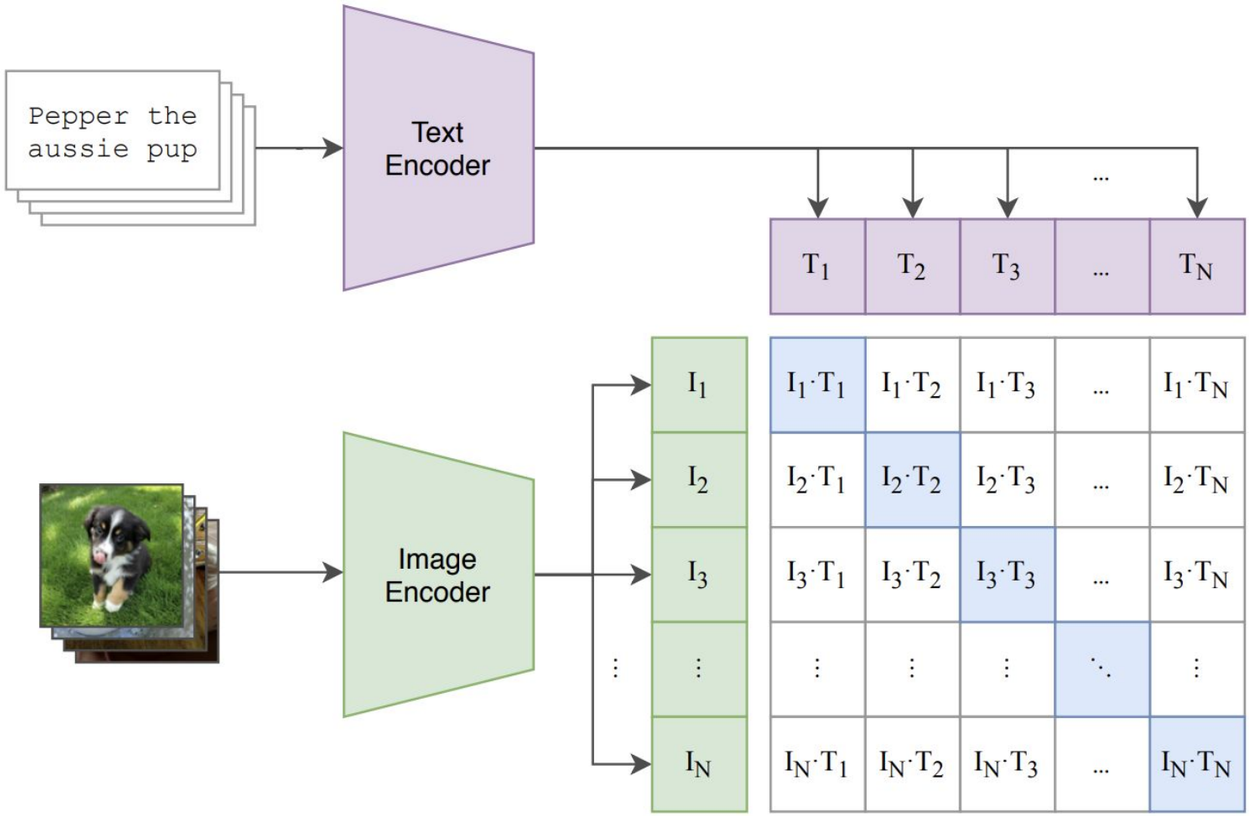


Radford et. al, *Learning Transferable Visual Models From Natural Language Supervision*, ICML 2021

CLIP: Contrastive Language–Image Pre-training

Key idea

- **Data:** Trade off quality for scale → scrape a large dataset (400M) of image-text pairs (image and corresponding alt-text) from the web.
- **Task**: In a batch containing multiple image-text pairs, make each image embedding `I` similar to its corresponding text embedding `T`, and dissimilar with all others.

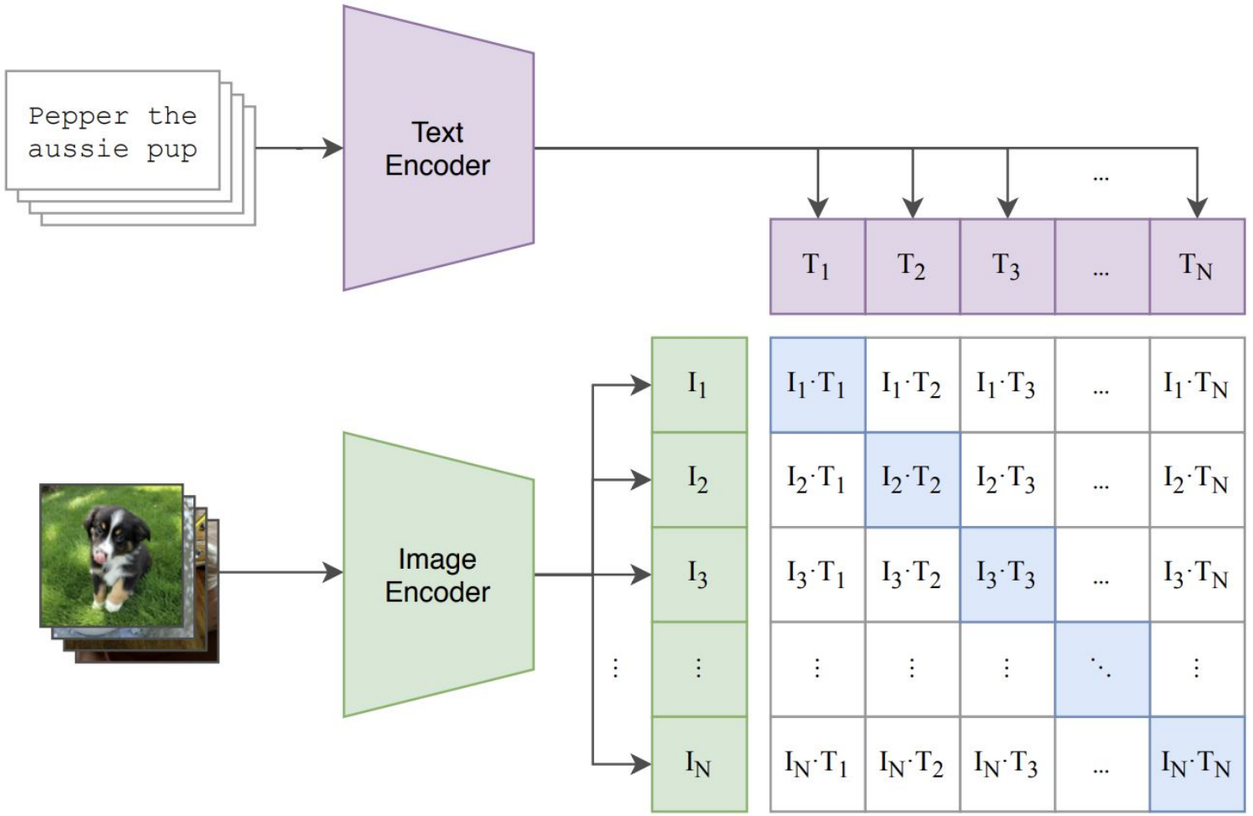


Radford et. al, *Learning Transferable Visual Models From Natural Language Supervision*, ICML 2021

CLIP: Contrastive Language–Image Pre-training

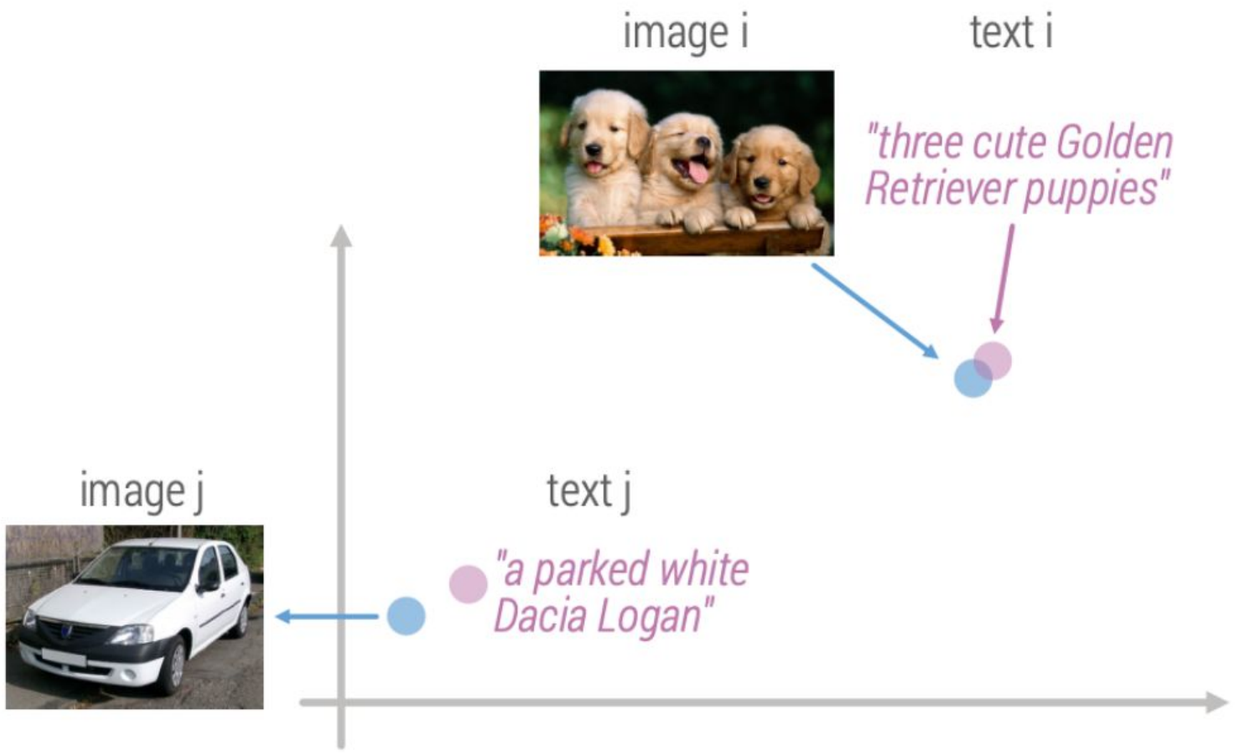
Key idea

- **Data:** Trade off quality for scale → scrape a large dataset (400M) of image-text pairs (image and corresponding alt-text) from the web.
- **Task**: In a batch containing multiple image-text pairs, make each image embedding `I` similar to its corresponding text embedding `T`, and dissimilar with all others.
- **Loss**: Contrastive loss.



Radford et. al, *Learning Transferable Visual Models From Natural Language Supervision*, ICML 2021

Contrastive Loss



- Consider two arbitrary (image, text) pairs, i and j.
- Want the embedding of image i to be closer to the embedding of text i than the embedding of text j.
- How: treat this as a classification problem where we try to guess which text corresponds to which image.

Contrastive Loss

Optimize this such that text `i` has the highest probability for image i.

We can do this using the cross-entropy loss:

$$-\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\sigma)}$$

x_i = image embedding for pair i

y_i = text embedding for pair i

y_j = text embedding for pair j

N = batch size

σ = temperature

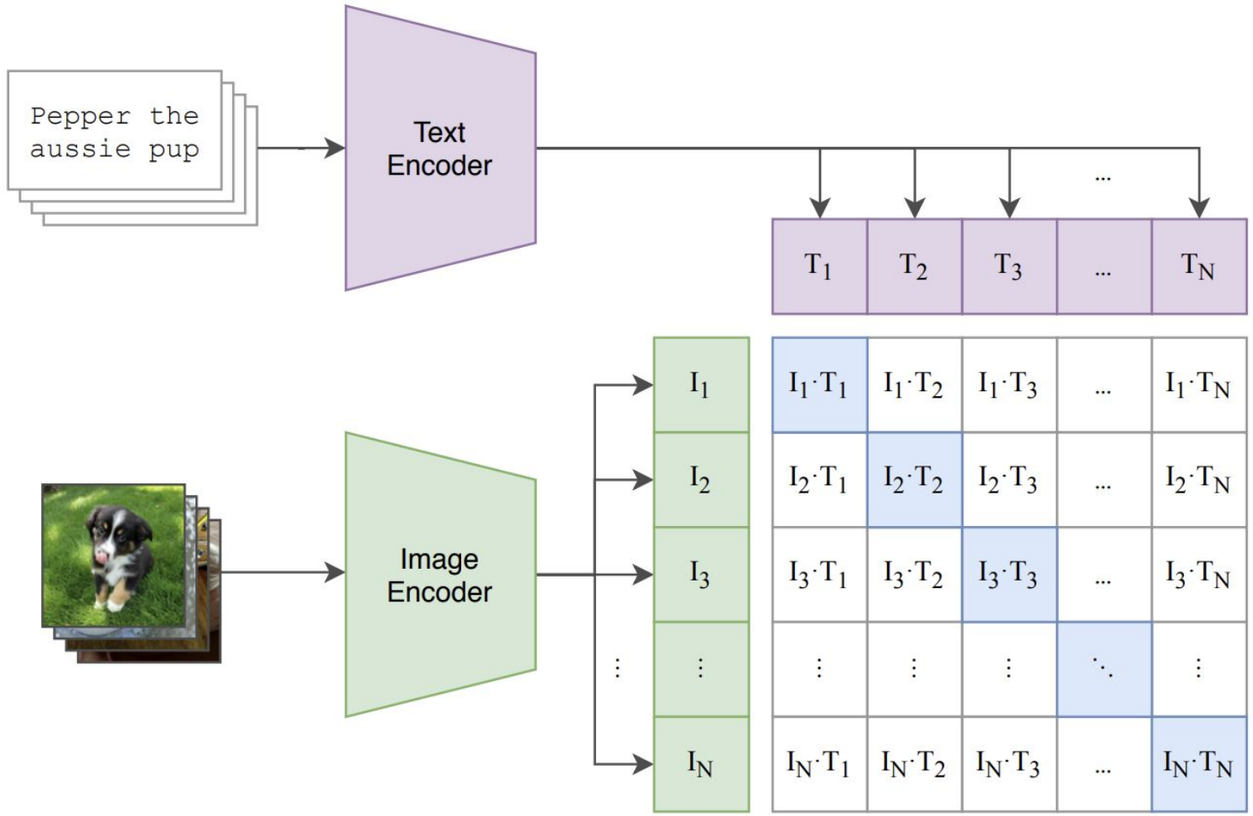
CLIP: Contrastive Language–Image Pre-training

Image-to-text contrastive objective:

$$L_{i2t} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\sigma)}$$

Text-to-image contrastive objective

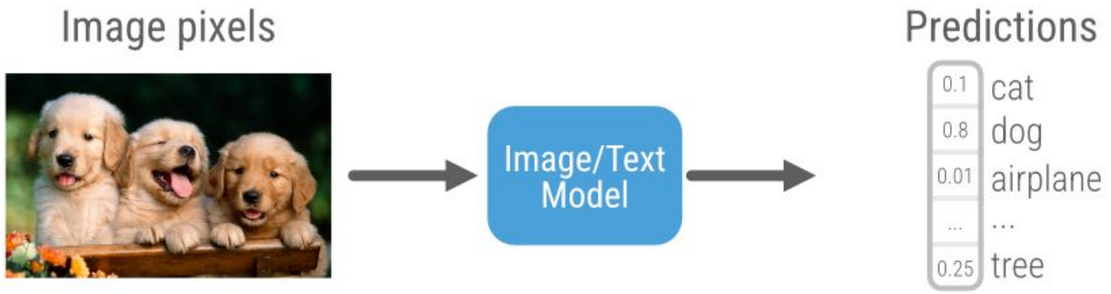
$$L_{t2i} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\sigma)}$$



Radford et. al, *Learning Transferable Visual Models From Natural Language Supervision*, ICML 2021

Applications of dual-encoder models

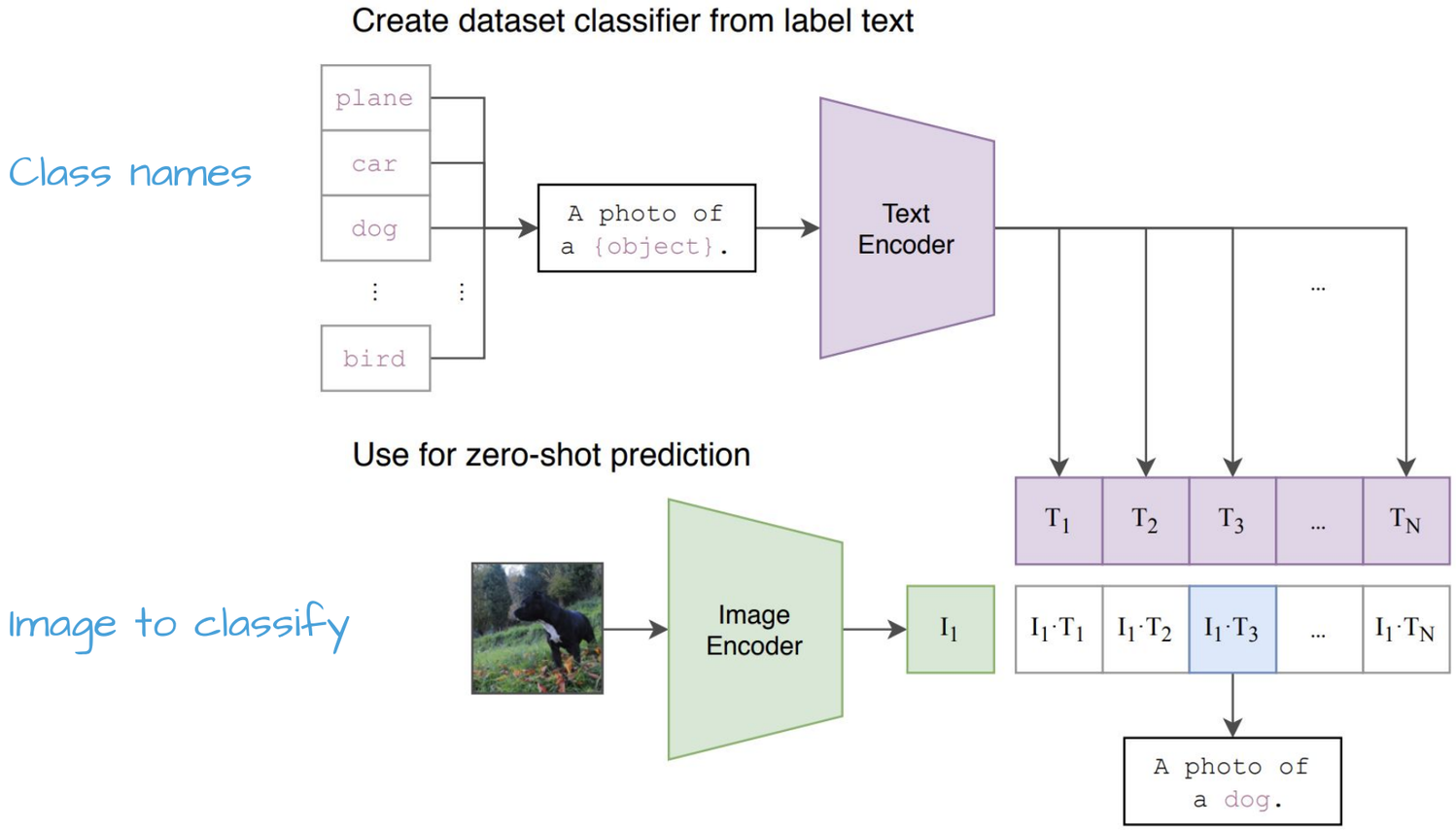
Before CLIP / ALIGN:



fixed set of classes

What the user needs:

Applications of dual-encoder models



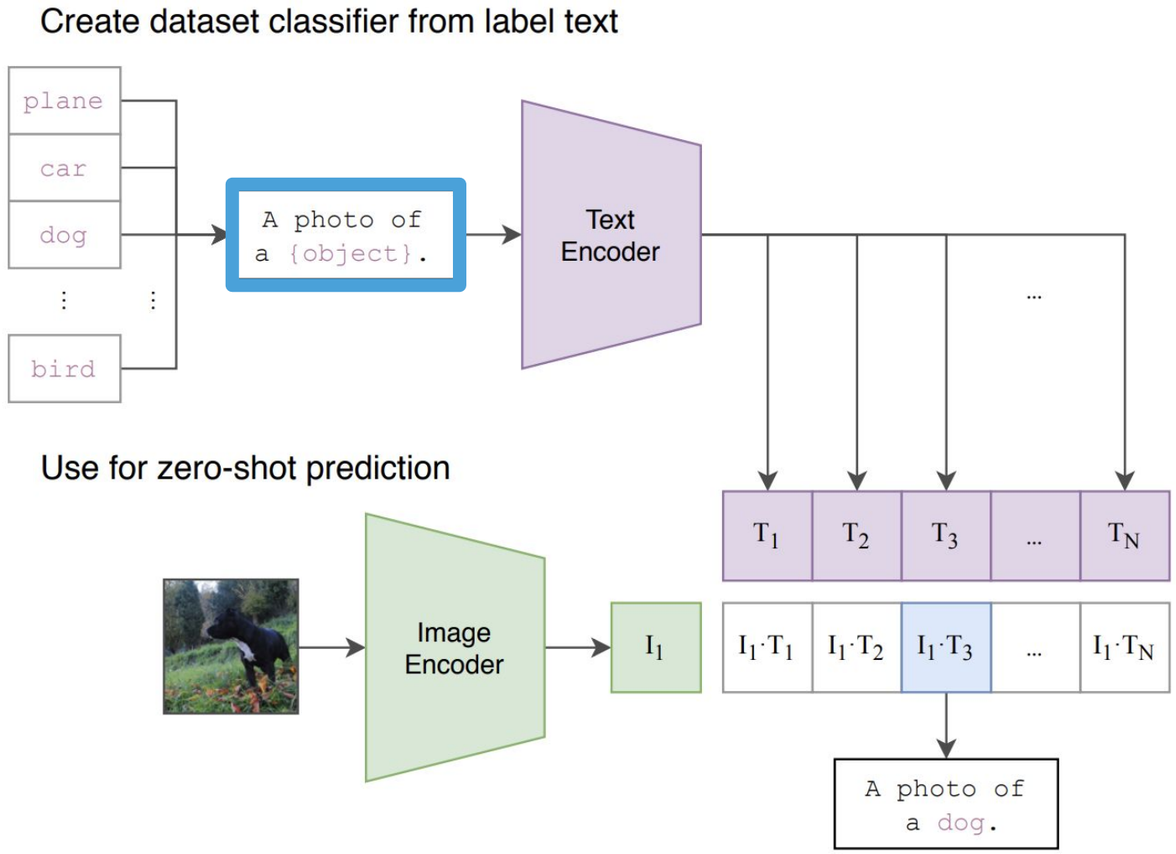
Open-Vocabulary Image Classification

Radford et. al, *Learning Transferable Visual Models From Natural Language Supervision*, ICML 2021

Applications of dual-encoder models

Prompt Engineering

- Radford et. al, 2021 found that using the template "`A photo of a {label}.`" improves accuracy on ImageNet by 1.3%
  → **bridges the train/test data distribution gap.**
- Customizing the prompt template based on the task also helps. E.g., "`A photo of a {label}, a type of pet.`" worked well for fine-grained classification
  → **provides more context.**
- Still important for state-of-the-art models today!

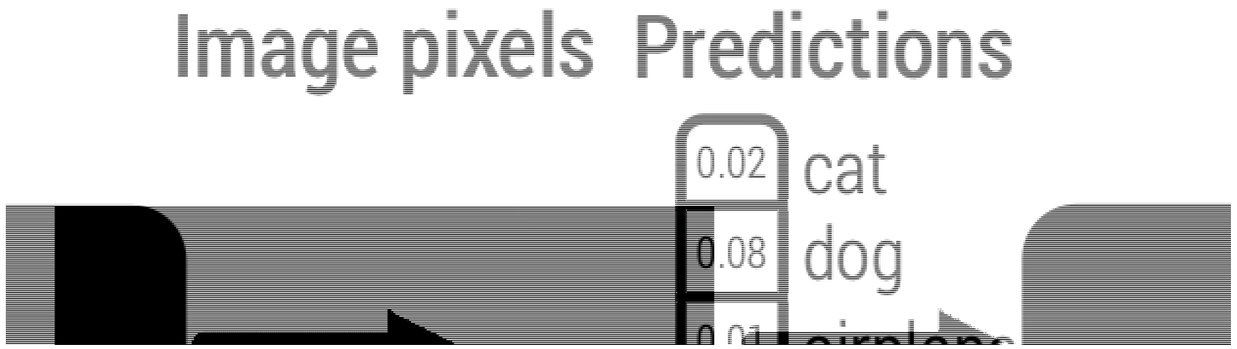


Open-Vocabulary Image Classification

Radford et. al, *Learning Transferable Visual Models From Natural Language Supervision*, ICML 2021

Applications of dual-encoder models

With CLIP / ALIGN:



What the user needs:

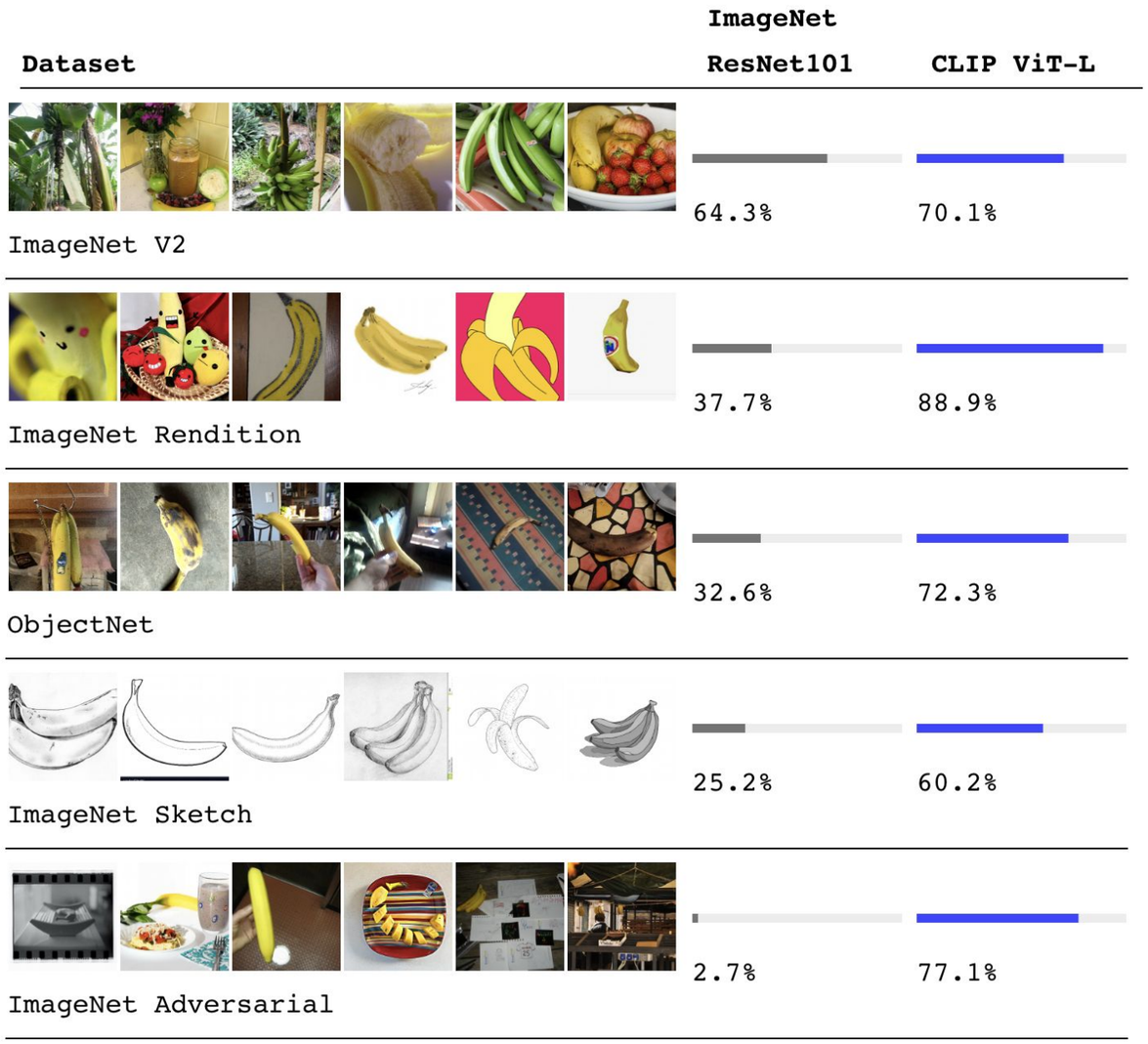
Applications of dual-encoder models

Dataset	ImageNet ResNet101	ImageNet CLIP ViT-L
 ImageNet	76.2%	76.2%

Matches the accuracy of ResNet 101 (trained on human-labeled data) with no human labels at all!

Radford et. al, *Learning Transferable Visual Models From Natural Language Supervision*, ICML 2021

Applications of dual-encoder models



Radford et. al, *Learning Transferable Visual Models From Natural Language Supervision*, ICML 2021

How do weakly labeled models beat supervised models?

	ImageNet Resnet 101		CLIP ViT-L
# Parameters	44.5 M	X 7 →	307 M
Data	1.2 M	X 300 →	400 M

Scale!

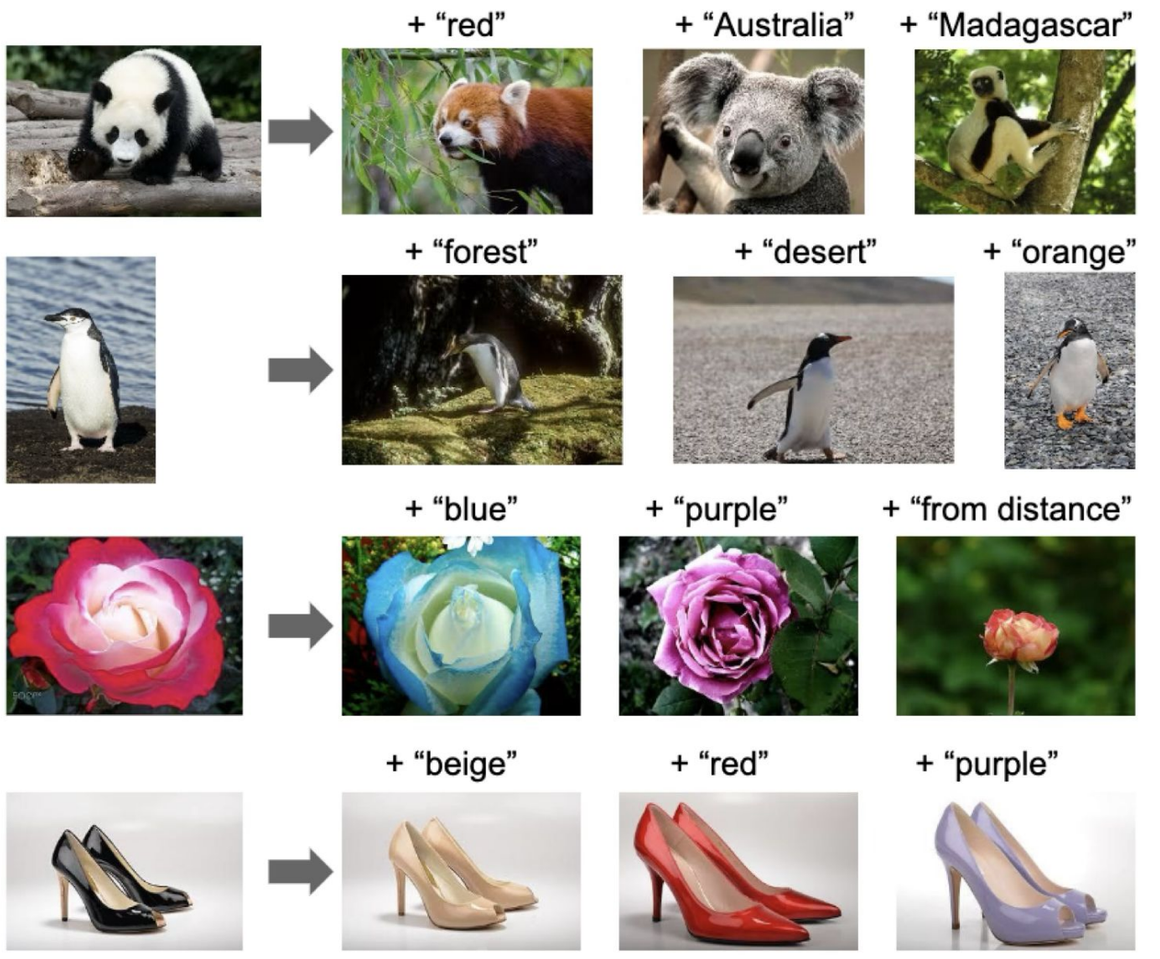
Applications of dual-encoder models

Zero-Shot Image Retrieval

ALIGN

Jia et. al, *Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision*, ICML 2021

Applications of dual-encoder models



Compositional Relations

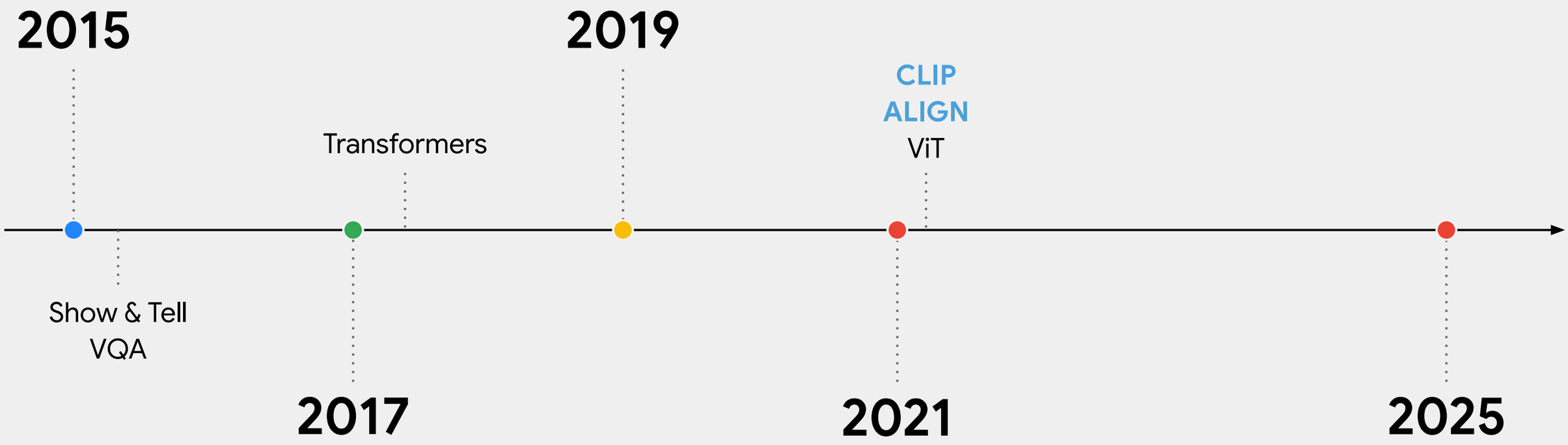
ALIGN

Jia et. al, *Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision*, ICML 2021

CLIP Adaptation to Various Domains

Domain	Example Paper	Achievement
General Biomedical	BMCA-CLIP [Lozano et. al, CVPR 2025]	SoTA on 40 biomedical tasks
Remote Sensing	GRAFT [Mall et. al, ICLR 2024]	Zero-shot tasks for satellite images
Robotics	Robotic-CLIP [Nguyen et. al, ICRA 2025]	Language-driven robotic tasks
Medical Imaging	MedCLIP-SAM [Koleilat et. al, MICCAI 2024]	Medical image segmentation using natural language prompts
Generative Art	StyleCLIP [Patashnik et. al, ICCV 2021]	Text-driven image editing without new annotations.

Vision-Language Models Timeline



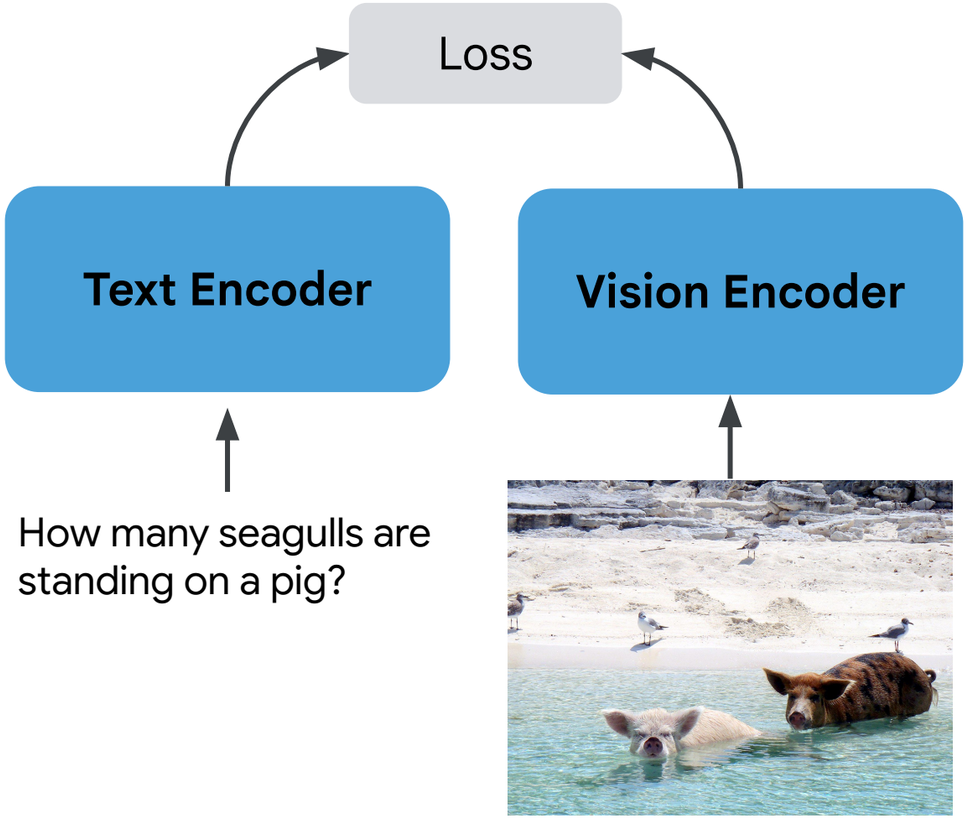
Dual-Encoder Models with Contrastive Loss

Advantages

- Open-vocabulary (zero-shot generalization)
- Fast & Efficient (e.g., can compute image and text embeddings in parallel)
- Scalable (e.g., fast retrieval over 5B images)

Weaknesses

- Sensitive to prompt format



Dual-Encoder Models with Contrastive Loss

Advantages

- Open-vocabulary (zero-shot generalization)
- Fast & Efficient (e.g., can compute image and text embeddings in parallel)
- Scalable (e.g., fast retrieval over 5B images)

Weaknesses

- Sensitive to prompt format
- **Image-level captions are insufficient supervision**

living room

pillows

couch

coffee table

Lacks spatial granularity

Dual-Encoder Models with Contrastive Loss

Advantages

- Open-vocabulary (zero-shot generalization)
- Fast & Efficient (e.g., can compute image and text embeddings in parallel)
- Scalable (e.g., fast retrieval over 5B images)

Weaknesses

- Sensitive to prompt format
- Image-level captions are insufficient supervision
- **Lacks scene-level alignment between image and text**

there is a mug in some grass

there is some grass in a mug

CLIP cannot distinguish

Thrush et. al, *Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality,* CVPR 2022

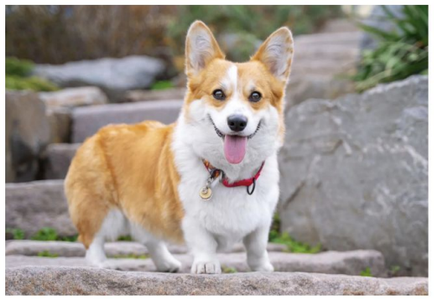
Dual-Encoder Models with Contrastive Loss

Advantages

- Open-vocabulary (zero-shot generalization)
- Fast & Efficient (e.g., can compute image and text embeddings in parallel)
- Scalable (e.g., fast retrieval over 5B images)

Weaknesses

- Sensitive to prompt format
- Image-level captions are insufficient supervision
- Lacks scene-level alignment between image and text
- **Performance depends on batch size**



Batch size: 4	“animal”
Batch size: 100	“dog”
Batch size: **32000**	“Welsh Corgi”

Increasing batch size allows for more fine-grained classification, but it is limited by computational resources.

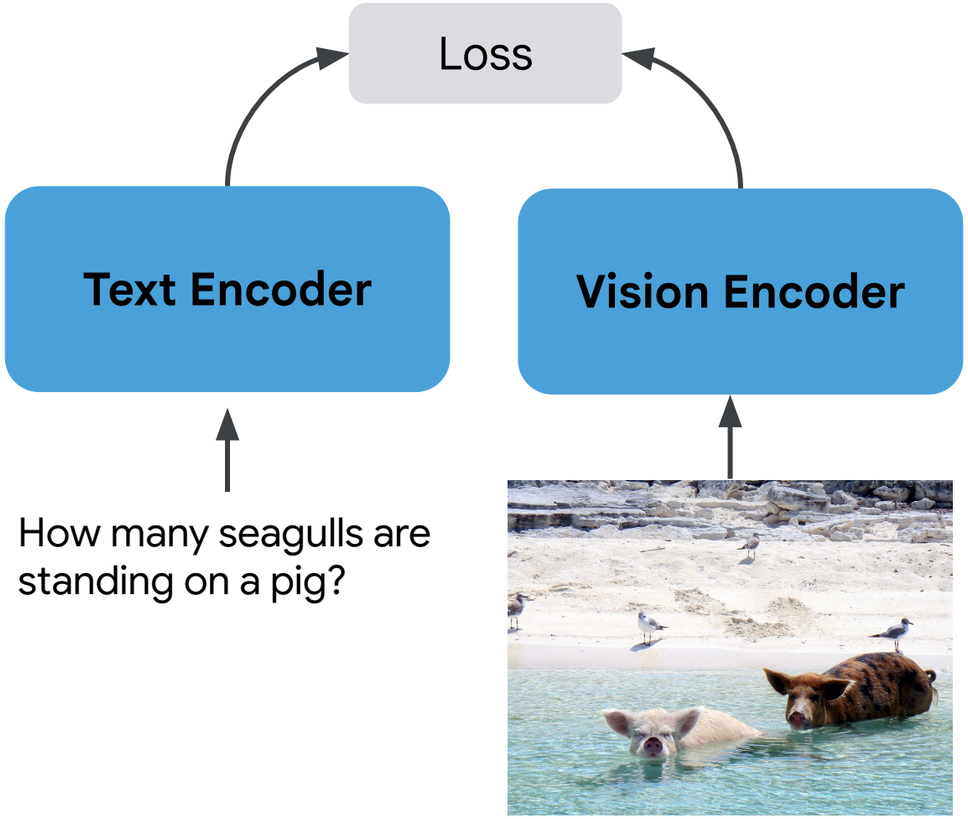
Dual-Encoder Models with Contrastive Loss

Advantages

- Open-vocabulary (zero-shot generalization)
- Fast & Efficient (e.g., can compute image and text embeddings in parallel)
- Scalable (e.g., fast retrieval over 5B images)

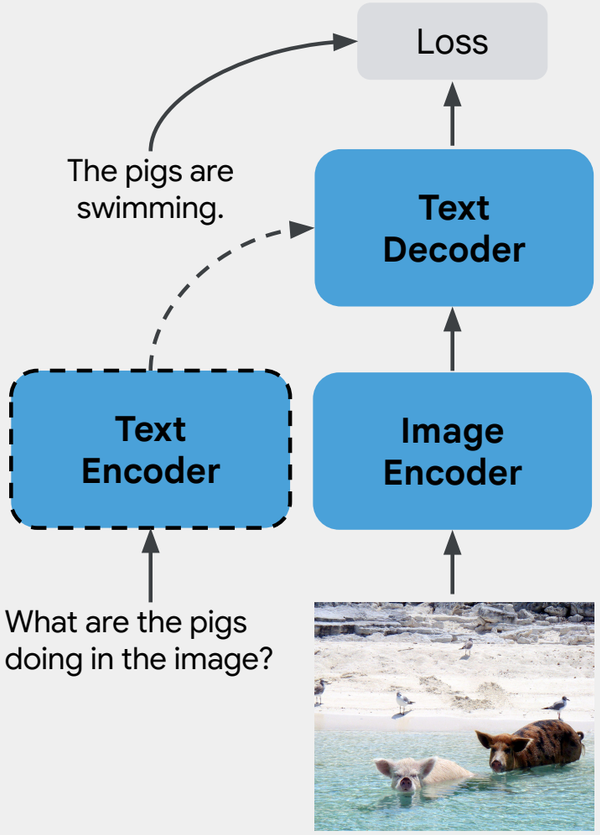
Weaknesses

- Sensitive to prompt format
- Image-level captions are insufficient supervision
- Lacks scene-level alignment between image and text
- Performance depends on batch size
- **Cannot generate text**

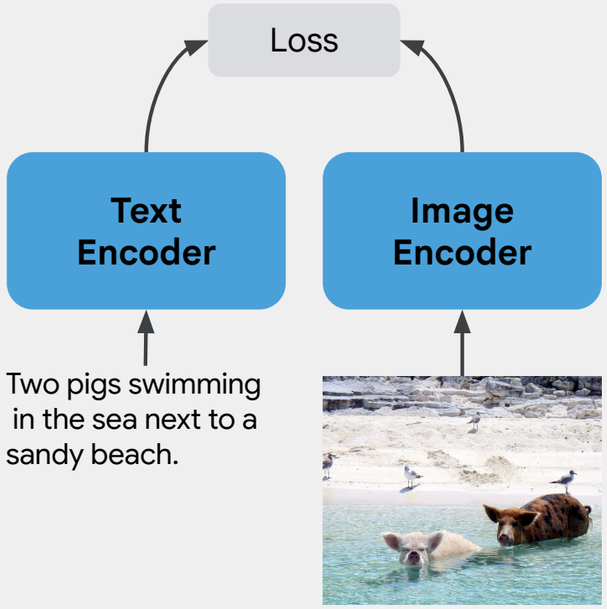


Vision-Language Model Architectures

Encoder-Decoder



Dual-Encoder



Cross-Modal

Natively Multimodal

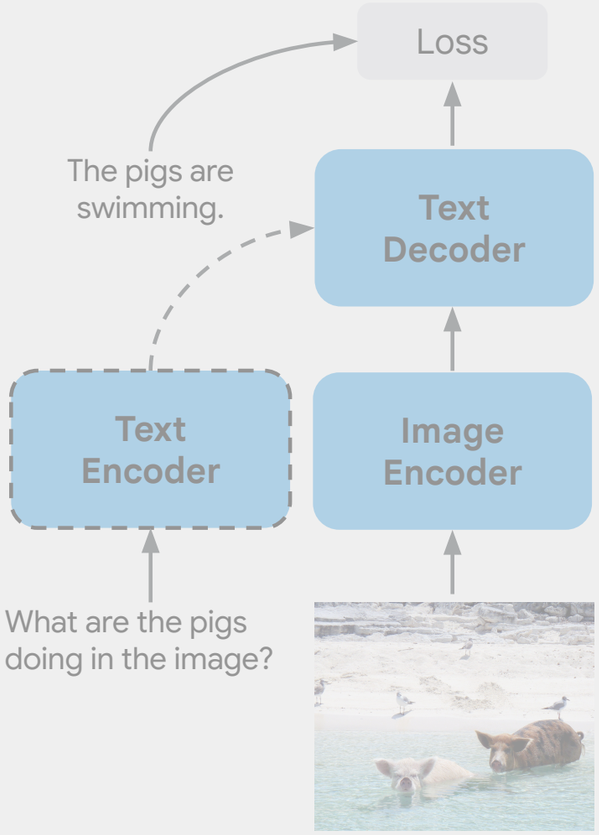
Vision-Language Model Architectures

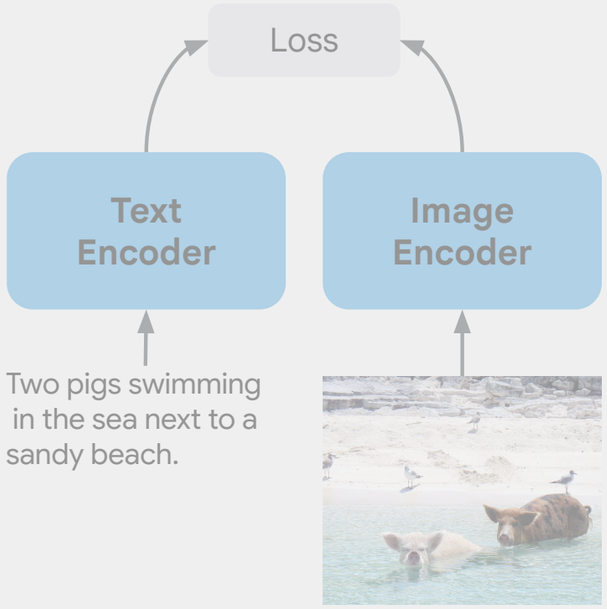
Encoder-Decoder

Dual-Encoder

Cross-Modal

Natively Multimodal





Cross-Modal Models: Motivation

Encoder-Decoder Models

1. No shared embedding space
2. No cross attention
3. Information bottleneck

thus they are not ideal for alignment tasks such as **image or video retrieval**, or tasks requiring grounding such as **counting**, **spatial reasoning**.

Dual-Encoder Models

1. No text generation
2. No cross attention

thus they cannot be applied to more complex vision-language tasks such as **visual question answering (VQA)** or **image captioning.**

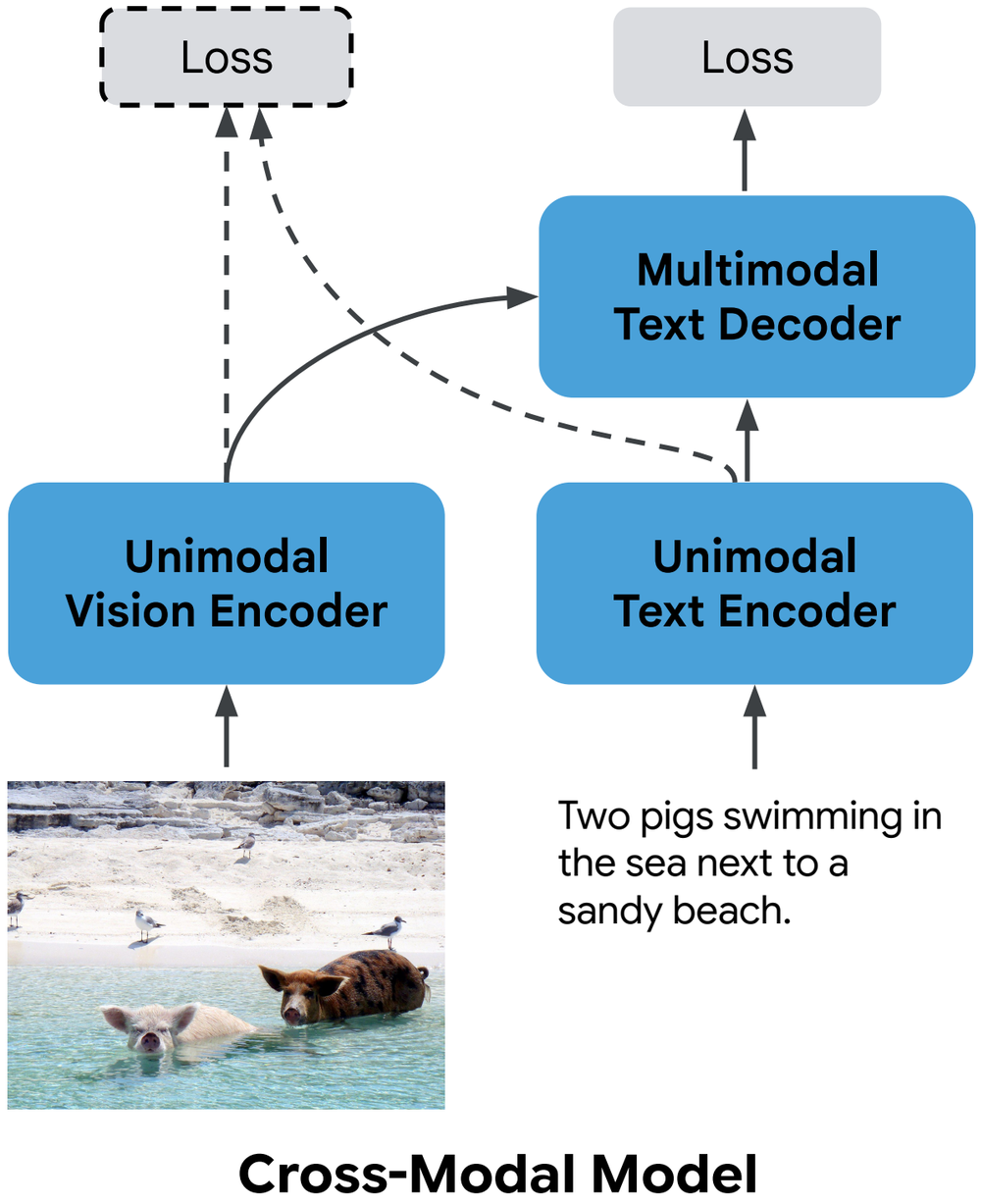
Cross-Modal Models

Key Idea

Merge the benefits of dual-encoder and encoder-decoder models

by

modeling more complex interactions between image and text to enable **fine-grained image-text** understanding and **more complex multimodal tasks.**



Cross-Modal Model

CoCa: Contrastive Captioners are Image-Text Foundation Models

Architecture

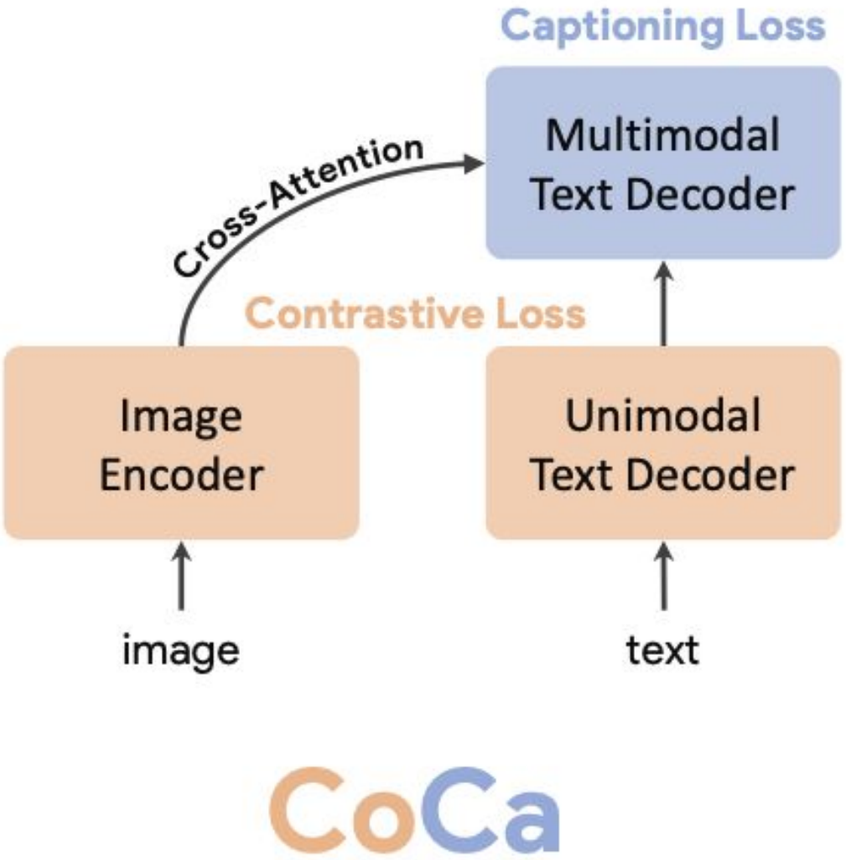
- Separate unimodal encoders with self-attention
- Joint multimodal decoder with cross-attention across modalities.

Training strategy

The optimization objective consists of two parts:

- A contrastive loss → learns global features
- A captioning loss → learns more fine-grained, local features

Everything is trained from scratch.



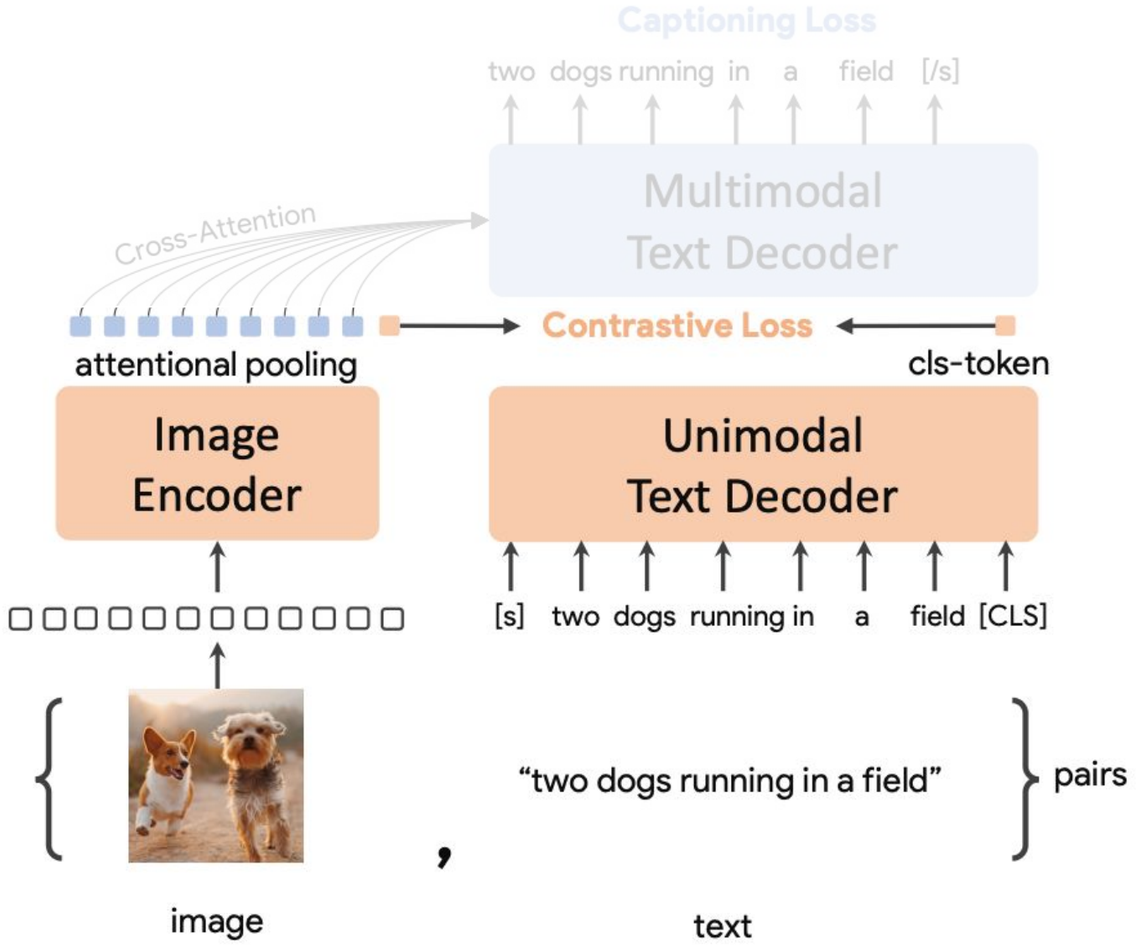
Yu et. al, *CoCa: Contrastive Captioners are Image-Text Foundation Models*, TMLR 2024

CoCa: Pretraining

$$\mathcal{L}_{\text{CoCa}} = \lambda_{\text{Con}} \cdot \mathcal{L}_{\text{Con}} + \lambda_{\text{Cap}} \cdot \mathcal{L}_{\text{Cap}}$$

contrastive loss captioning loss

$$\mathcal{L}_{\text{Con}} = -\frac{1}{N}\Bigg(\underbrace{\sum_i^N \log \frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^N \exp(x_i^\top y_j/\sigma)}}_{\text{image-to-text}} + \underbrace{\sum_i^N \log \frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^N \exp(y_i^\top x_j/\sigma)}}_{\text{text-to-image}}\Bigg)$$



similar to CLIP / ALIGN

Yu et. al, *CoCa: Contrastive Captioners are Image-Text Foundation Models*, TMLR 2024

CoCa: Pretraining

$$\mathcal{L}_{\text{CoCa}} = \lambda_{\text{Con}} \cdot \mathcal{L}_{\text{Con}} + \lambda_{\text{Cap}} \cdot \mathcal{L}_{\text{Cap}}$$

contrastive loss

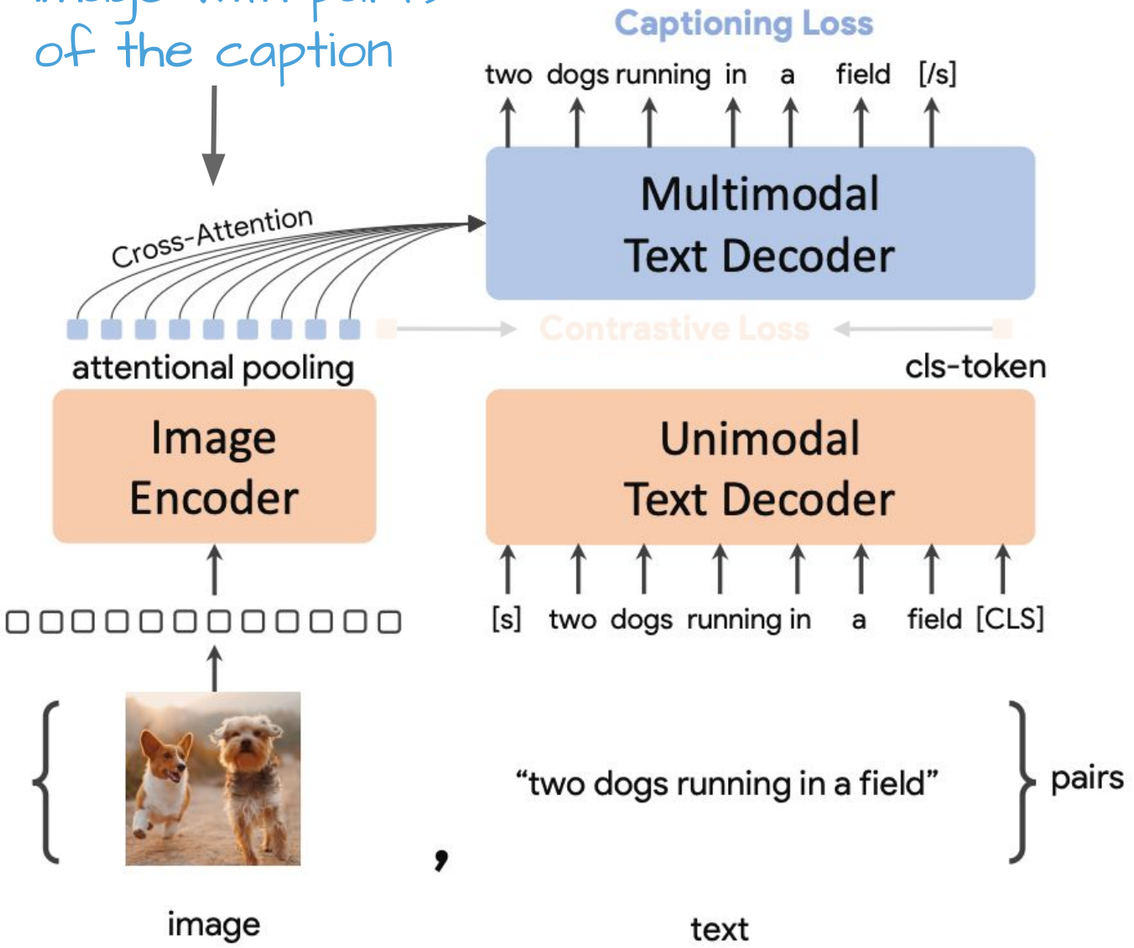
captioning loss

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

text caption

image encoding

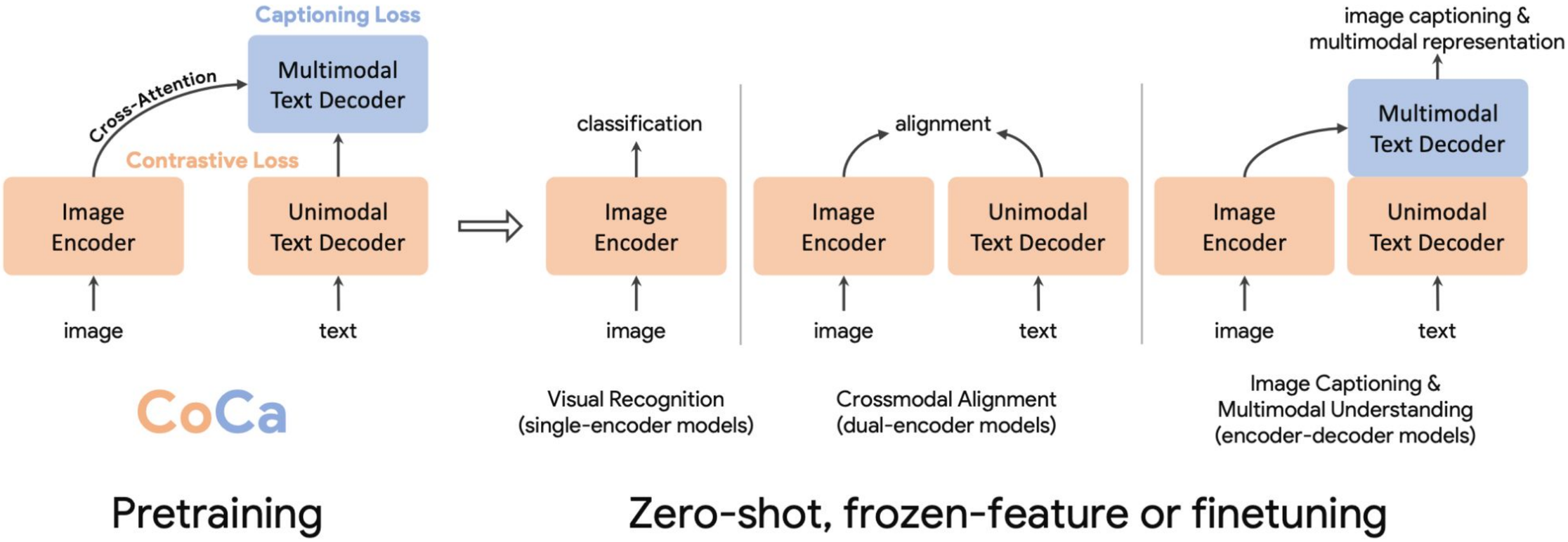
Enables aligning parts of the image with parts of the caption



similar to Show & Tell

Yu et. al, *CoCa: Contrastive Captioners are Image-Text Foundation Models*, TMLR 2024

CoCa: Fine-Tuning & Inference on Various Tasks

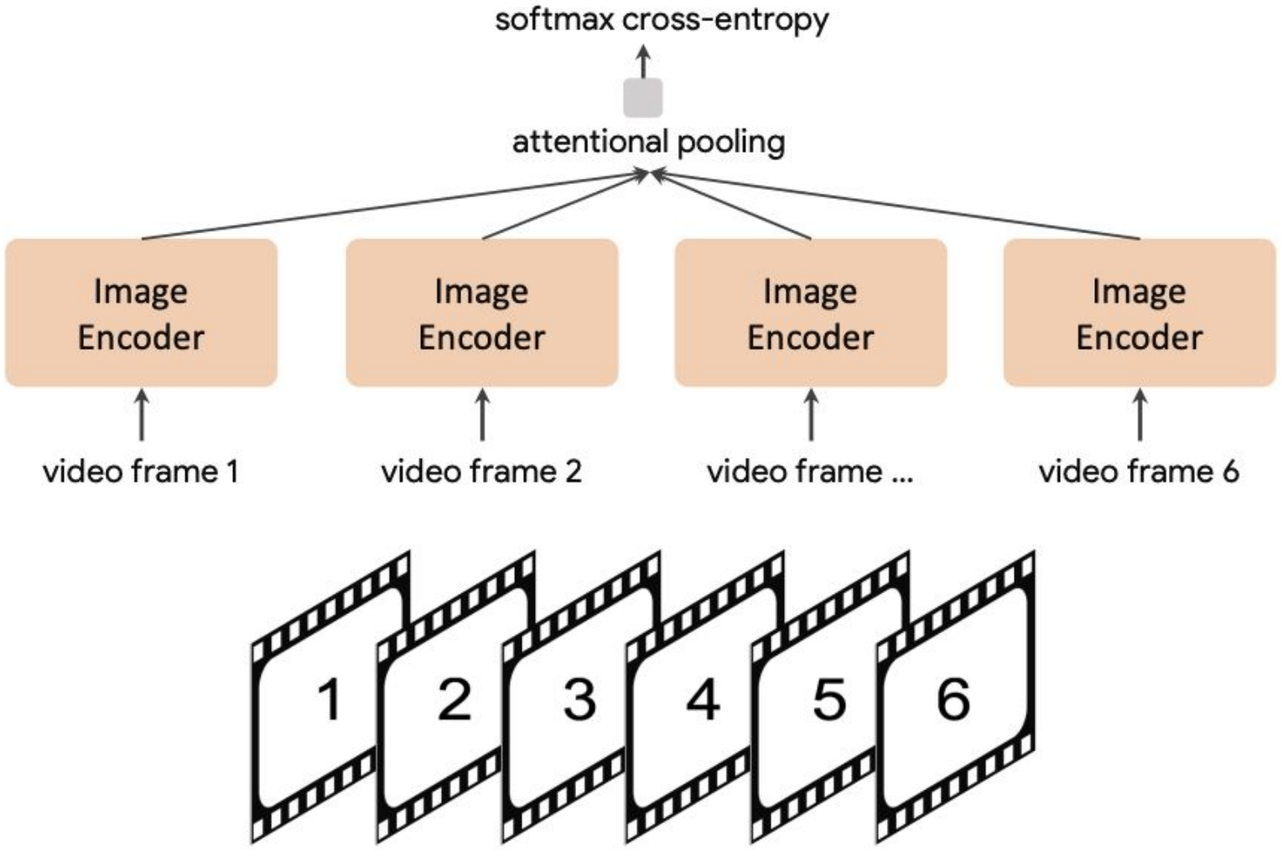


Yu et. al, *CoCa: Contrastive Captioners are Image-Text Foundation Models*, TMLR 2024

CoCa: Evaluation

SoTA (at the time) on many tasks, including:

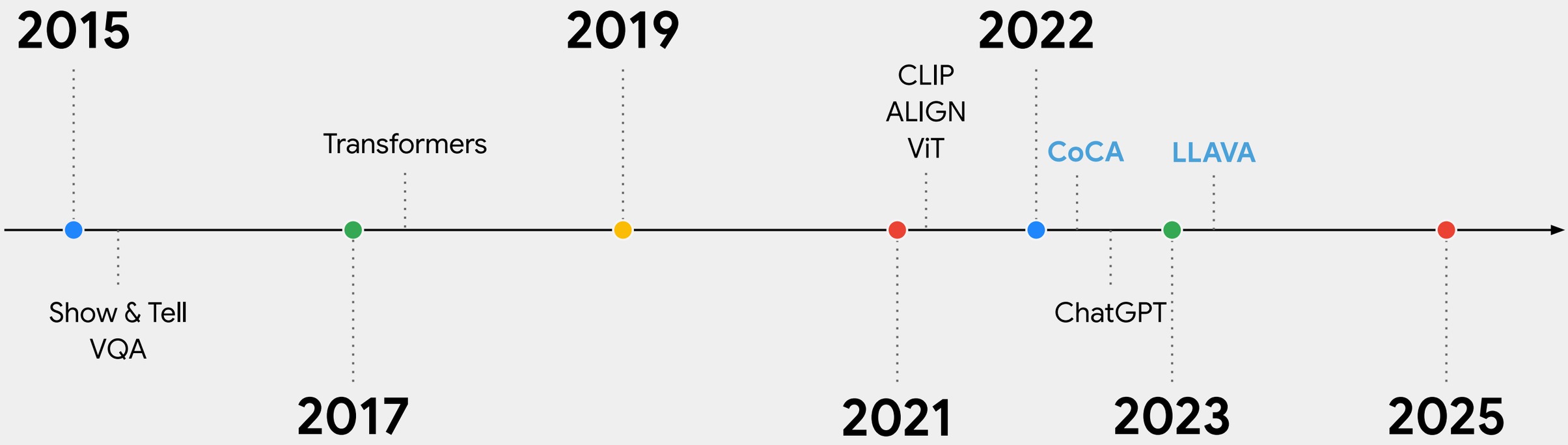
- Visual recognition (e.g., image classification)
- Multimodal understanding (e.g., VQA, captioning)
- Crossmodal Alignment Tasks (e.g., image retrieval)
- Video action recognition



Video encoding using CoCA

Yu et. al, *CoCa: Contrastive Captioners are Image-Text Foundation Models*, TMLR 2024

Vision-Language Models Timeline



LLaVA: Large Language and Vision Assistant

Motivation

- LLMs had become a game-changer for language tasks, due to their world knowledge and ability to follow instructions.
- Multimodal models were still limited.
- Can we make LLM *see*, quickly and affordably?

Key Idea

- Create a resource-efficient method for endowing **pre-existing LLMs** with sight.
- How? Convert images to “language”.
- **Key contribution**: automatic pipeline for creating language-image instruction-following data.
- Train multimodal model to follow human intent to complete visual tasks.

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

LLaVA: Large Language and Vision Assistant

Step 4

Fine-tune on vision-language tasks
(e.g., captioning, VQA).

Step 3

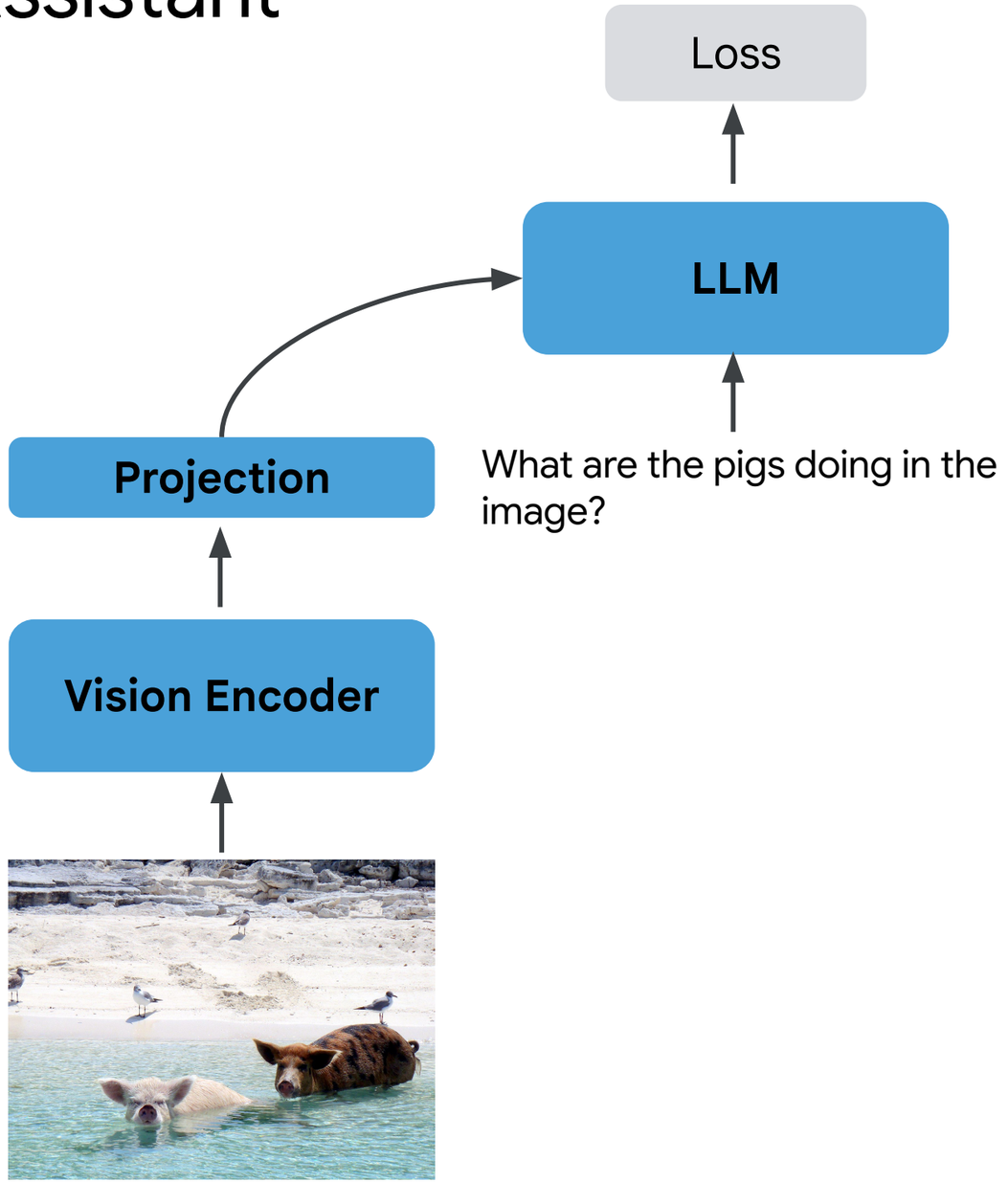
Pass projected image to the LLM.

Step 2

Project image representation into the same space as a pre-trained LLM's text representations.

Step 1

Encode image using a vision encoder.

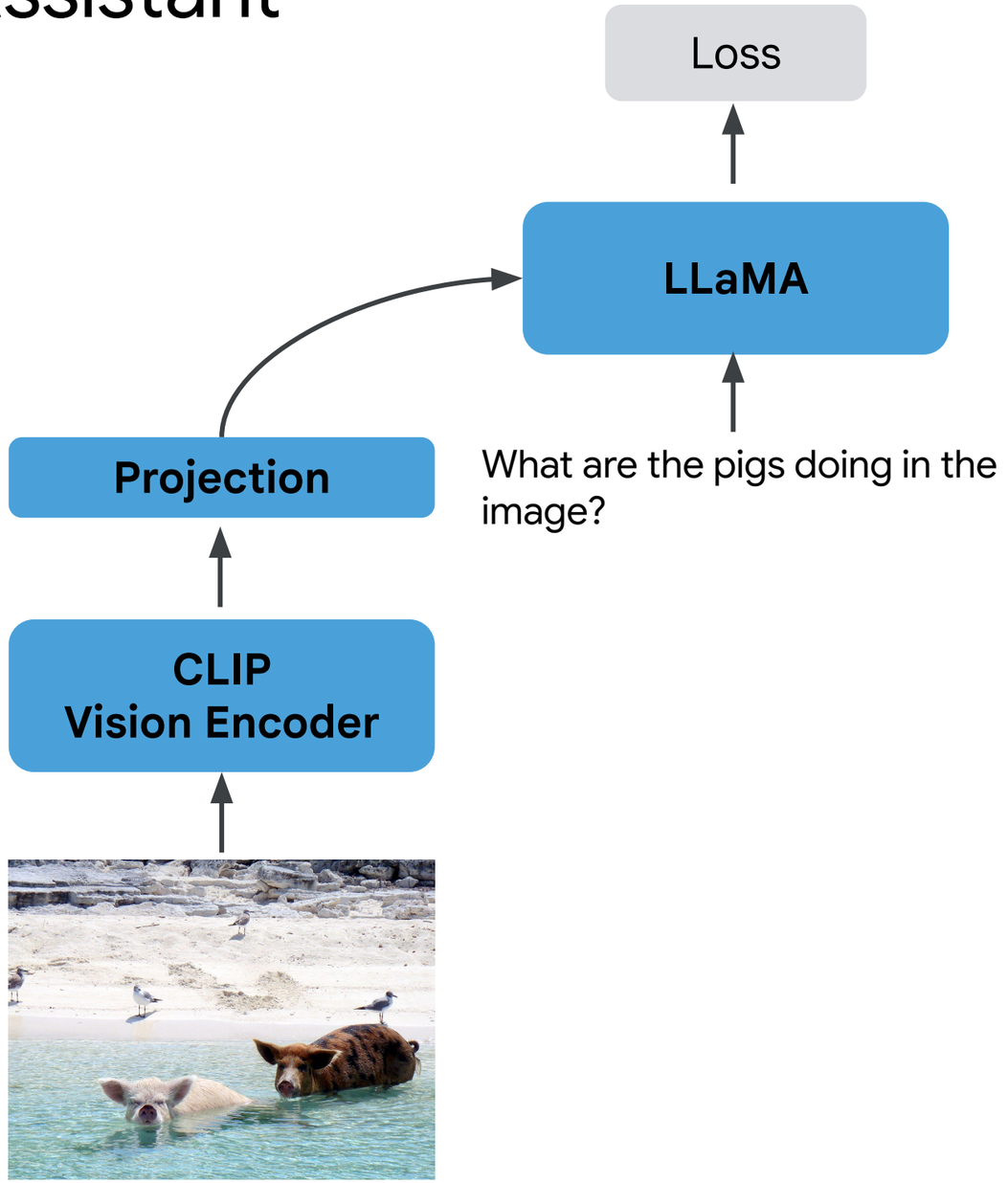


Liu et. al, *Visual Instruction Tuning,* NeurIPS 2023

LLaVA: Large Language and Vision Assistant

Architecture

- Vision Encoder: CLIP ViT-L/14
- Text encoder: LLaMA
- Projection: Linear



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

LLaVA: Training

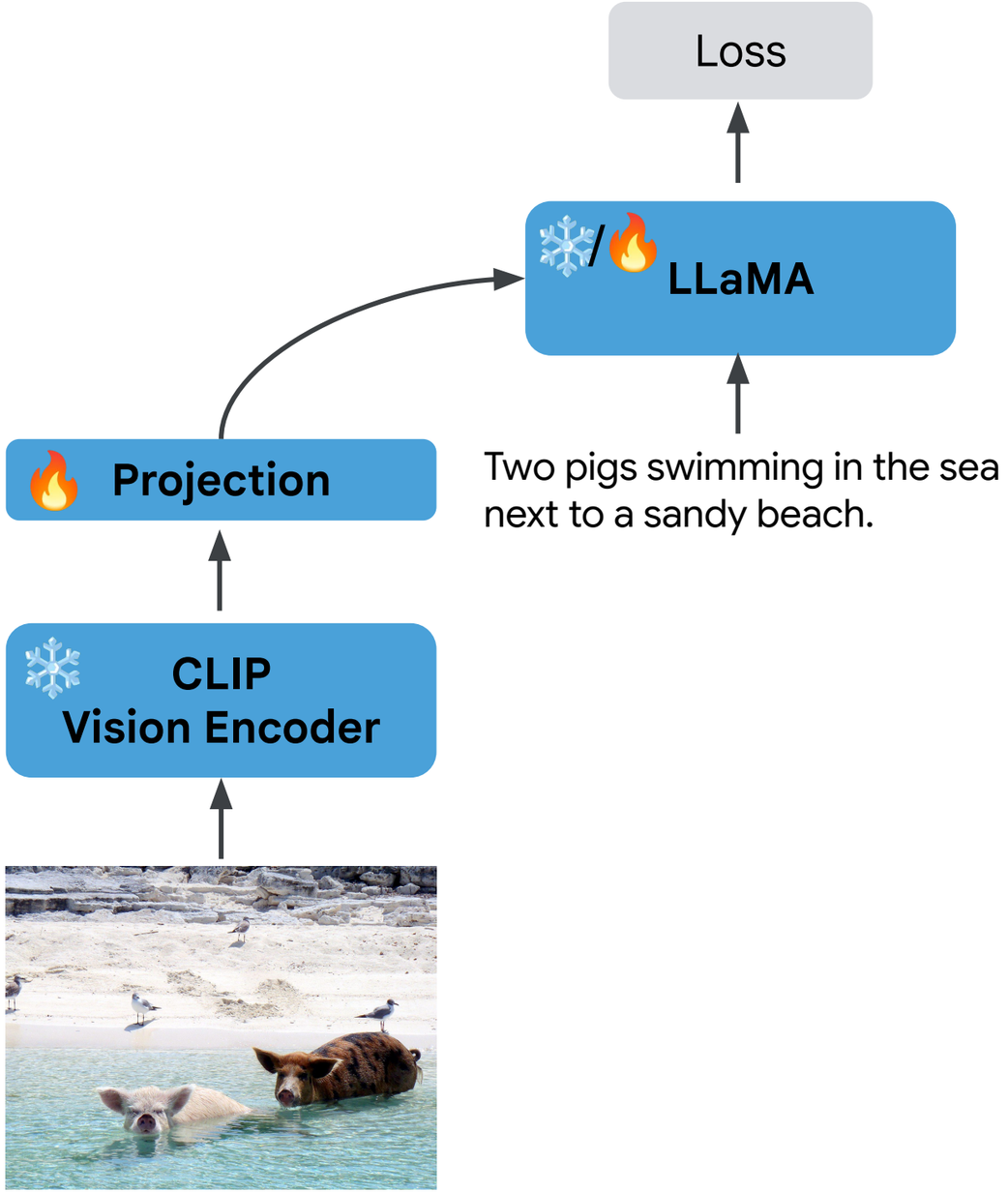
Training Strategy

1) Pre-training for Feature Alignment

Vision Encoder → ❄️ Pretrained
Text encoder → ❄️ Pretrained
Projector → 🔥 Trained from random init on captioning data

2) Fine-tuning End-to-End

Vision Encoder → ❄️ Pretrained
Text encoder → 🔥 Finetune
Projector → 🔥 Finetune on Visual Chat & Science QA data



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

LLaVA: Training

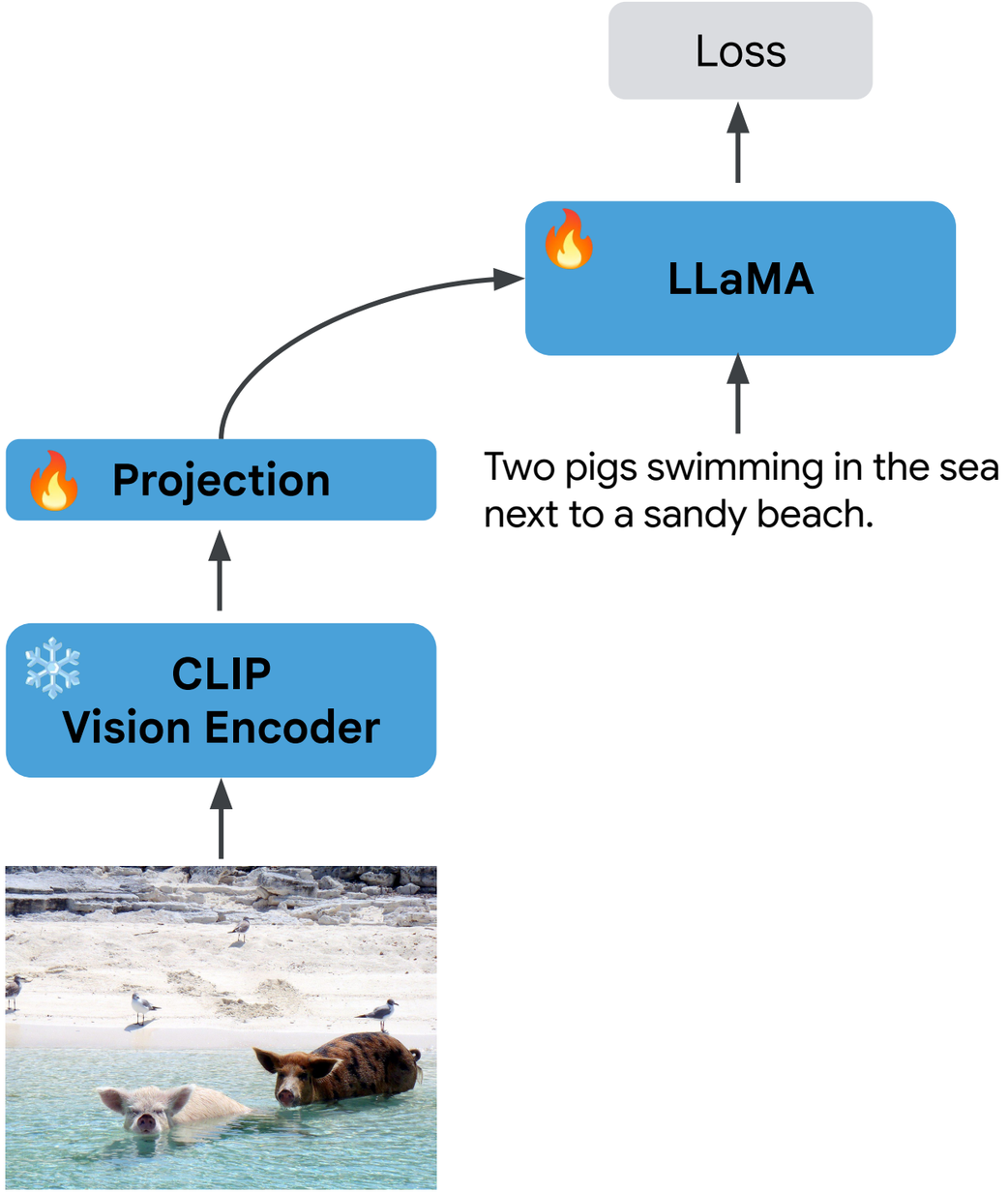
Training Strategy

1) Pre-training for Feature Alignment

Vision Encoder → ❄️ Pretrained
Text encoder → ❄️ Pretrained
Projector → 🔥 Trained from random init on captioning data

2) Fine-tuning End-to-End

Vision Encoder → ❄️ Pretrained
Text encoder → 🔥 Finetune
Projector → 🔥 Finetune on **Visual Chat** & Science QA data



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

LLaVA: Data

LLaVA's main contribution is actually the data!

LLaVA was not the first model with this type of architecture
(e.g., VL-T5 [Cho et al, 2021], GPV-2 [Kamath et al, 2022])

But it is the first attempt to use language-only GPT-4 to generate multimodal language-image instruction-following data.

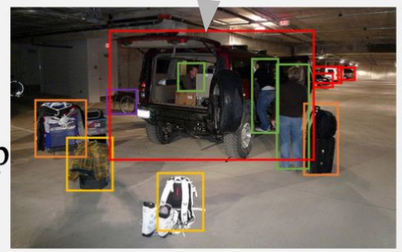
Focus is shifting from model architectures to data generation and training procedures.

LLaVA: Instruction Tuning Synthetic Data

images the COCO dataset

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>

Visual data translated into text, to prompt text-only GPT-4 model

---

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>

GPT-4 responses, used to create synthetic training data.

LLaVA-Instruct-150K dataset

LLaVA: Training

$\mathbf{X}_{\text{system-message}}$ <STOP>
Human : $\mathbf{X}^1_{\text{instruct}}$ <STOP> Assistant: $\mathbf{X}^1_{\text{a}}$ <STOP>
Human : $\mathbf{X}^2_{\text{instruct}}$ <STOP> Assistant: $\mathbf{X}^2_{\text{a}}$ <STOP> ...

input sequence used to train the model, by maximizing the likelihood:

$$p(\mathbf{X}_{\text{a}}|\mathbf{X}_{\text{v}}, \mathbf{X}_{\text{instruct}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\mathbf{X}_{\text{v}}, \mathbf{X}_{\text{instruct},<i}, \mathbf{X}_{\text{a},<i})$$

answer image prompt — current token — previous tokens in instruction — previous tokens in answer

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

LLaVA: Outputs

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 [36]	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

Follows user instructions rather than just describing the scene.

How do we evaluate a general-purpose VLM?

LLAVA Evaluation

Challenging examples from LLaVA-Bench (In-the-Wild):

	 ICHIRAN Ramen [source]	 Filled fridge [source]
Annotation	A close-up photo of a meal at ICHIRAN. The chashu ramen bowl with a spoon is placed in the center. The ramen is seasoned with chili sauce, chopped scallions, and served with two pieces of chashu. Chopsticks are placed to the right of the bowl, still in their paper wrap, not yet opened. The ramen is also served with nori on the left. On top, from left to right, the following sides are served: a bowl of orange spice (possibly garlic sauce), a plate of smoke-flavored stewed pork with chopped scallions, and a cup of matcha green tea.	An open refrigerator filled with a variety of food items. In the left part of the compartment, towards the front, there is a plastic box of strawberries with a small bag of baby carrots on top. Towards the back, there is a stack of sauce containers. In the middle part of the compartment, towards the front, there is a green plastic box, and there is an unidentified plastic bag placed on it. Towards the back, there is a carton of milk. In the right part of the compartment, towards the front, there is a box of blueberries with three yogurts stacked on top. The large bottle of yogurt is Fage non-fat yogurt, and one of the smaller cups is Fage blueberry yogurt. The brand and flavor of the other smaller cup are unknown. Towards the back, there is a container with an unknown content.
Question 1	What's the name of the restaurant?	What is the brand of the blueberry-flavored yogurt?
Question 2	Describe this photo in detail.	Is there strawberry-flavored yogurt in the fridge?

Diverse but very small (24 images, 60 questions)

LLaVA-Bench (In-the-Wild)

VLM Evaluation Today

Nowadays we have more general-purpose benchmarks and arenas, capturing how users really use models (e.g., “arenas”)

- MMMU (Massive Multi-discipline Multimodal Understanding): tests knowledge, perception; multiple-choice format
- VQA, GQA, OK-VQA, TextVQA, ScienceQA: tests visual reasoning; question answering format
- MathVista: tests visual mathematical reasoning
- Vision Arena (extension of Chatbot Arena): human ranks two VLM responses; ELO rating system

...

LLaVA: Summary

- Among the first to use a vision encoder along with an LLM
- Generated synthetic training data using LLMs
- Used multimodal instruction tuning to created **multimodal chatbot**
- Fine-tuned on a wide variety of downstream tasks

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

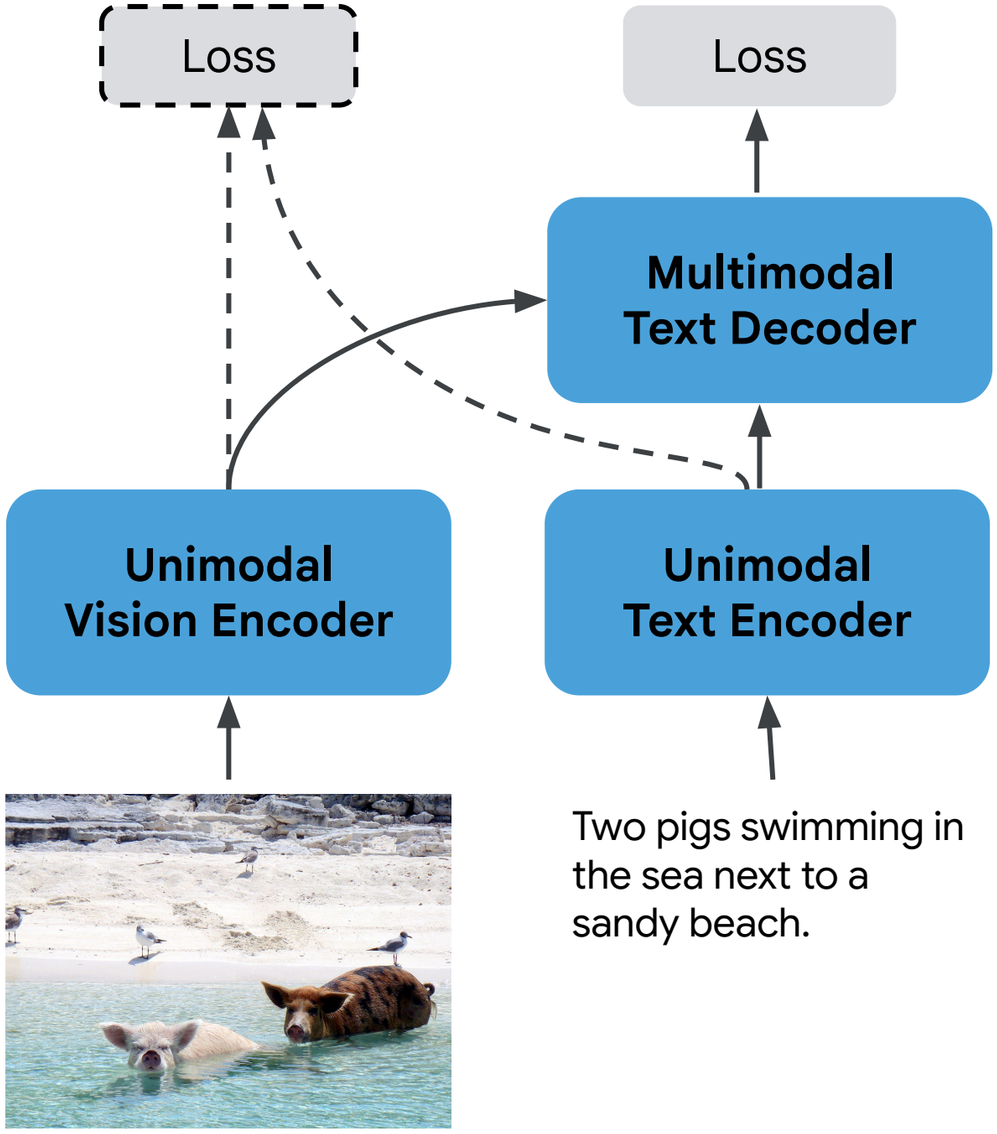
Cross-Modal Models

Advantages

- Better performance on complex vision-language tasks.

Weaknesses

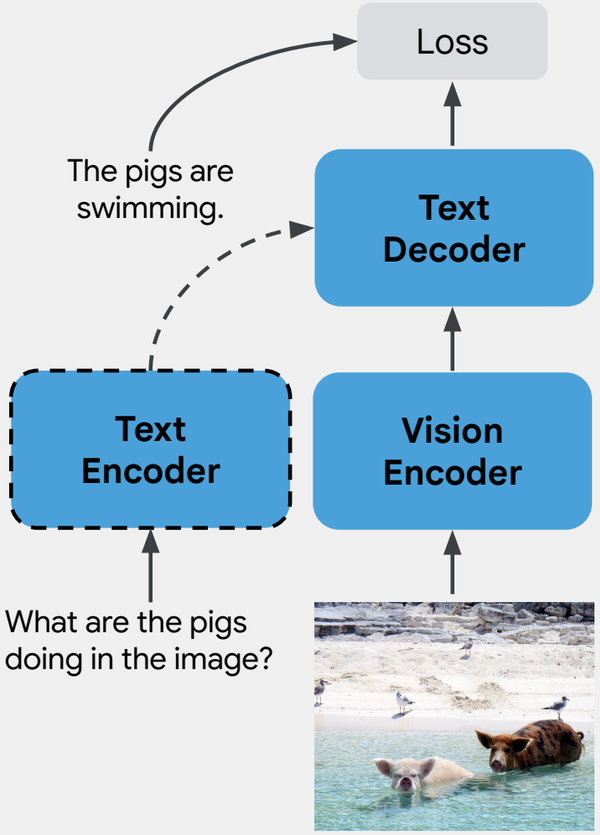
- Higher computational cost. Requires a full model pass for every input pair.
- High latency. Unsuitable for large-scale retrieval.
- No cross-attention during vision & text encoding.



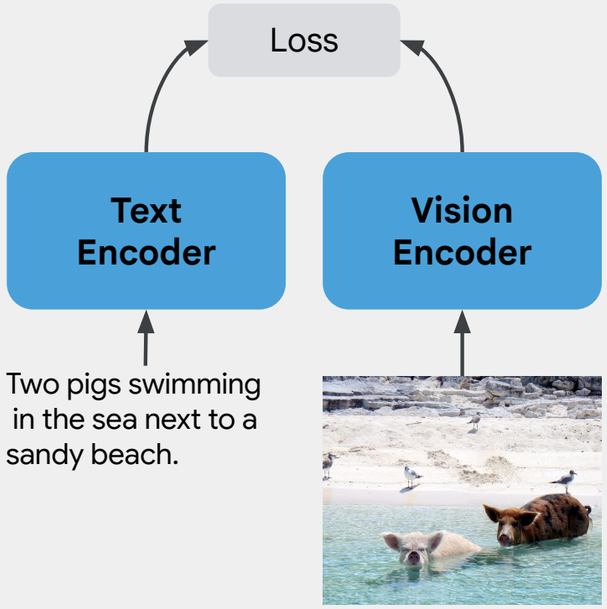
Cross-Modal Model

Vision-Language Model Architectures

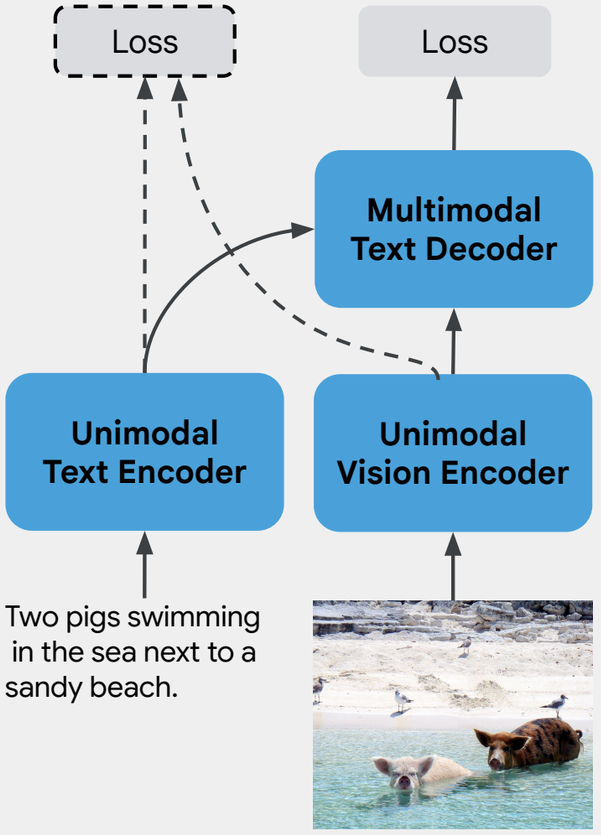
Encoder-Decoder



Dual-Encoder



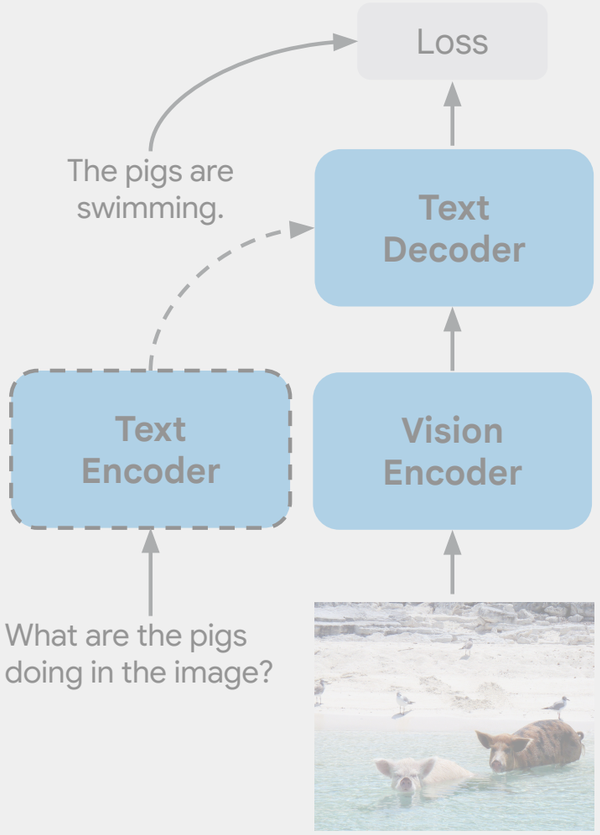
Cross-Modal



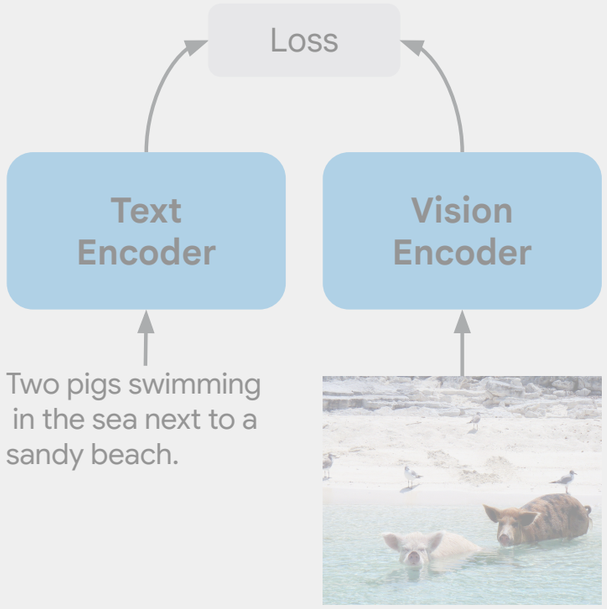
Natively Multimodal

Vision-Language Model Architectures

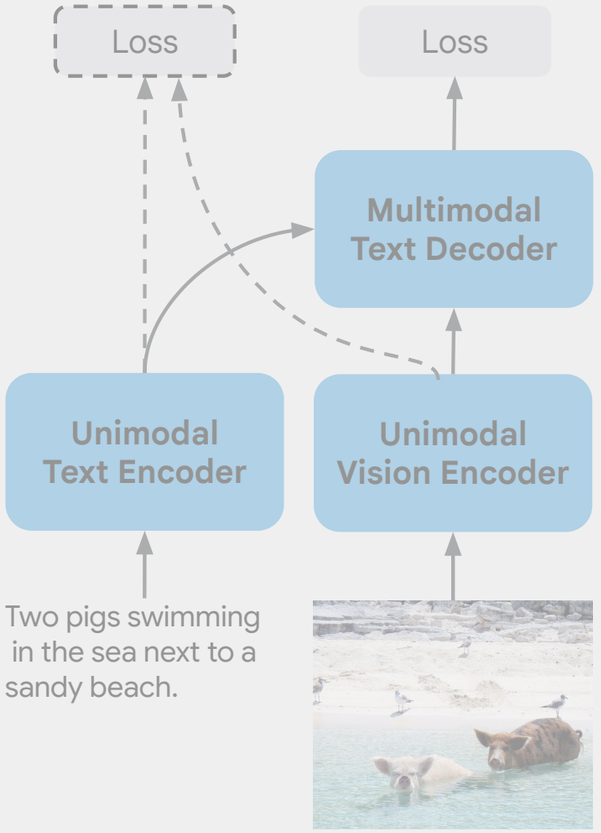
Encoder-Decoder



Dual-Encoder



Cross-Modal



Natively Multimodal

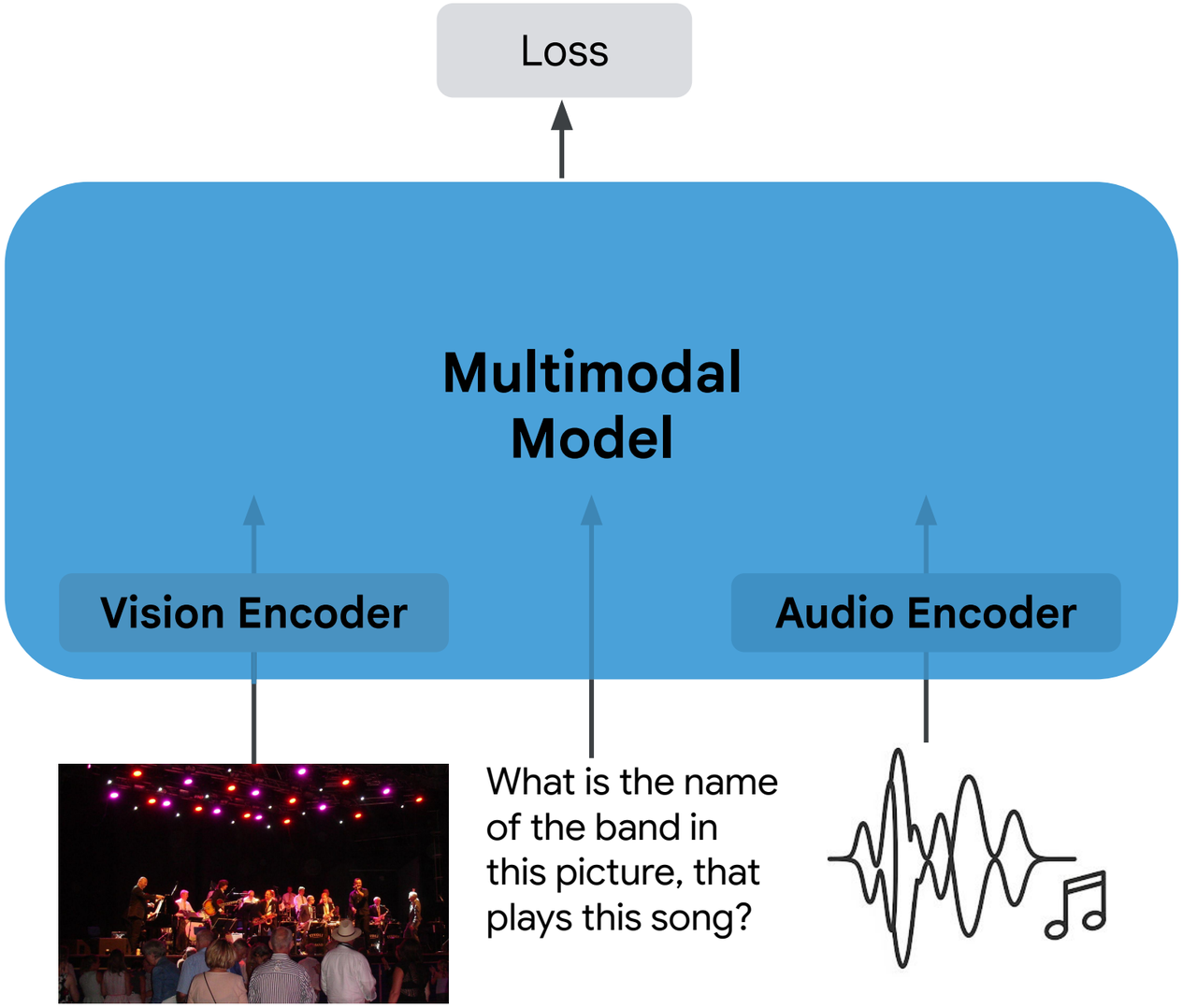
Natively Multimodal Models

Key Ideas

- Support interleaving data of different modalities.
- Key design choice: **early fusion.**
- All input modalities are converted into a common representational format, typically a sequence of tokens or embeddings.

Example Models:

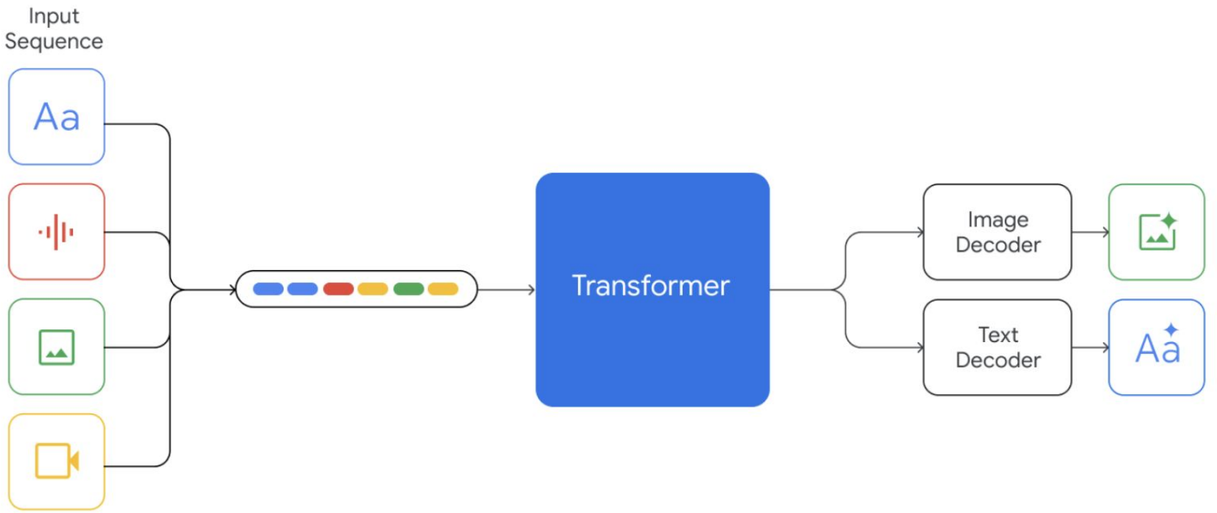
- Gemini
- GPT-4o
- LLAMA 4



Gemini

Architecture

- Interleaved sequences of text, image, audio, and video as inputs.
- Built on top of Transformer **decoders**.
- Can also natively output images using discrete image tokens.



Training

- Large scale pretraining on multimodal, multilingual dataset.
- Supervised fine-tuning (SFT) on demonstration data.
- Reinforcement Learning from Human Feedback (RLHF) [Bai et al., 2022] to align the model's outputs with human preferences.

Gemini Team @ Google, *Gemini: A Family of Highly Capable Multimodal Models*, 2023

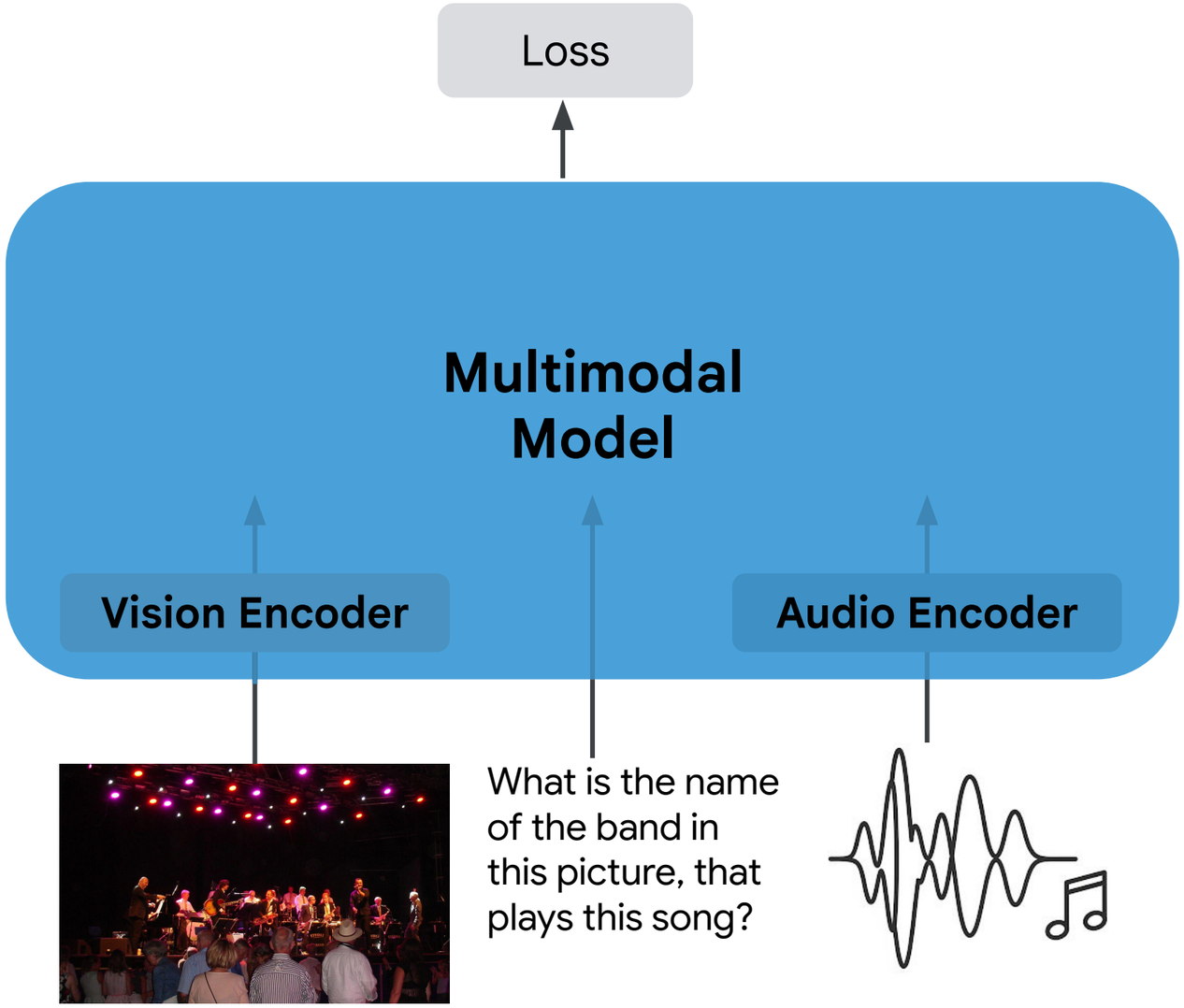
Natively Multimodal Models

Advantages

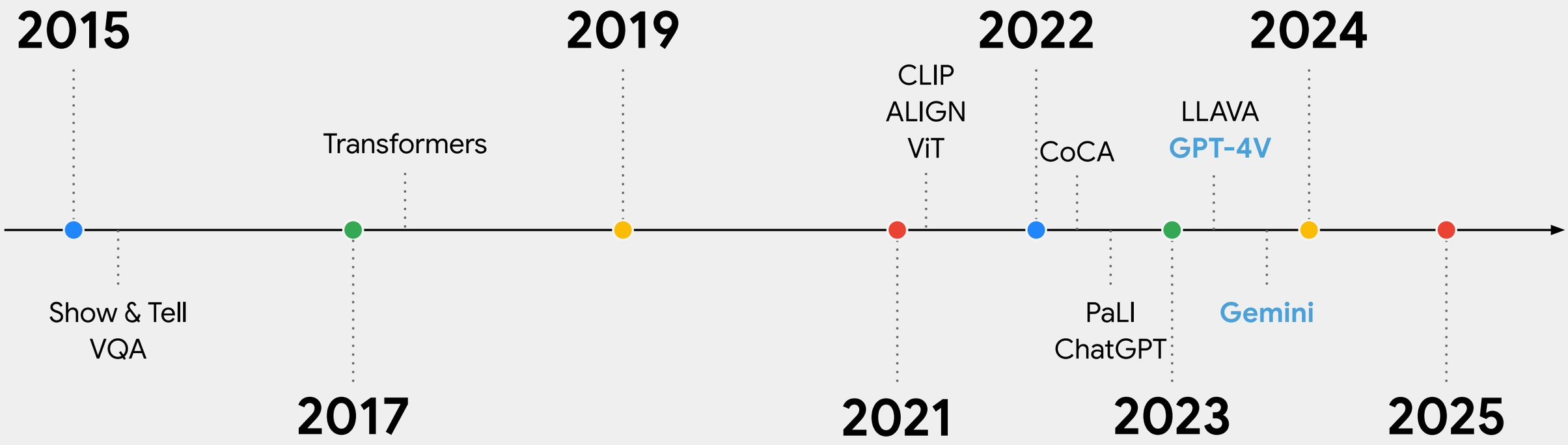
- More nuanced understanding
- Enhanced reasoning capabilities
- Emergent abilities that the model was not trained for

Weaknesses

- Very expensive to train
- Require large amounts of multimodal data
- Harder to understand how models arrive at their conclusions

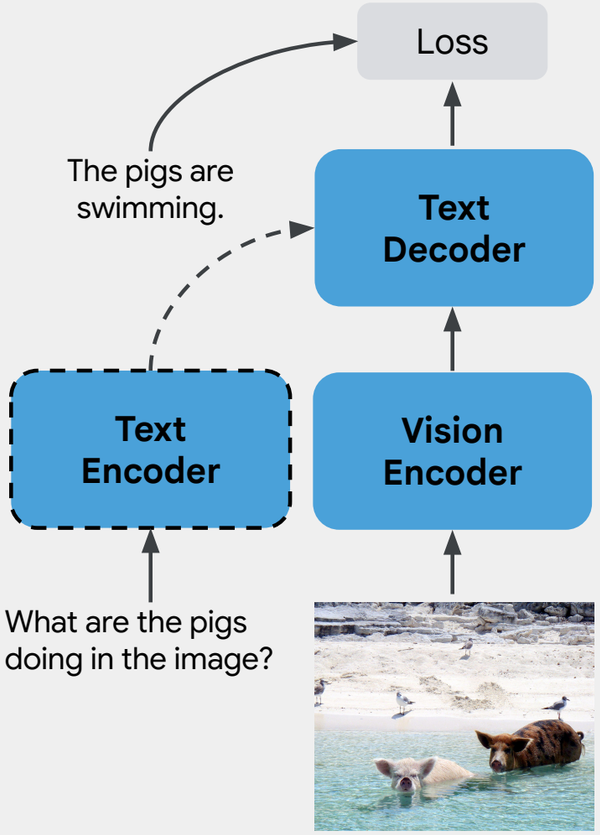


Vision-Language Models Timeline

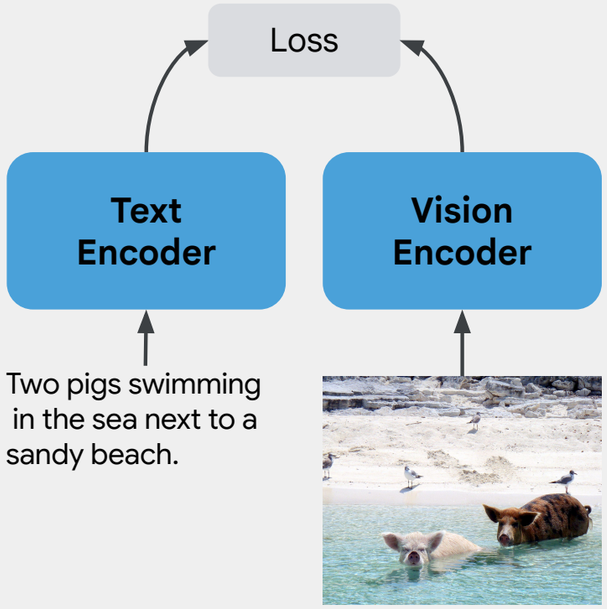


Vision-Language Model Architectures

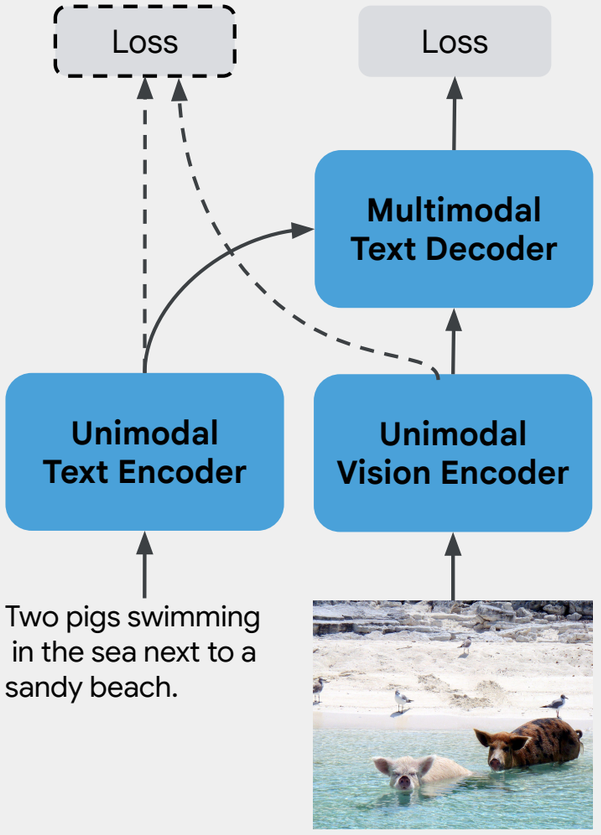
Encoder-Decoder



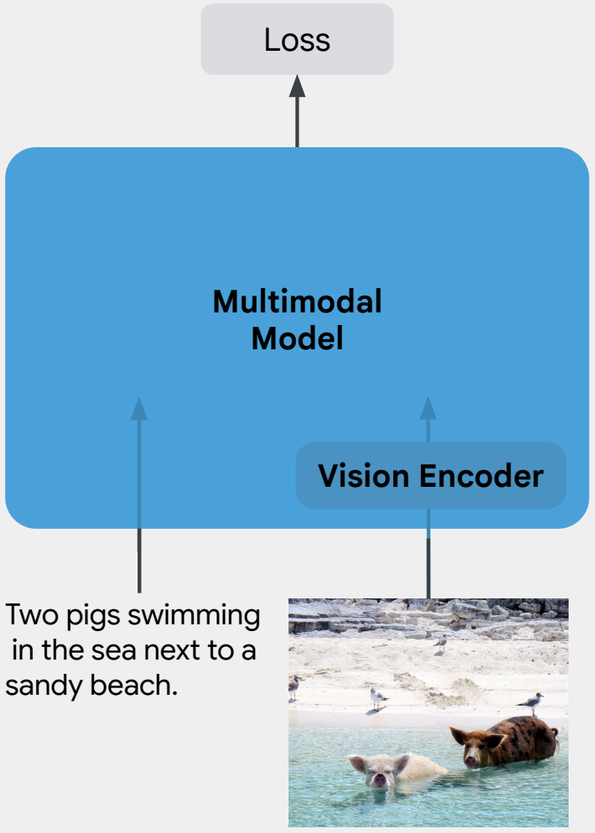
Dual-Encoder



Cross-Modal



Natively Multimodal



At A Glance:

Other Multimodal Topics

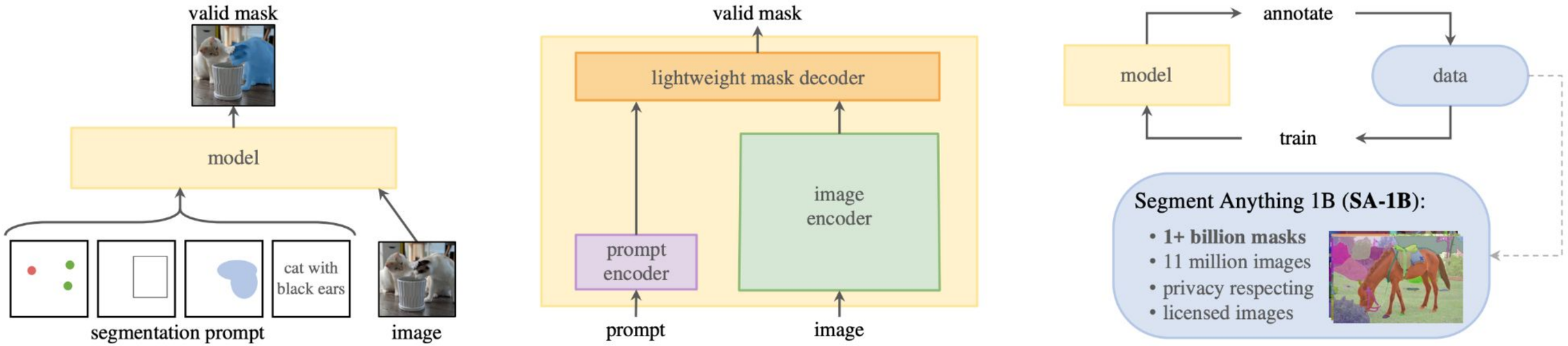
Open Vocabulary Segmentation: Segment Anything (SAM)

Segmentation foundation model outputs masks of any objects prompted by the user.

Key contribution: data annotation engine → efficient method for collecting new data using weaker version of the model to retrain a better version.

Kirillov et al., *Segment Anything*, ICCV 2023

Open Vocabulary Segmentation: Segment Anything (SAM)



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Figure 1: We aim to build a foundation model for segmentation by introducing three interconnected components: a prompt-able segmentation *task*, a segmentation *model* (SAM) that powers data annotation and enables zero-shot transfer to a range of tasks via prompt engineering, and a *data* engine for collecting SA-1B, our dataset of over 1 billion masks.

Kirillov et al., *Segment Anything*, ICCV 2023

Open Vocabulary Segmentation: Segment Anything (SAM)

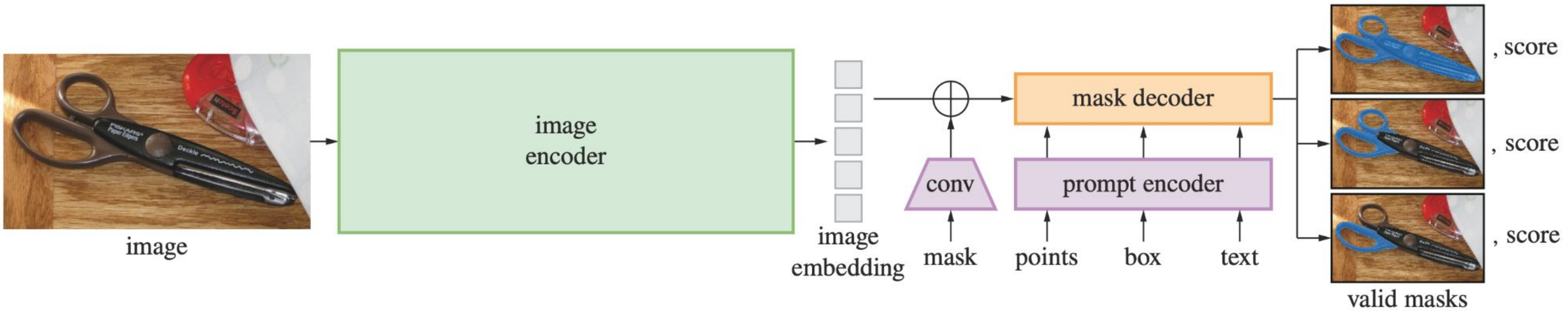


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

Kirillov et al., *Segment Anything*, ICCV 2023

Text-to-Image Generation: Diffusion Models

Can generate high-quality images given a prompt.

Key components:
Transformer-based LLM to understand text, followed by a cascade of diffusion models for image generation.

Prompt: Create a cinematic, photorealistic medium shot capturing the nostalgic warmth of a late 90s indie film. The focus is a young woman with brightly dyed pink hair (slightly faded) and freckled skin, looking directly and intently into the camera lens with a hopeful yet slightly uncertain smile [...]

Imagen 4 from Google, 2025

Video-Language Models: Veo 3

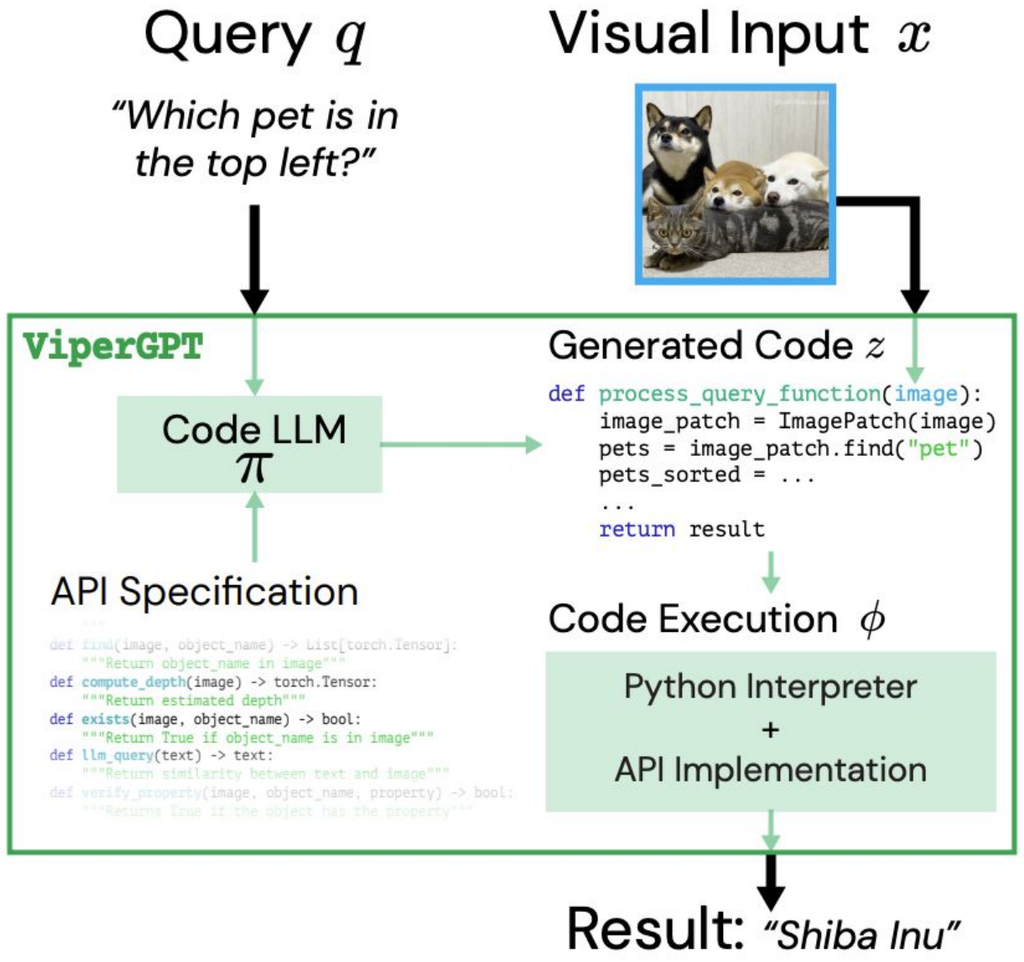
Prompt: A medium shot frames an old sailor, his knitted blue sailor hat casting a shadow over his eyes, a thick grey beard obscuring his chin. He holds his pipe in one hand, gesturing with it towards the churning, grey sea beyond the ship's railing. "This ocean, it's a force, a wild, untamed might. And she commands your awe, with every breaking light"

[Google Veo 3 demo; 2025]

VLMs with Tools: Visual Programming

Step 1: Given a query,
have an LLM generate Python code that
solves the task with vision specialists!

Step 2: Execute the code.

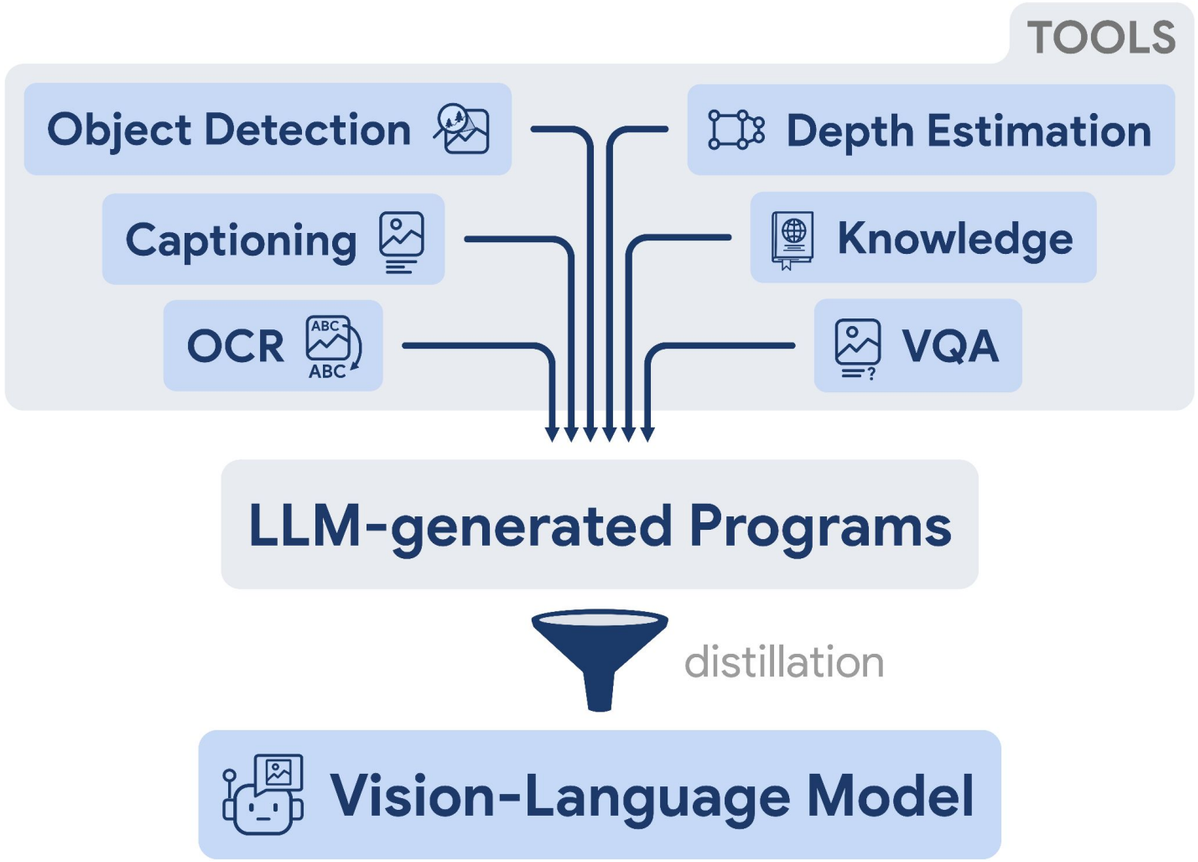


Gupta et al., *Visual Programming: Compositional visual reasoning without training*. CVPR 2023. (Best Paper)
Sur'is et.al., *ViperGPT: Visual Inference via Python Execution for Reasoning*. ICCV 2023.

VLMs with Tools: Visual Program Distillation

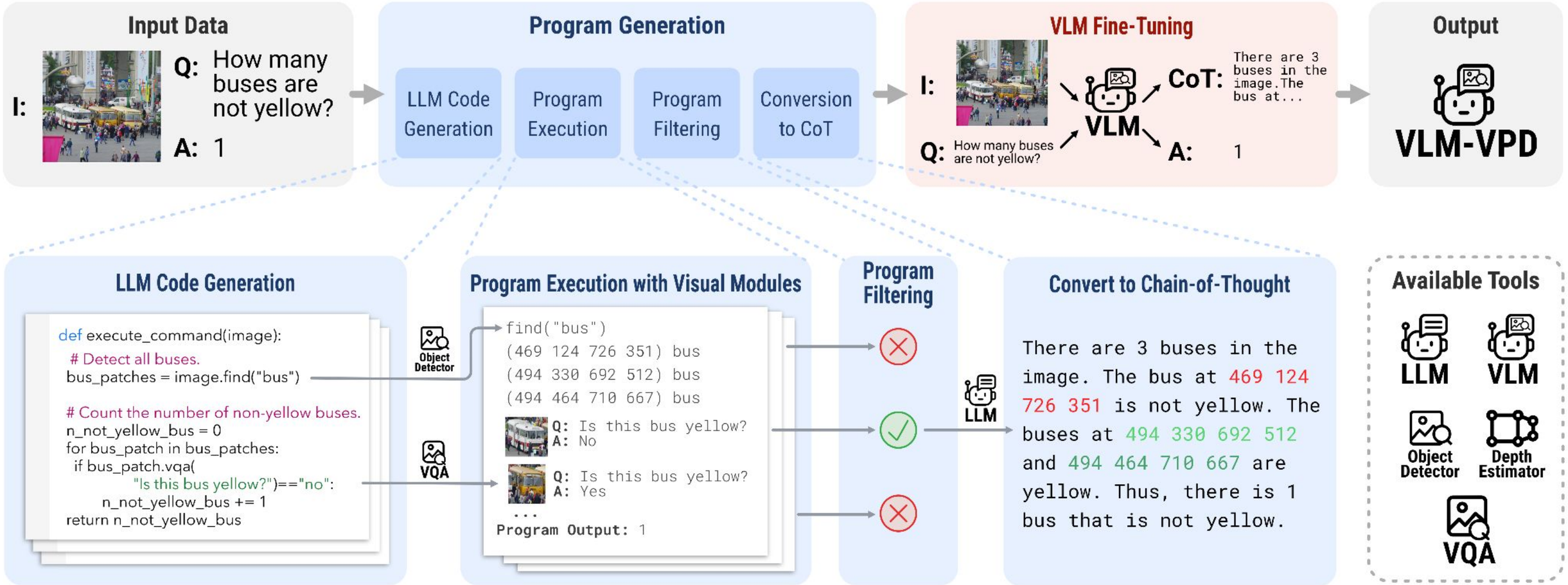
Step 1: Generate training data using programs.

Step 2: Fine-tune VLMs on the generated data.



Hu, **Stretcu** et al., *Visual Program Distillation: Distilling Tools and Programmatic Reasoning into Vision-Language Models, CVPR 2024*

VLMs with Tools: Visual Program Distillation



Hu, **Stretcu** et al., *Visual Program Distillation: Distilling Tools and Programmatic Reasoning into Vision-Language Models, CVPR 2024*

Summary & Conclusions

VLM Key Components

Model Architecture

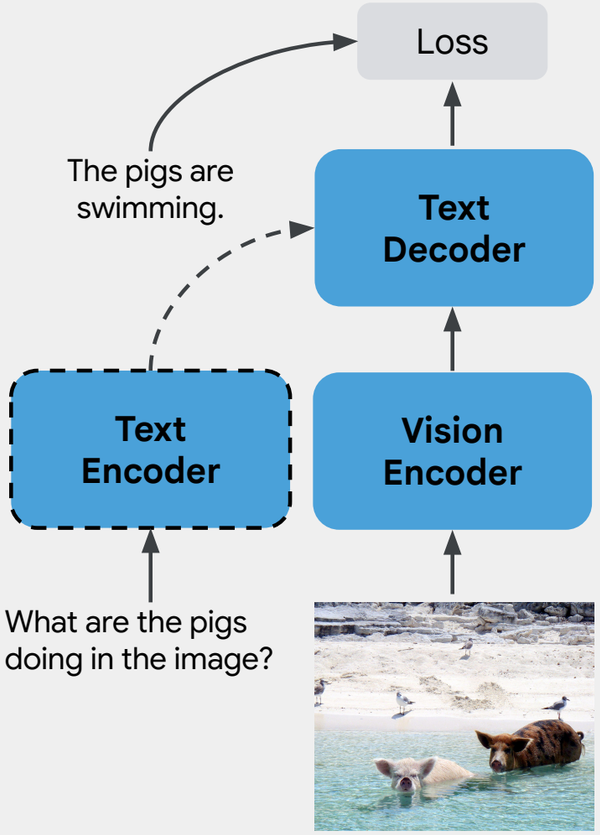
- How is each modality is encoded?
- How do we fuse the different modalities?

Training Strategy

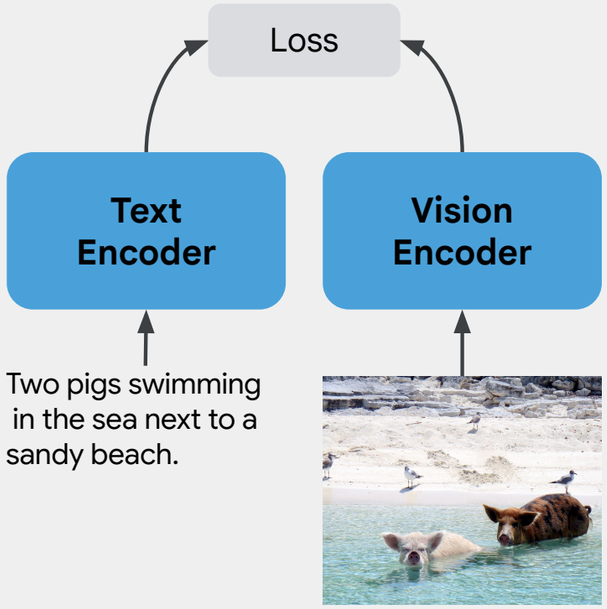
- Loss function
- Initialization (pretrained weights or from scratch?)
- Training stages (e.g., pretraining, fine-tuning)
- Data
  - Types of tasks (e.g., VQA, captioning)
  - Supervision (where do the target outputs come from?)

Vision-Language Model Architectures

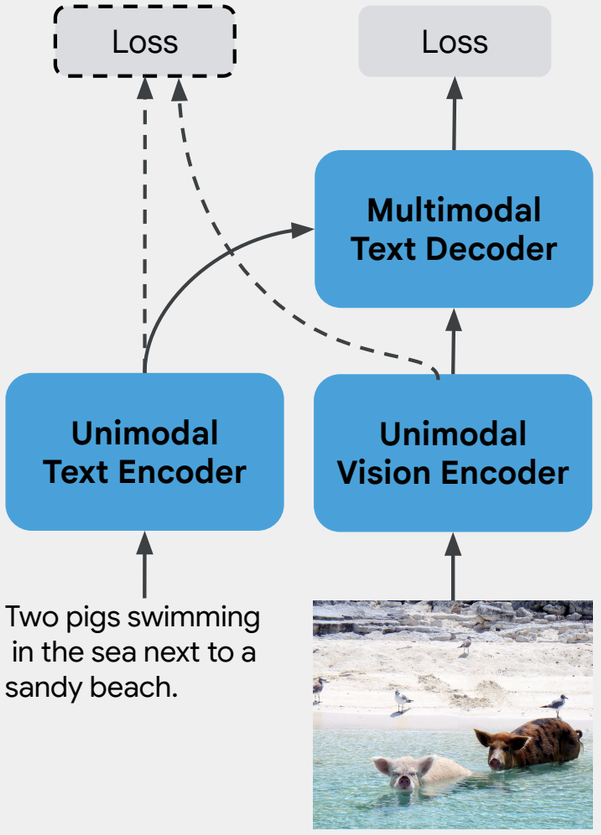
Encoder-Decoder



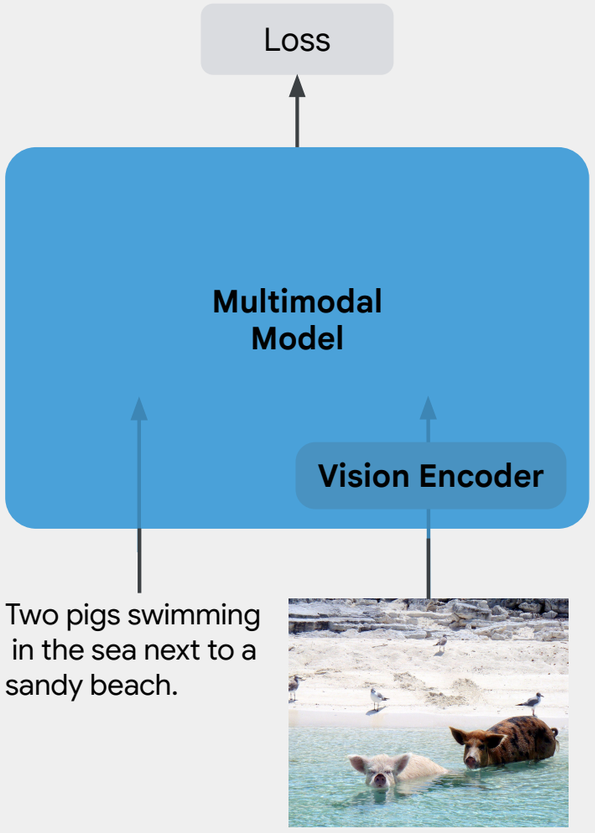
Dual-Encoder



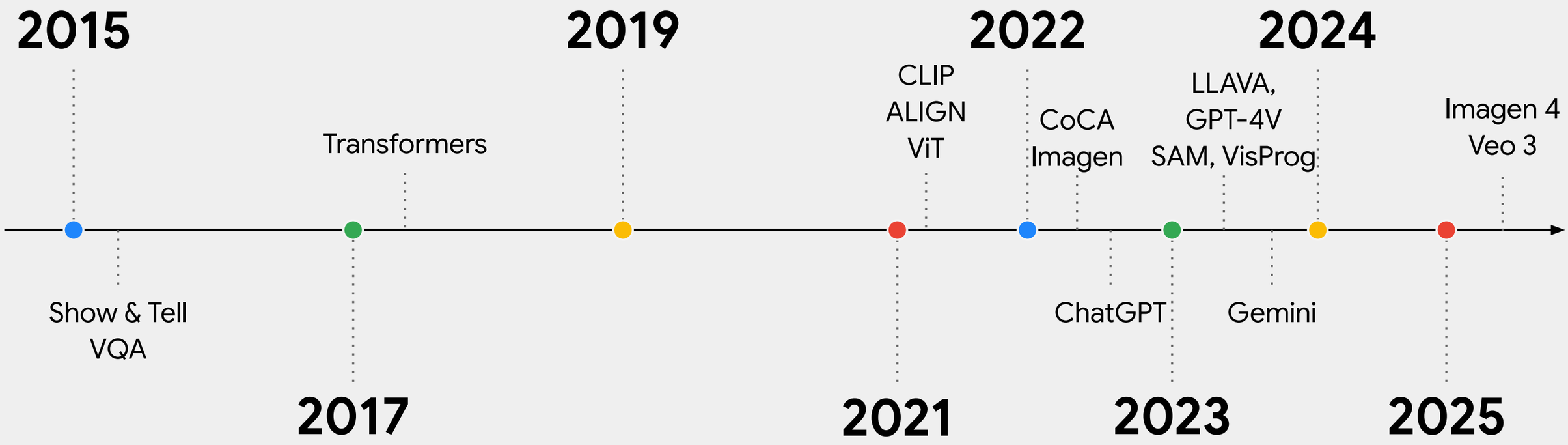
Cross-Modal



Natively Multimodal



Vision-Language Models Timeline



Open Problems

- What is missing from the training data? How do you fill the gaps?
- What can scale solve? What can't it solve?
- Is there a better way to encode images than a sequence of patches?
- Interpretability
- Making models more efficient, e.g., via distillation
- What are our evaluations missing? Are we setting the wrong research targets?
- When to use tools?

Slide credit: Ranjay Krishna, Amita Kamath, University of Washington, Multimodal Foundation Models Lecture.

Thank You!

Questions?

Otilia Stretcu • Google Research • otiliastr@google.com