Multimodal NLP

Vision Language Models and Text-to-Image Models

Aaron Mueller

CAS CS 505: Introduction to Natural Language Processing

Spring 2026

Boston University

Admin

- Your **midway reports** are due tonight at 11:59pm!
  - Remember: you can use your free late days if you still have any remaining
  - The rubric is available on Gradescope

- Course evaluation season is here!
  - If we can get a >75% response rate by Apr. 29, I'll add 1 point of extra credit to everyone's grades, and I'll bring coffee/tea and donuts the last day of class (Apr. 30)

https://go.blueja.io/RQKUN1S8XUS-oEax3H05AQ

What is a modality?

- A “modality” is a type of data
  - Text
  - Images
  - Audio

- Some use “modality” to refer to different modes of writing within text
  - Math
  - Code
  - Language

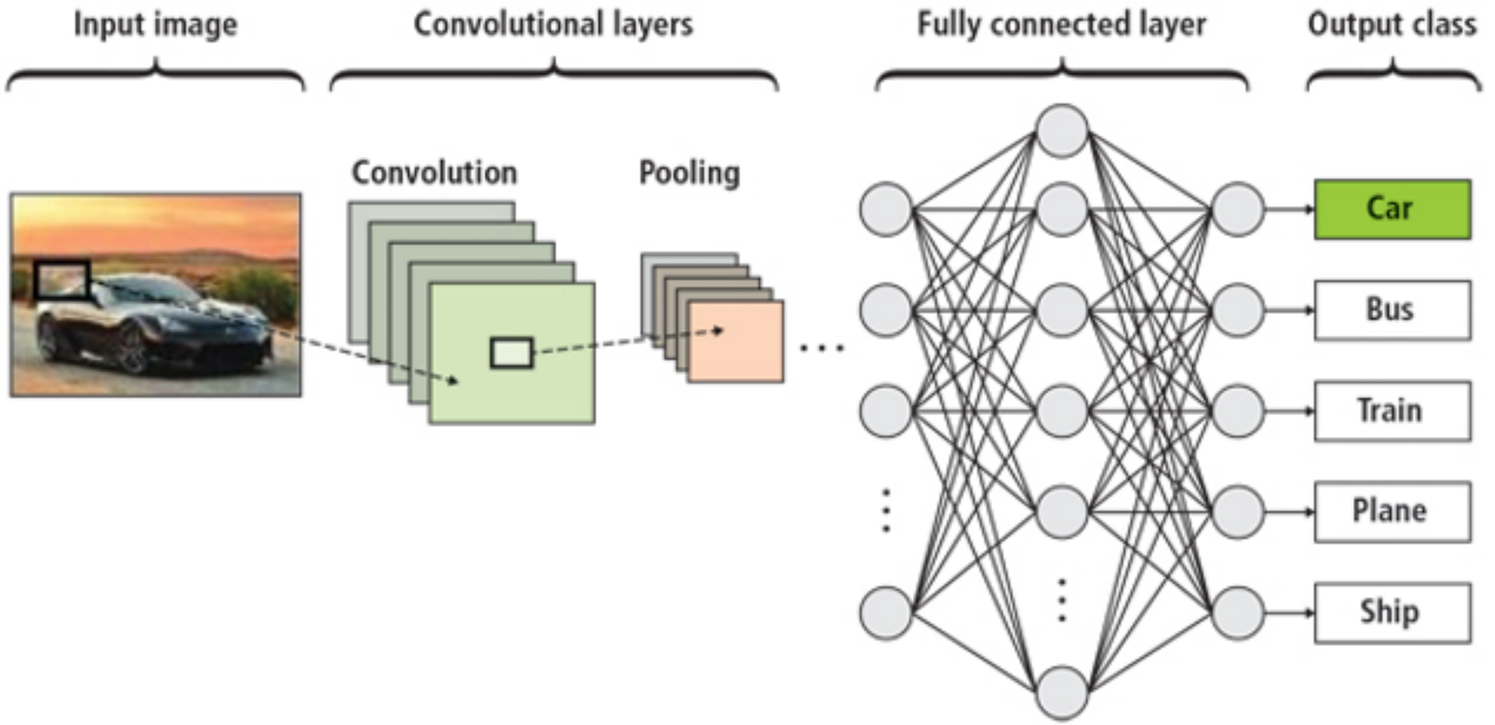
Outline

1. A quick intro to vision models

2. Vision-and-text models
    1. Vision language models: Vision Transformers (ViT), LlaVa
    2. Image-and-text pre-training with CLIP
    3. Image generation models

3. Simulating the human language learning environment

Vision Models



Goal: to produce some meaningful representation of images for use in downstream tasks.

Convolutional neural nets (CNNs) were the canonical architecture for a long time.

These days, we have more Transformer-based and diffusion-based models.

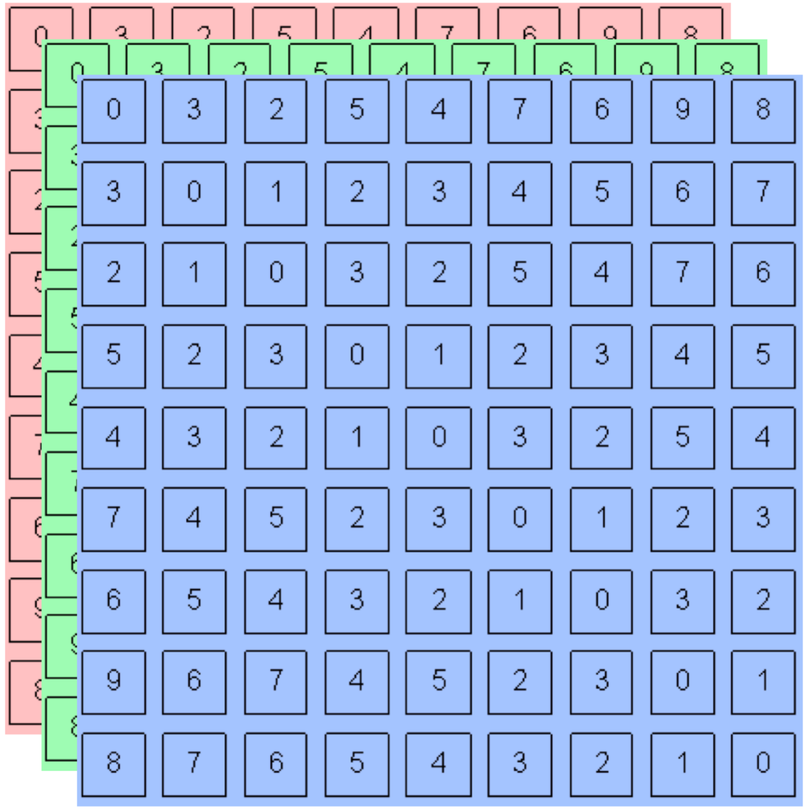
How do we represent images?

0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

An image can be thought of as a 2D tensor of pixel values.

Greyscale pixels only have one value dimension: light/dark.

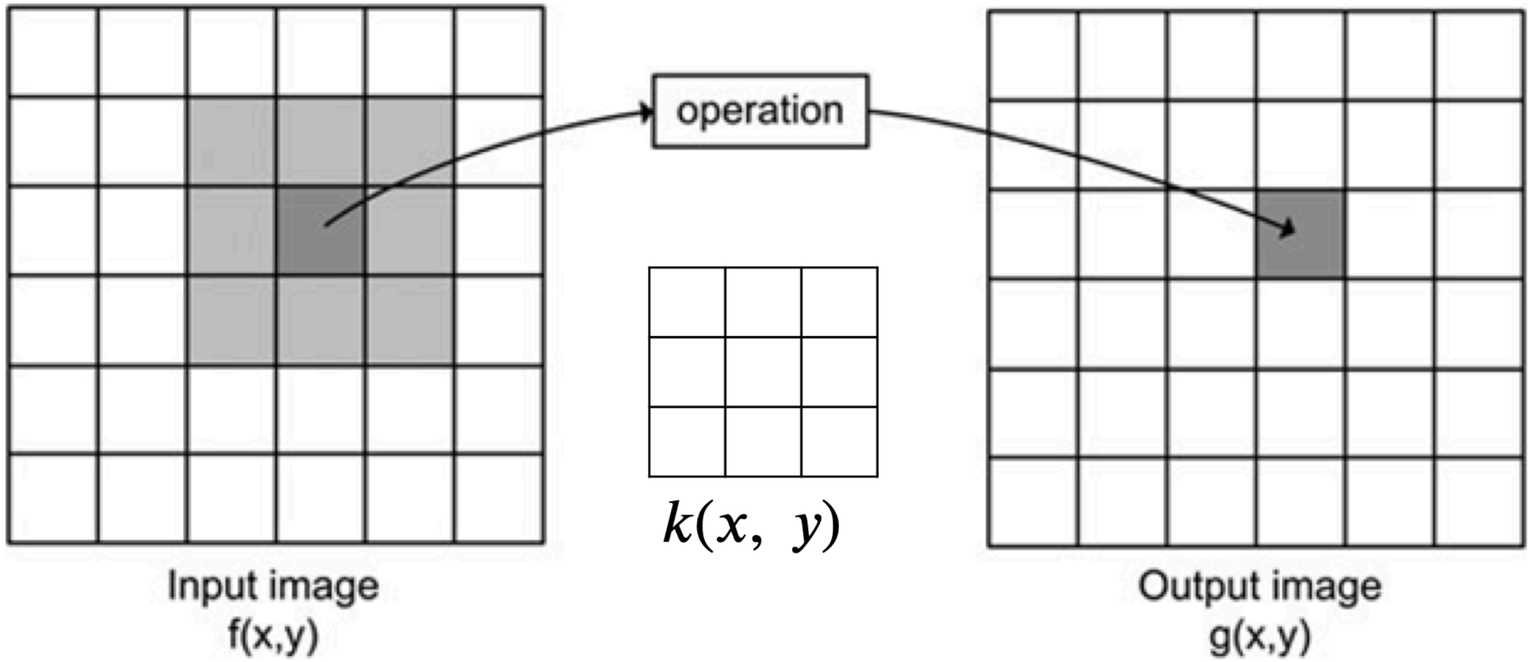
How do we represent images?



Color pixels have 3 value dimensions per pixel: **red**, **green**, **blue**.

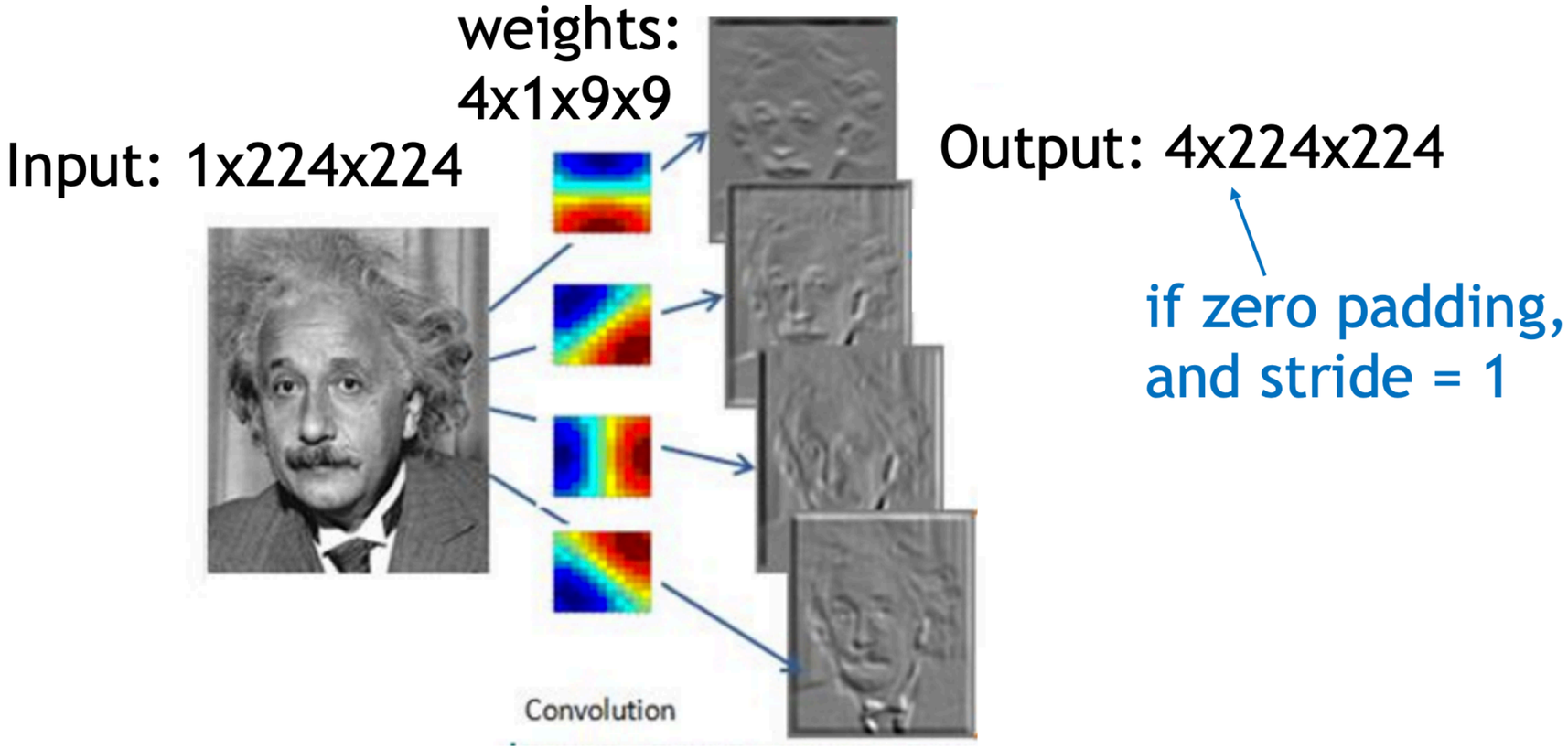
Thus, color images are 3D tensors: height x width x channels

Convolutions



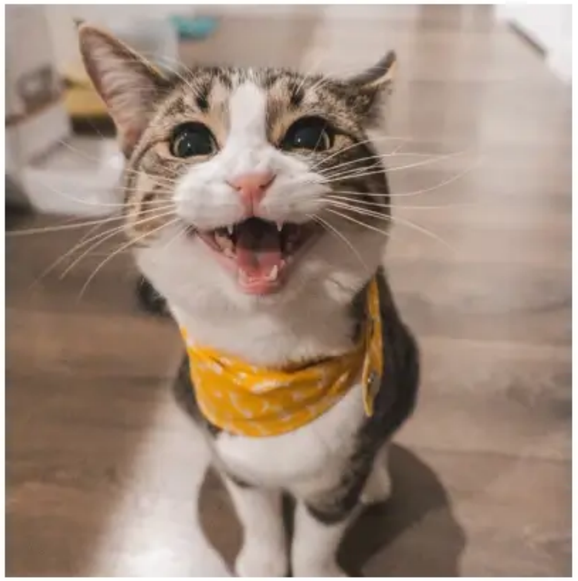
A **convolution** involves a sliding multiplication and addition of regional patches by a small array of numbers known as a **kernel/filter**.

Convolutions



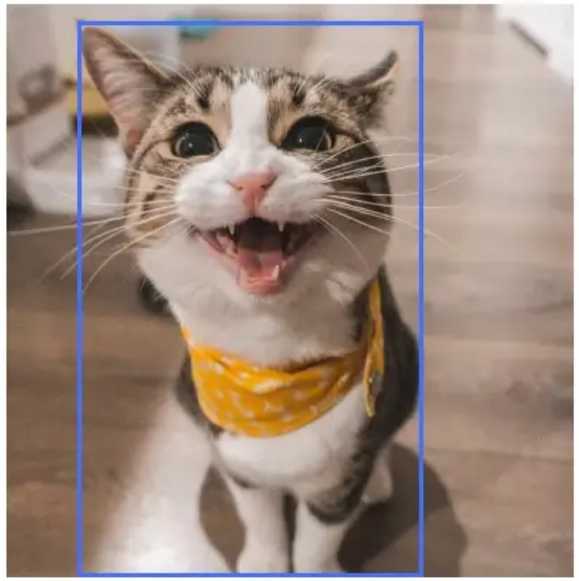
Each filter learns to attend to particular features of an image, like edges, corners, shapes, etc.

Vision Tasks



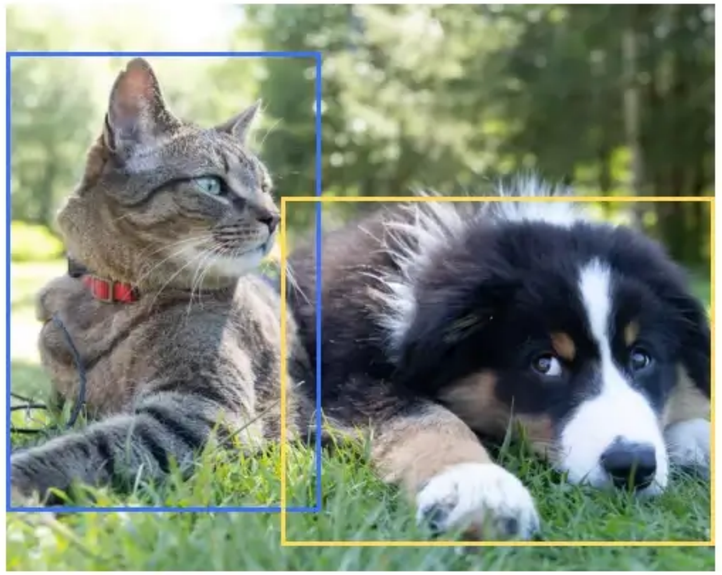
Classification

Cat



Classification, Localization

Cat



Object Detection

Cat, Dog

Tasks like classification, segmentation, and detection can be done with no language representations.

AlexNet

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

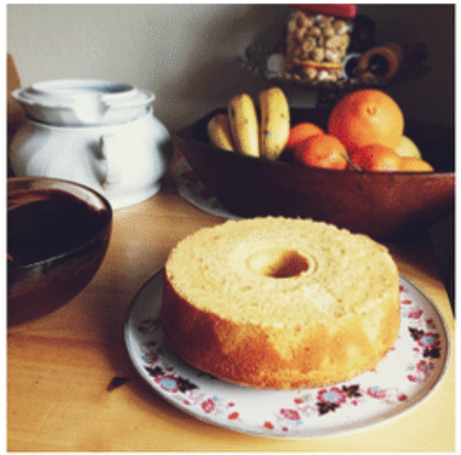
Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Before 2012, deep learning was not mainstream.

In 2012, **AlexNet** (an 8-layer convolutional neural net) was released. It achieved state-of-the-art performance by a pretty significant margin on image classification.

Vision and Language Tasks

Visual Question Answering



Question: *What's in the bowl behind the cake?*

Template: *The **[mask]** is in the bowl behind the cake.*

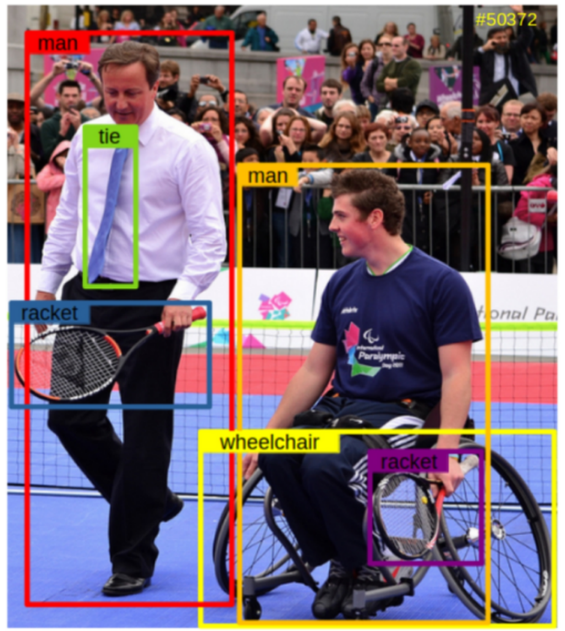
*bread, **fruit**, soup, spoon, donut, dessert...*

Prompts: *The **bread** is in the bowl behind the cake.*
*The **fruit** is in the bowl behind the cake.*
……

Answer: ***fruit***

A big challenge: **grounding**.

How do we ground language to content in other modalities?



(a) image captioning

Output:
Men playing tennis on a tennis court

(b) dense captioning

Output:
Men playing a game of tennis
A man riding a wheelchair on a tennis court
A man standing on a tennis court
Men playing tennis on a tennis court
Men playing tennis on a court
People sitting on a bench
A man standing on a tennis court holding racket

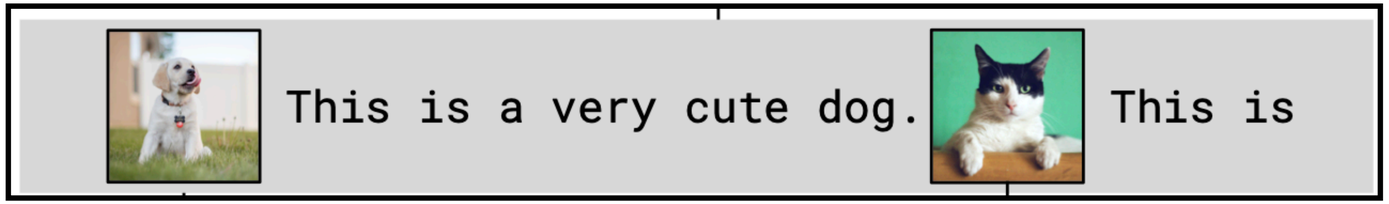
(c) RefCap

Prompt:
right man

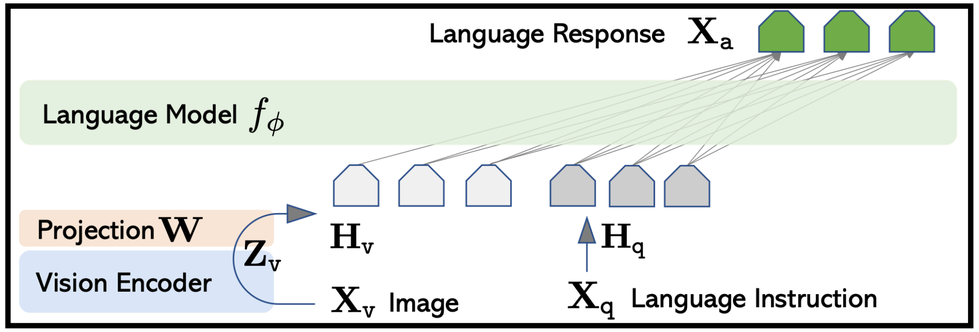
Output:
A man riding a wheelchair with holding a racket

Representing Images and Text

- Two main approaches:
  - Create embeddings for images, and feed them to a model as part of a sequence of tokens

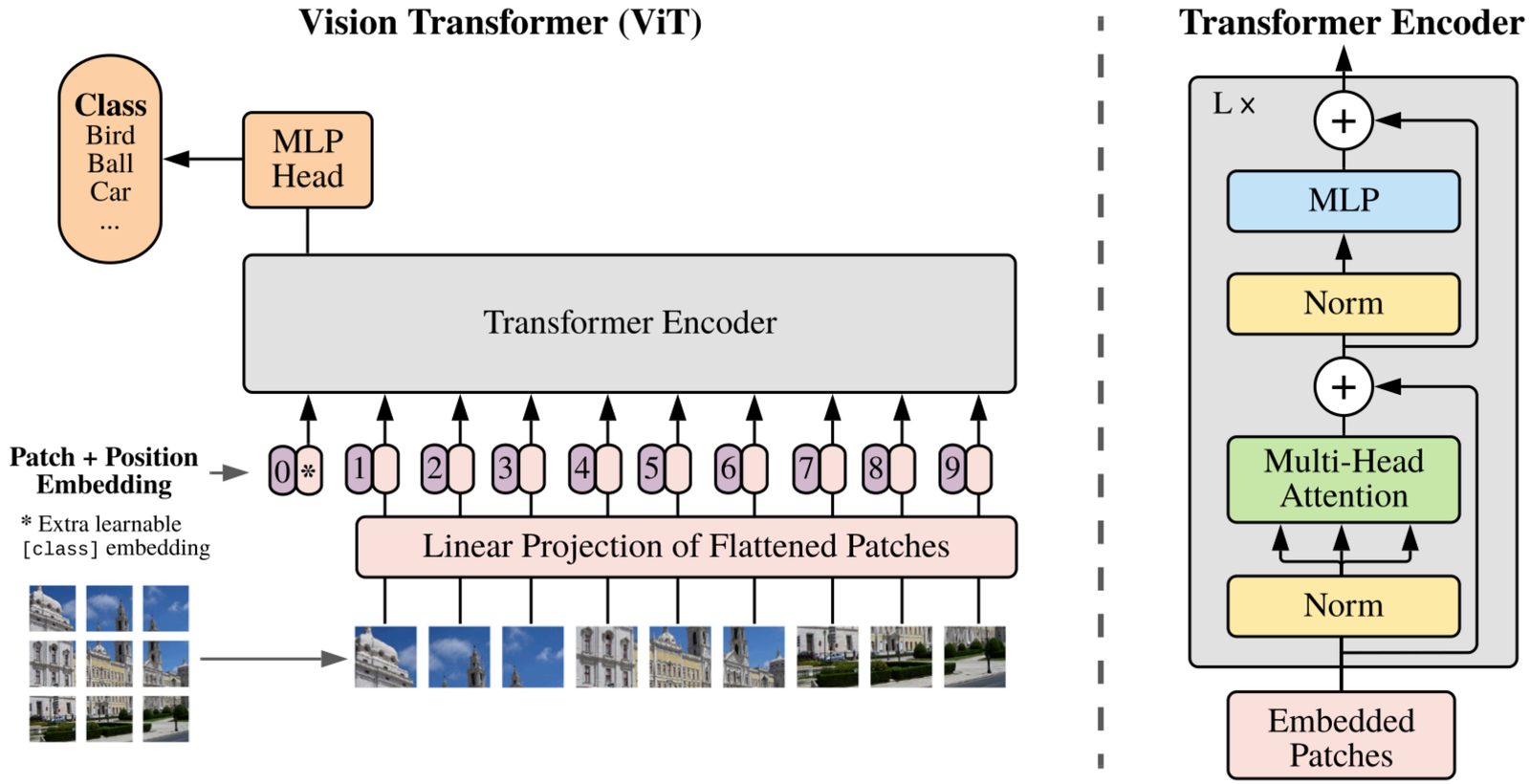


  - Separately encode the image and text, and train some other model that takes both representations and does something with them



Vision Transformers (ViT)

- *Idea:* divide image into patches, flatten the patches into vectors, use a standard Transformer to encode it



Vision Transformers (ViT)

Pixel height *Pixel width* *Pixel colors (R, G, B)*

$$x_{\text{image}} \in \mathbb{R}^{H \times W \times C}$$

Divide into patches

$$x_{\text{p}} \in \mathbb{R}^{N \times (P^2 \cdot C)}$$

$$x = Wx_{\text{p}}$$

Embed into vectors

$$x \in \mathbb{R}^{N \times D}$$

Let's assume there are 120x120 pixels.

We want 9 patches of 40x40 each.

$$H = 120,\ W = 120,\ P = 40$$

$$x_{\text{p}} \in \mathbb{R}^{9 \times (40^2 \cdot 3)}$$

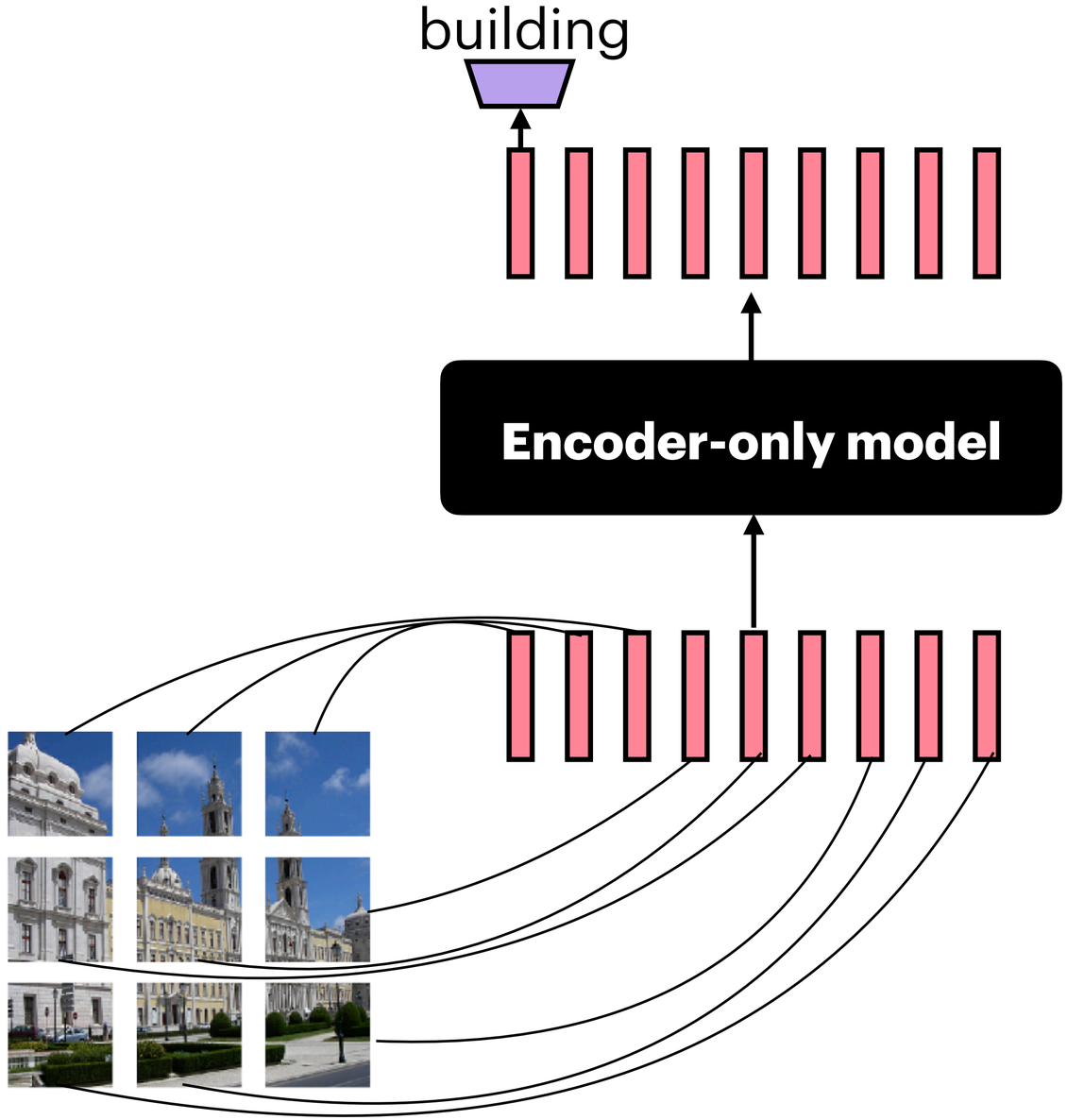
Each of these gets embedded into a vector of dimensionality D, which is a hyperparameter.

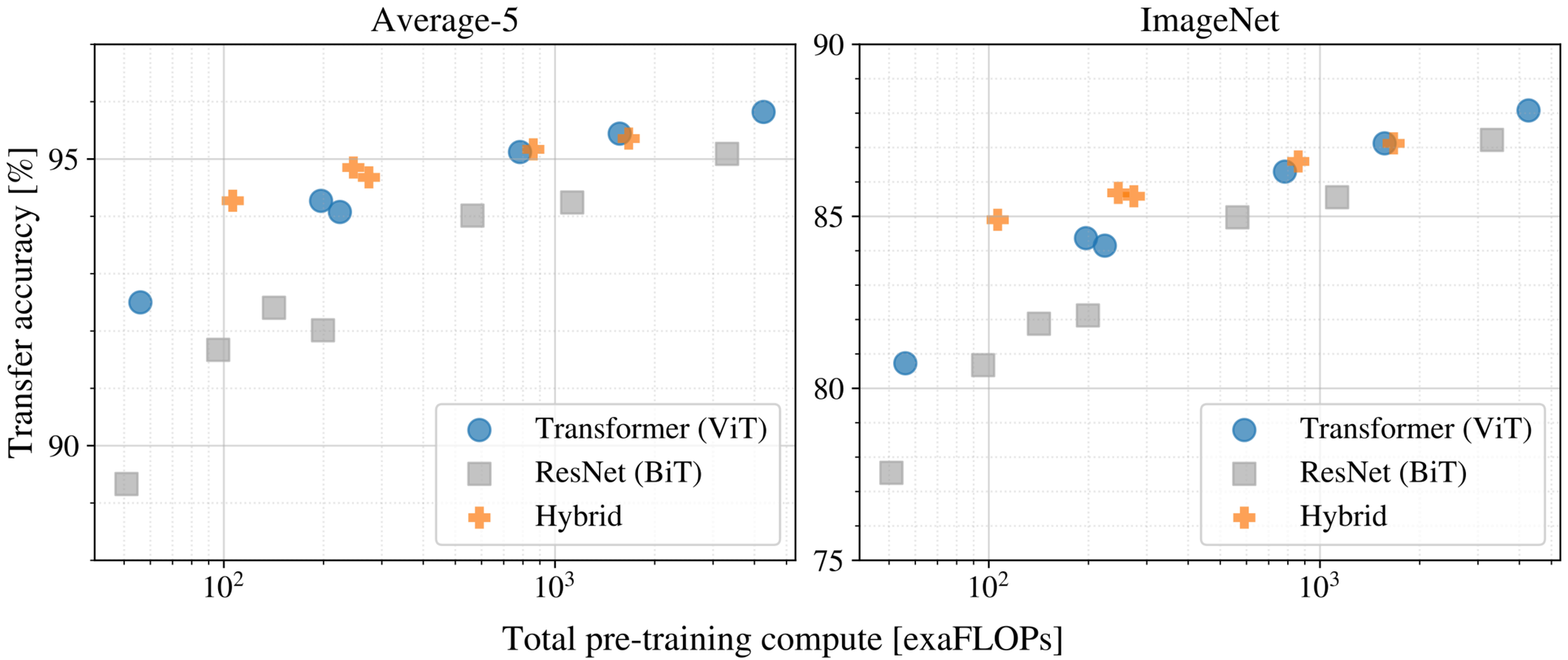
Vision Transformers (ViT)

- The ViT embeds patches into vector representations $\mathbf{x}_1, \ldots, \mathbf{x}_N$

- We can train this model using the exact same architecture we used for language modeling

- We could also concatenate text embeddings to image embeddings in the same input sequence!





Vision transformers generally outperform ResNets at the same compute budgets.

Hybrid methods do even better, but the gap narrows at larger compute budgets.

Learning Image Representations

- What would be a good pre-training objective for learning image representations?

- In computer vision, one often pre-trains using image classification: uses some textual supervision from the class label
  - Can also pre-train with captions; provides a much richer signal
  - Not super scalable: image pre-training largely limited to hand-labeled data

Contrastive Language–Image Pre-training (CLIP)

- Idea: learn image and text representations jointly in a shared embedding space
- Learn image encoder $f_I(x) = \mathbf{h}_I$
- Learn a separate text encoder $f_T(x) = \mathbf{h}_T$
- The representations for an image and its paired text in the training data should be similar
    - Representations for an unpaired image and text should be far apart
- Apply over a large corpus of (image, text) pairs

CLIP

Pre-training Data

- **Conceptual Captions (CC3M/12M):** extracts, filters, processes candidate image-caption pairs from the Internet

by Joi Ito

the trail climbs steadily uphill most of the way.

by Danail Nachev

the stars in the night sky.

by Justin Higuchi

musical artist performs on stage during festival.

by Viaggio Routard

popular food market showing the traditional foods from the country.

- **LAION-5B**: an open and very large-scale image-text dataset

2.3B images w/ English descriptions

Blue Beach Umbrellas, Point Of Rocks, Crescent Beach, Siesta Key - Spiral Notebook

2.3B images w/ non-English descriptions

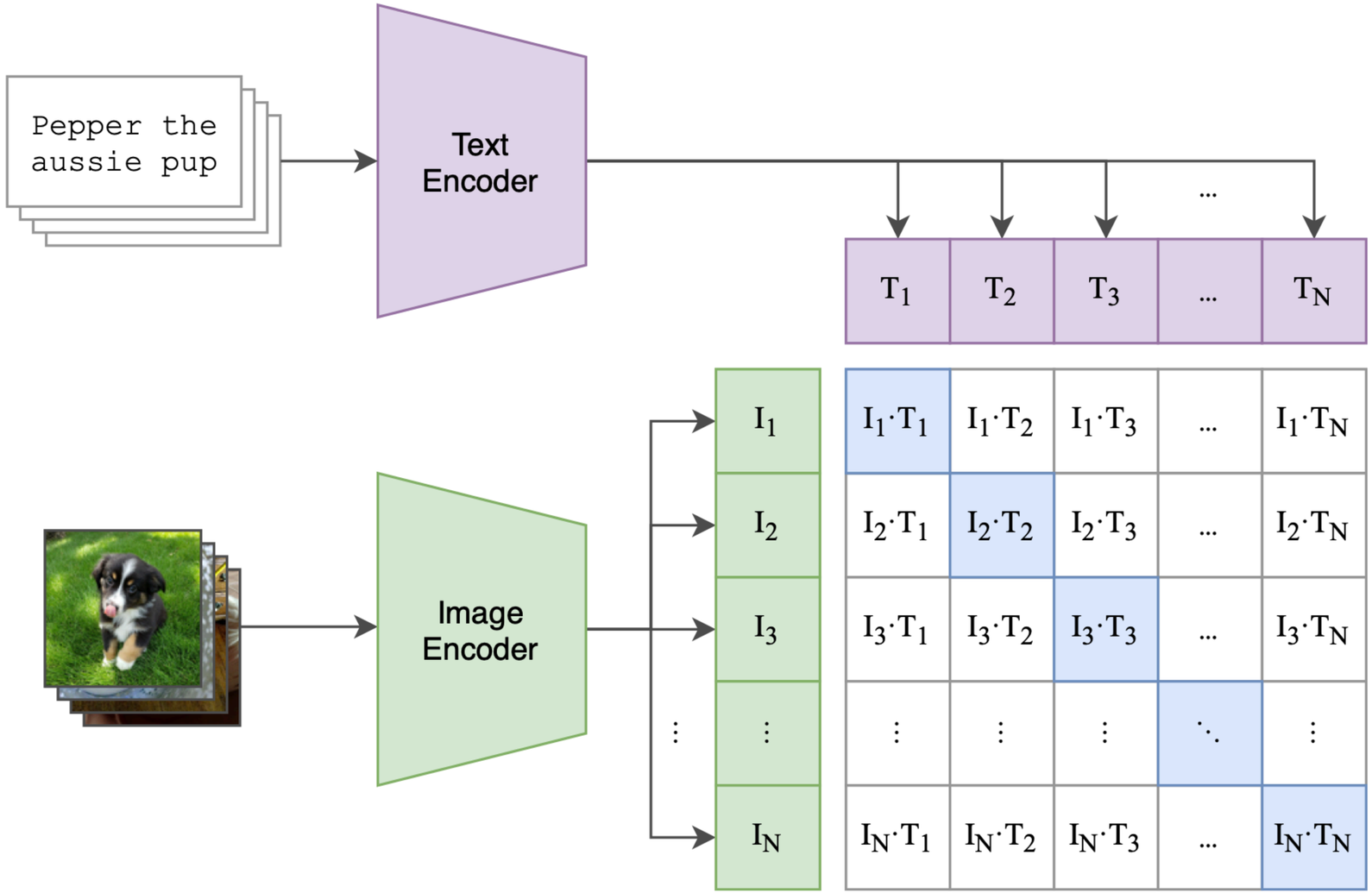
Episcopia Ortodoxa a Maramuresului si Satmarului are un nou Arhiereu vicar

1.3B images where language of description could not be determined

Europe, Italy

CLIP

Given N (image, text) pairs per training step, learn embeddings that make it easy to classify which image is paired with which text



CLIP

Image-text pair

$$L((x_1, y_1), \ldots, (x_N, y_N)) =$$

$$-\frac{1}{2}\sum_{n=1}^{N}\left[\log \frac{\exp\left(f_I(x_n)^\top f_T(y_n)\right)}{\sum_j \exp\left(f_I(x_j)^\top f_T(y_n)\right)} + \log \frac{\exp\left(f_I(x_n)^\top f_T(y_n)\right)}{\sum_j \exp\left(f_I(x_n)^\top f_T(y_j)\right)}\right]$$

Softmax over images

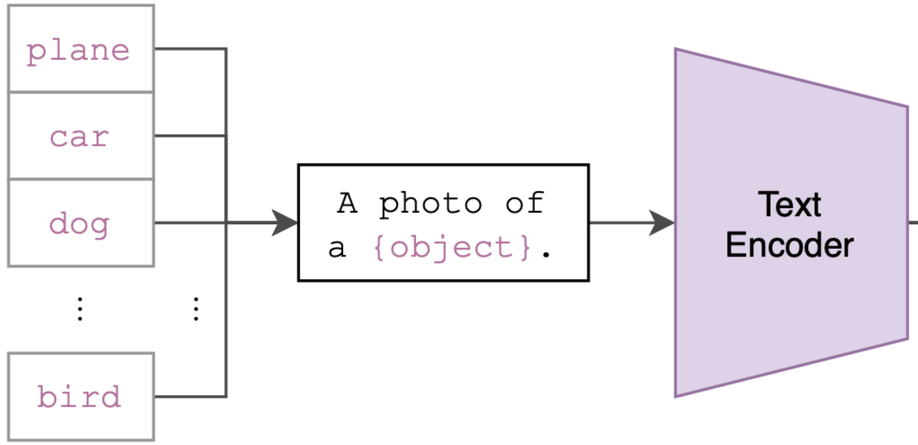
Softmax over text

Incentivizes the dot product of embeddings of an image-text pair to be high

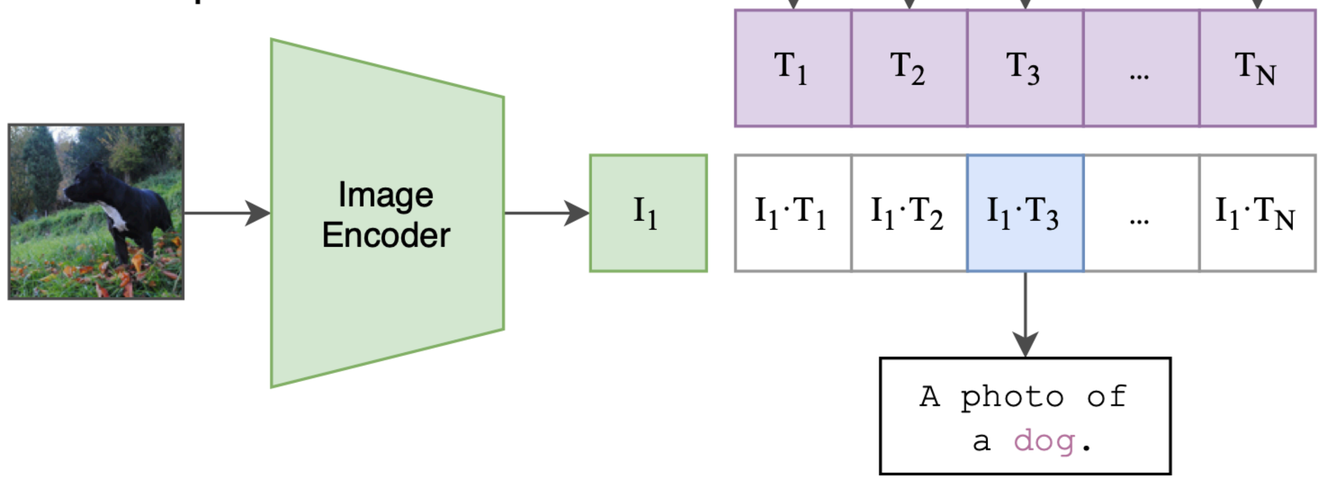
Incentivizes the dot product of embeddings of non-paired image and text to be low

CLIP

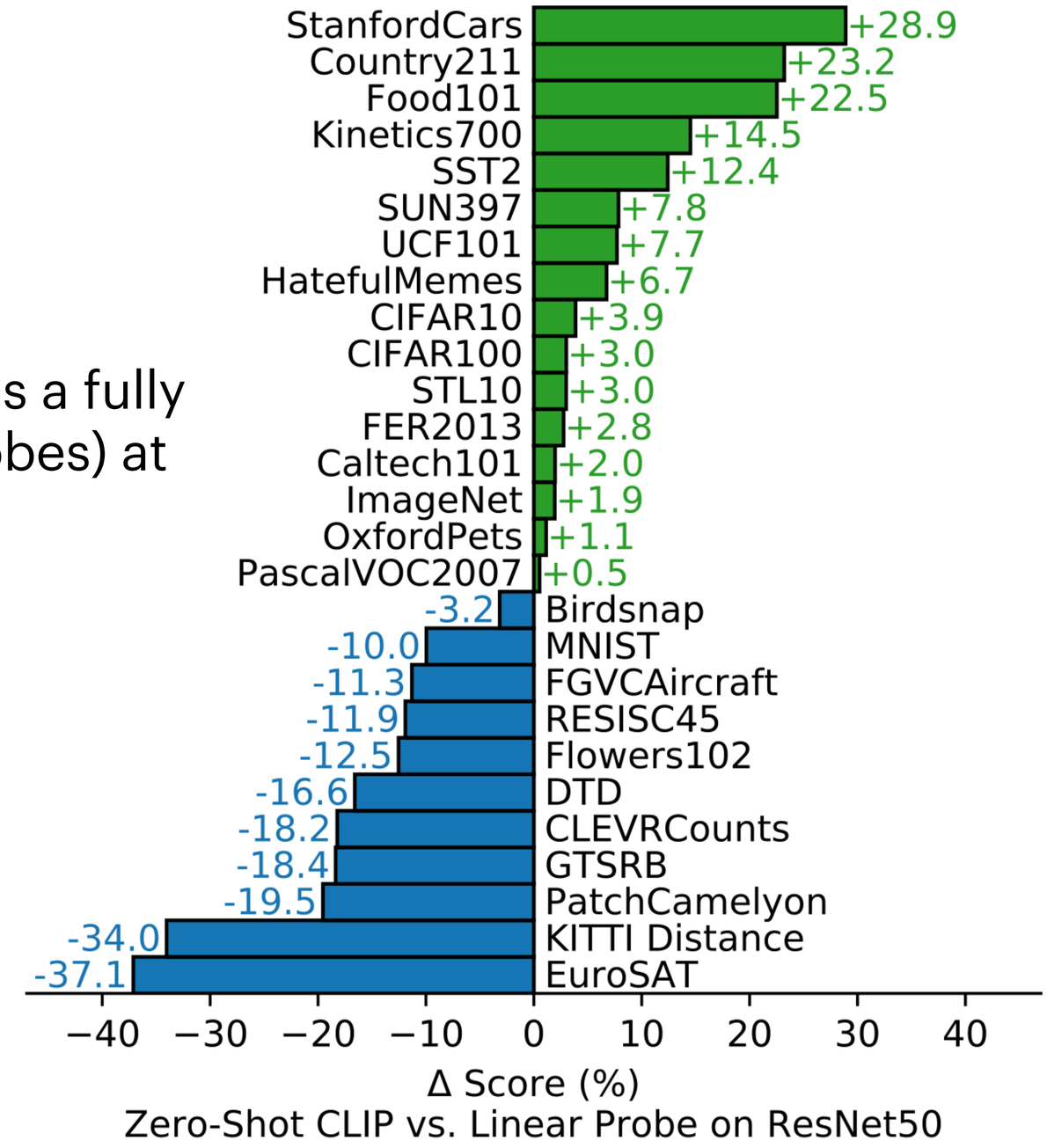
(2) Create dataset classifier from label text



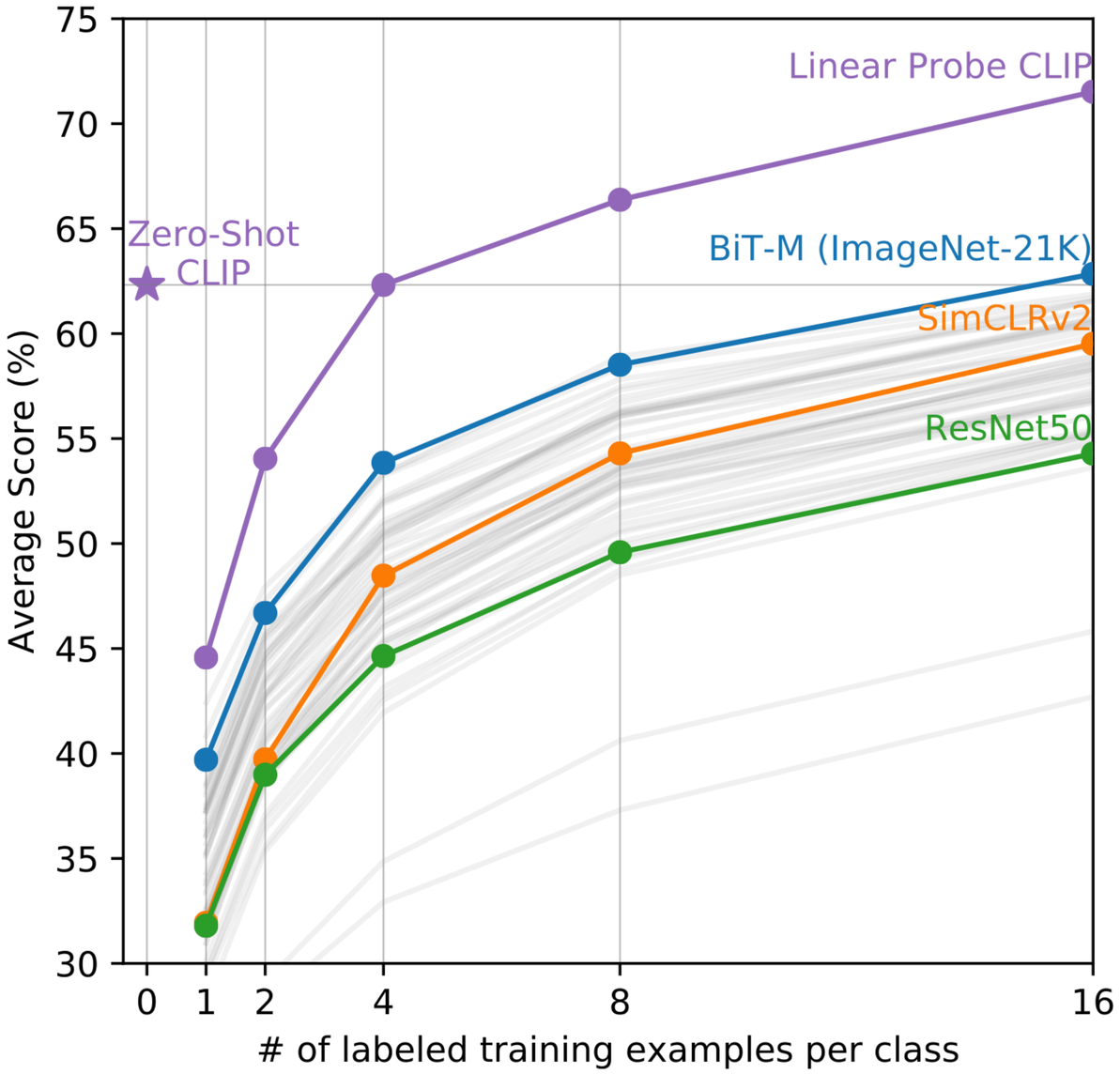
(3) Use for zero-shot prediction



On average, CLIP's zero-shot performance is about as good as a fully supervised approach (linear probes) at classifying images



CLIP is very good at few-shot learning.

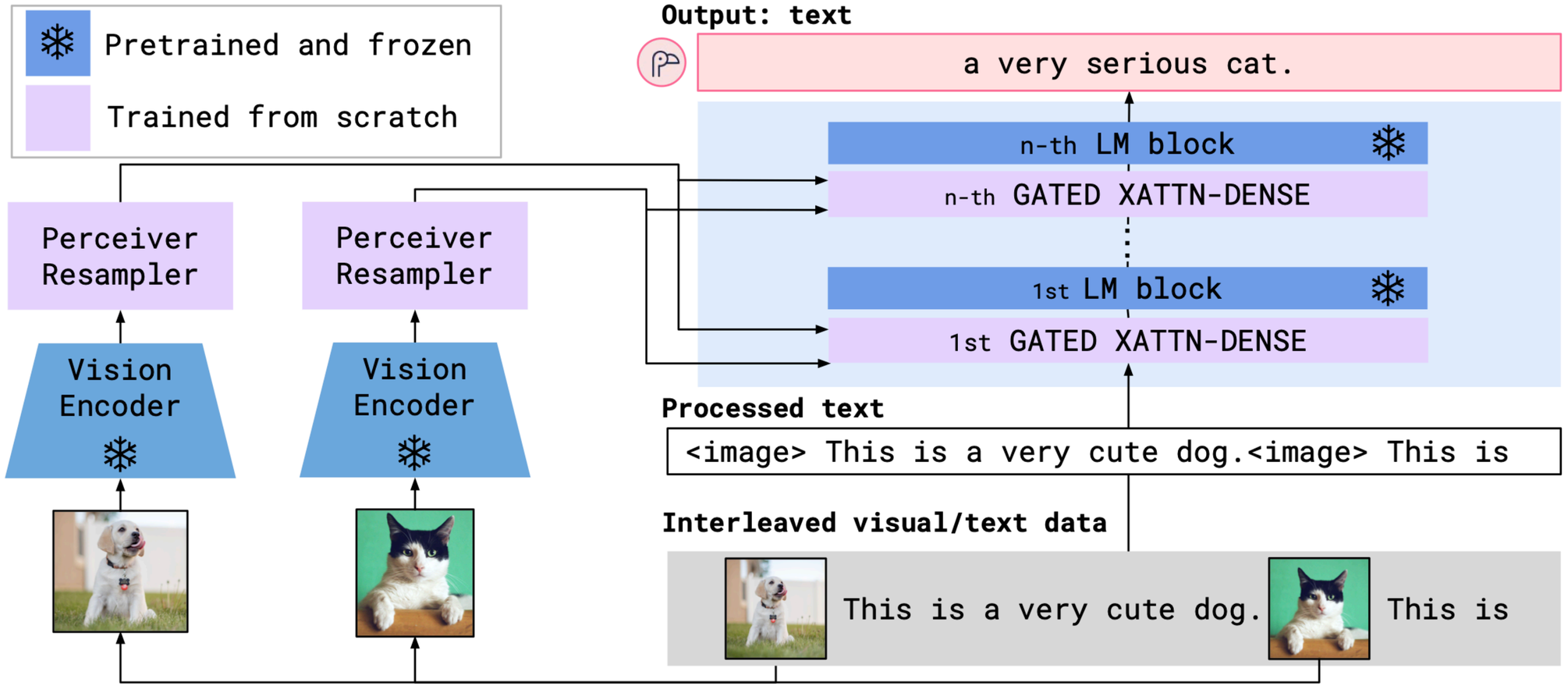


Classes of Multimodal NLP System

- Vision language models: take in images and text, generate text
  - Flamingo
  - LLaVA

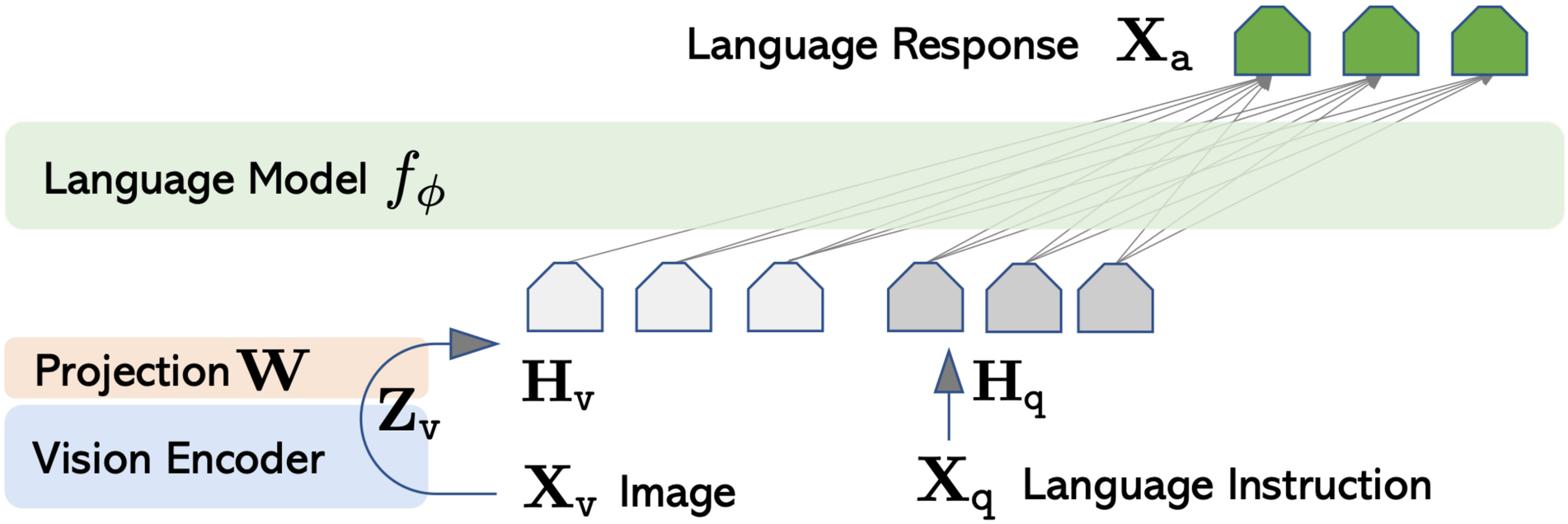
- Image generation: take in text, generate an image
  - Stable Diffusion

Flamingo



- Flamingo interweaves image tokens with text tokens
- Takes a pre-trained image encoder (e.g., NFNet, or, more commonly these days, DINO) and pre-trained LM. Model only trains components that help it use these representations together

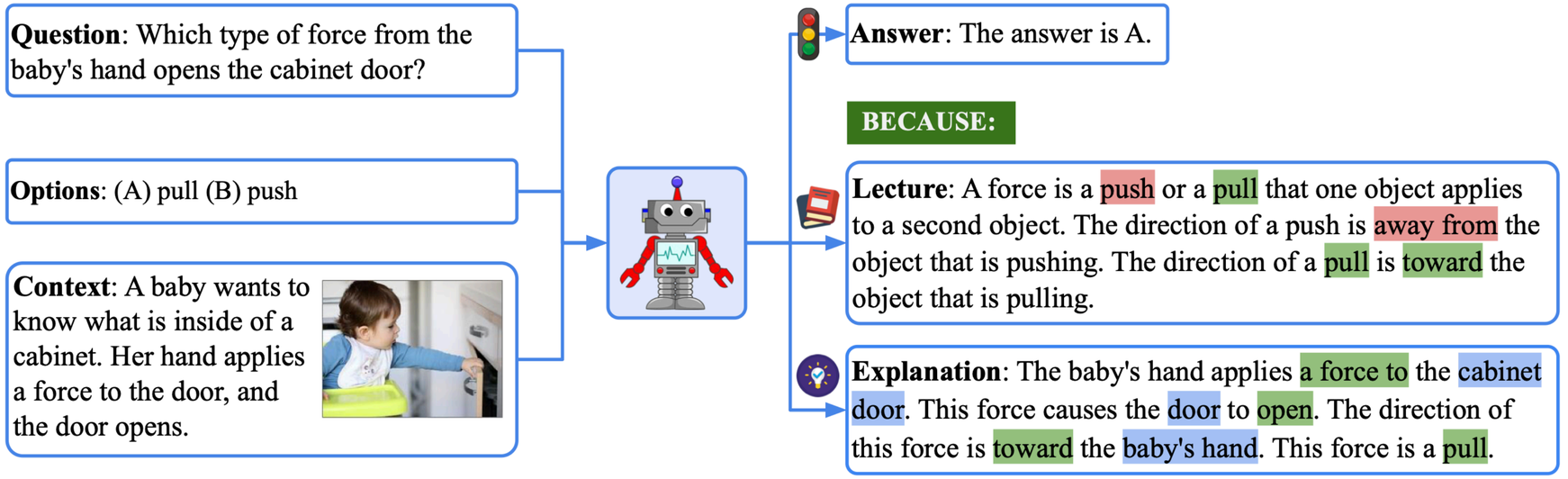
Large Language and Vision Assistant (LLaVA)



- Architecturally simpler than Flamingo: just projects the vision embedding into the language space using a learned MLP

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.
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
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.



Accuracy on ScienceQA is relatively high.

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 [34]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [34]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [34]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
LLaMA-Adapter [59]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT$_{Base}$ [61]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{Large}$ [61]	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**

Using Vision Language Models

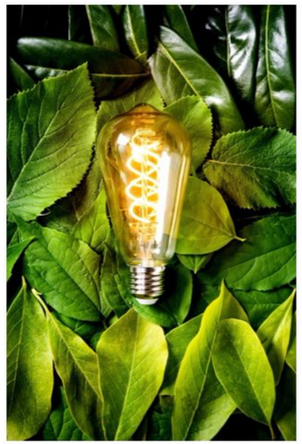
- Image preprocessing: split image into patches, flatten
- Image encoding: use CLIP, or get ViT vectors from last layer of model
- Give encodings to LLMs: e.g., transform vectors to be LM's embedding dimensionality
- Train or fine-tune on data with text and images
    - VQA, Image captioning

Vision Language Model Tasks

Winoground

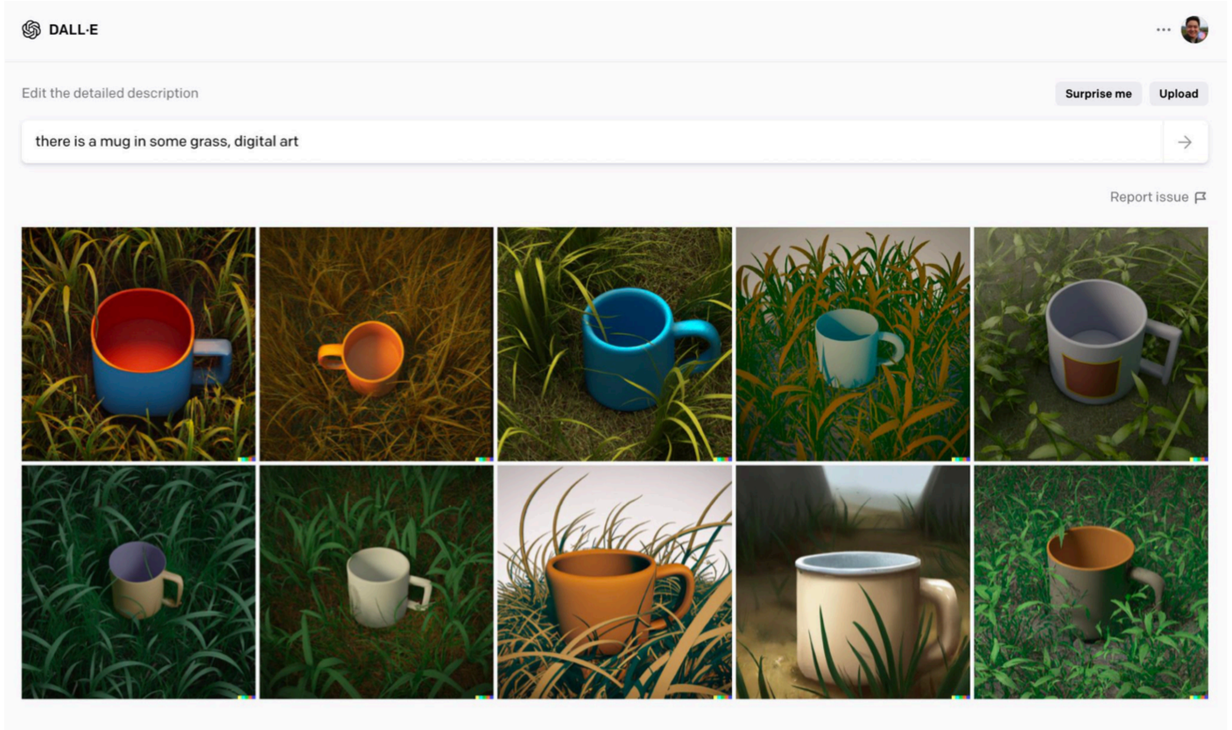
- Task: assign a higher probability to the correct caption given an image
    - Includes paired samples where (e.g.) a containment relation is reversed

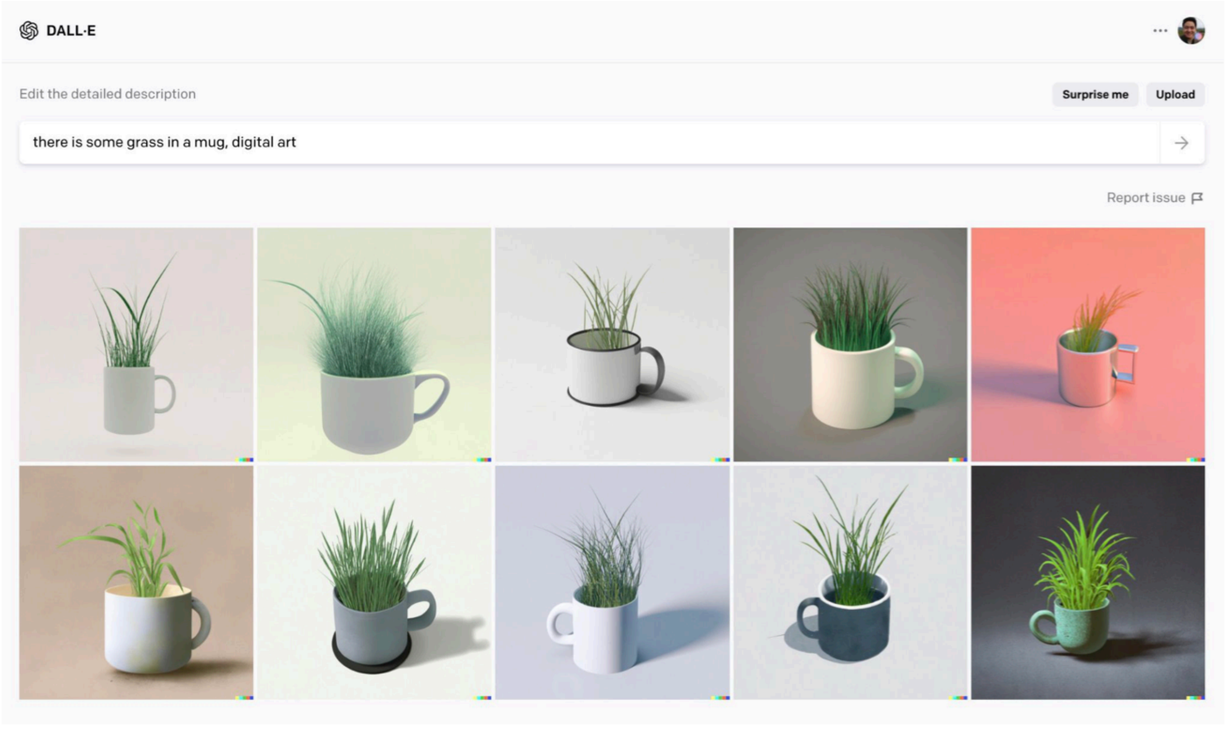
- Very hard task, even for state-of-the-art models!

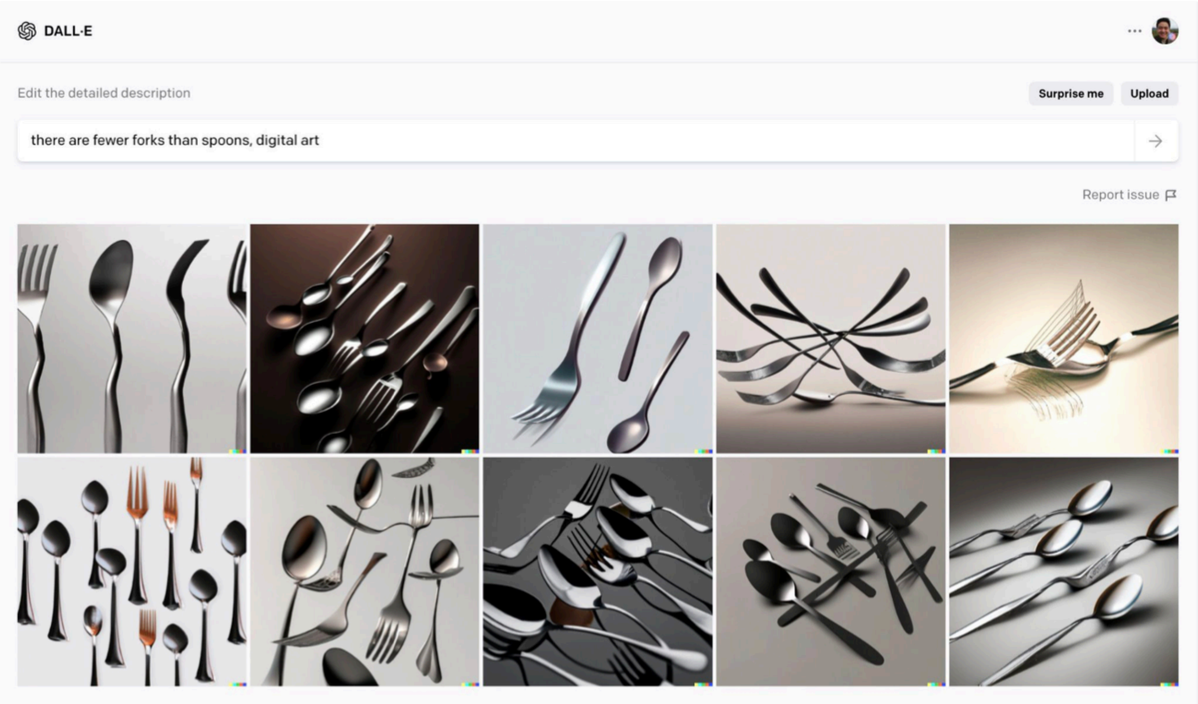


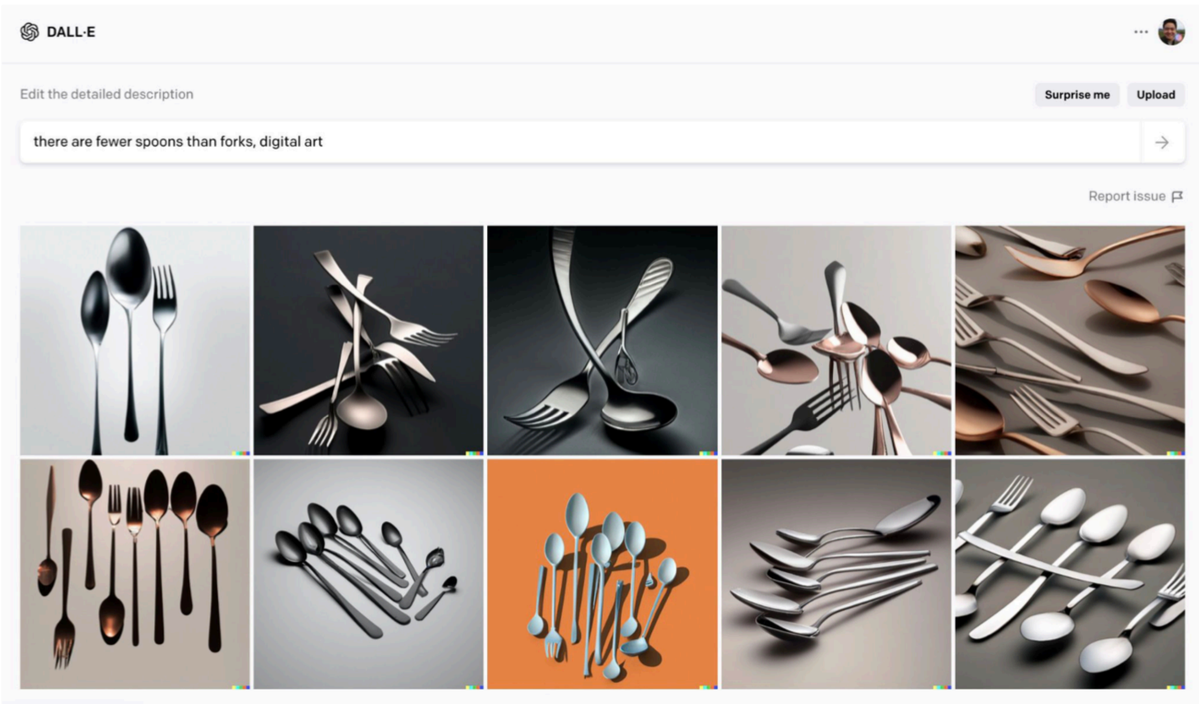
(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants


DALL·E
Edit the detailed description
Surprise me
Upload
there is a mug in some grass, digital art
Report issue


DALL·E
Edit the detailed description
Surprise me
Upload
there is some grass in a mug, digital art
Report issue





Even very good VLMs are still weak at compositional reasoning and quantifier semantics.

Vision Language Model Tasks

Visual Question Answering

Rank	Participant team	yes/no (↑)	number (↑)	Last submission at
1	B. Zacharie (c_q7_m8)	97.35	81.17	1 month ago
2	Allen Institute for AI (Molmo2-8B)	96.91	81.35	4 months ago
3	PaliGemma 2 (10B, 448px, finetune)	97.19	77.77	1 year ago
4	Allen Institute for Artificial Intelligence (Molmo-72B)	96.91	78.49	2 years ago
5	PaLI-X - Google Research (Single Generative Model)	96.78	74.14	3 years ago
6	PaliGemma-3B (finetune, 448px)	96.39	76.29	2 years ago
7	LQM	96.17	75.42	4 months ago
8	quanmin	96.02	74.64	4 months ago
9	Zhipu AI	95.71	71.97	3 years ago

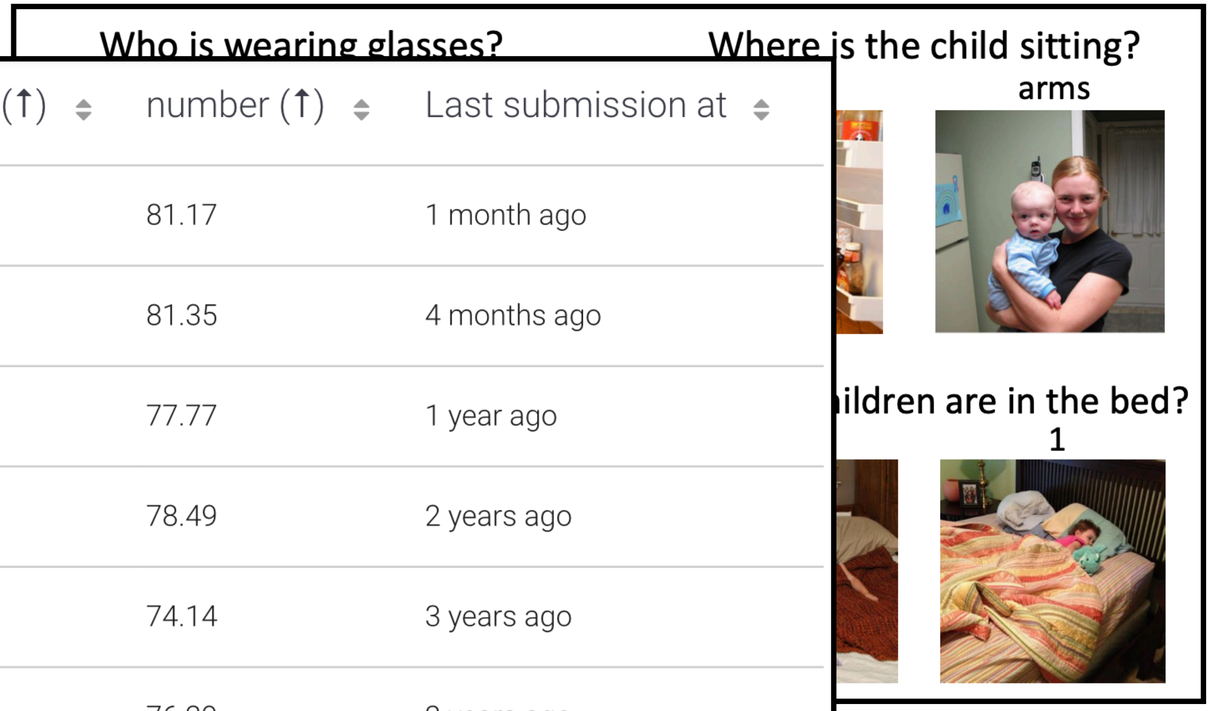


Image Generation

- So far, we've discussed how to take images and generate text with them.
- How can we go the opposite direction? Take text, generate images

- The most common architecture in this setting is *diffusion:*

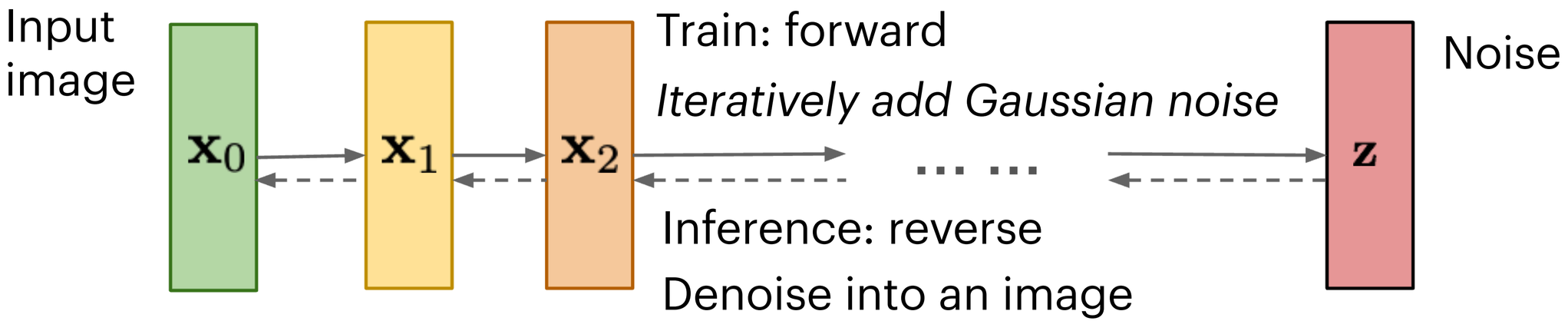
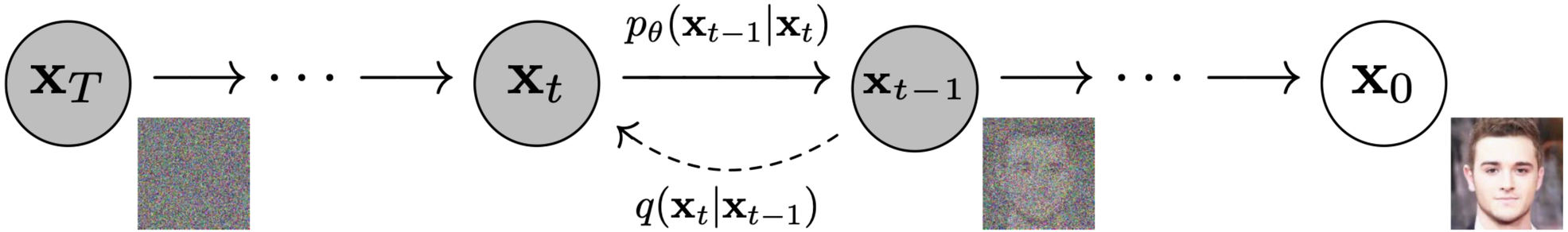
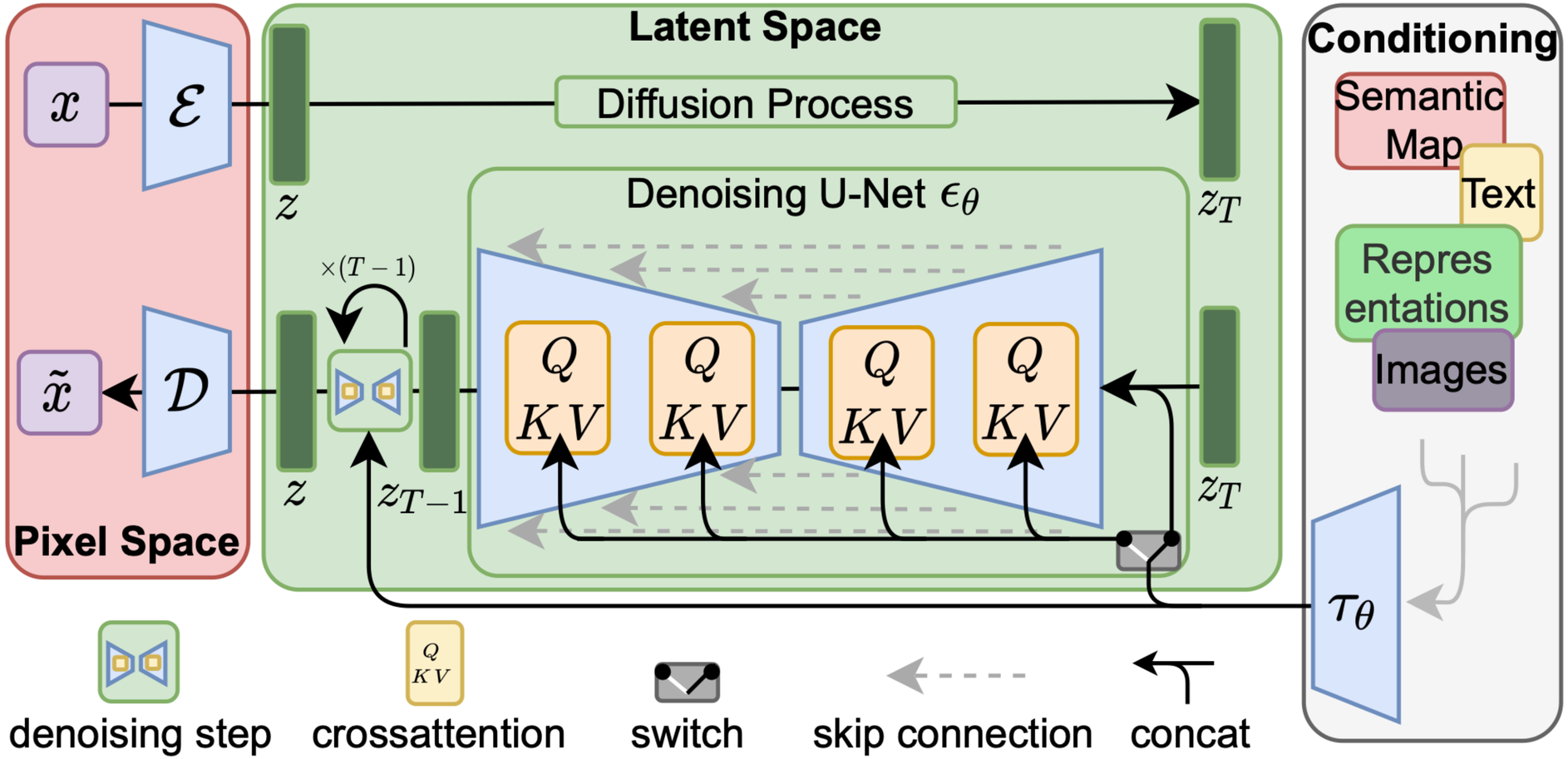


Image Generation with Diffusion

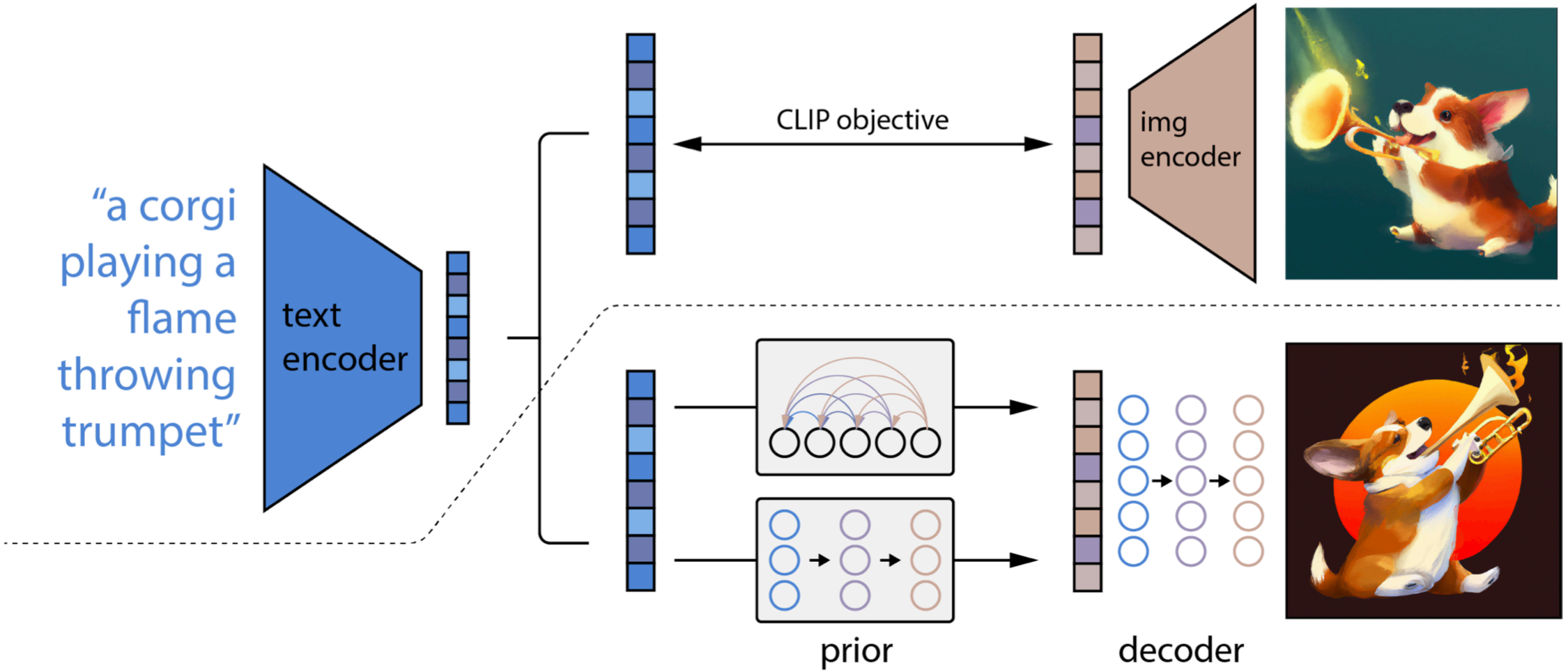


- Diffusion works very differently from language models
- During training, we take an image and text, and progressively add noise to the image
- During inference, we reverse this process: start with noise, and iteratively remove noise to obtain an image

Stable Diffusion

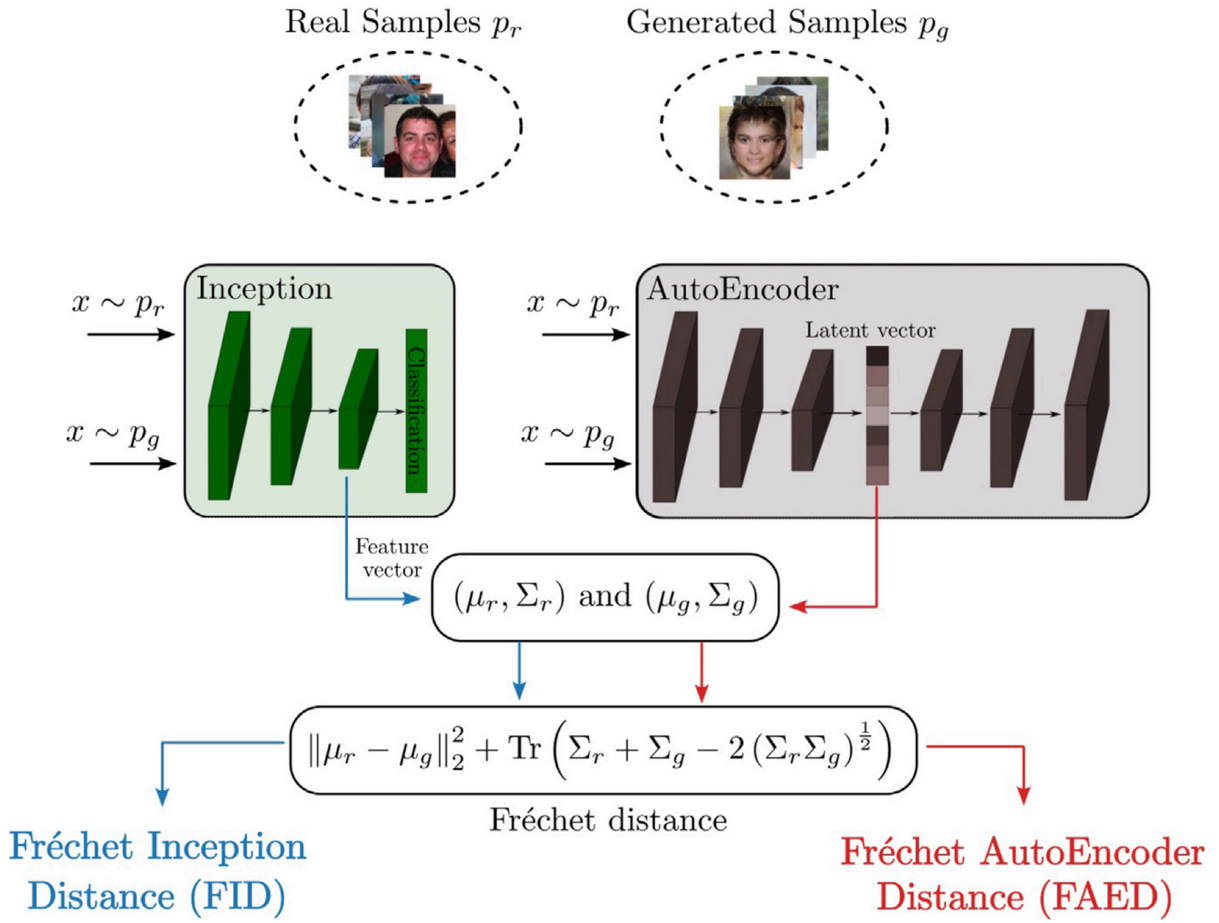


DALL-E 2


"a corgi playing a flame throwing trumpet"
text encoder
CLIP objective
img encoder
prior
decoder

Evaluating Image Generators

- How do we evaluate the quality of a generated image?
- **Fréchet Inception Distance (FID)**: measures the similarity of a dataset of real images vs. a dataset of generated images
  - Extract feature distributions from images, compute how different the train set is from the generated set



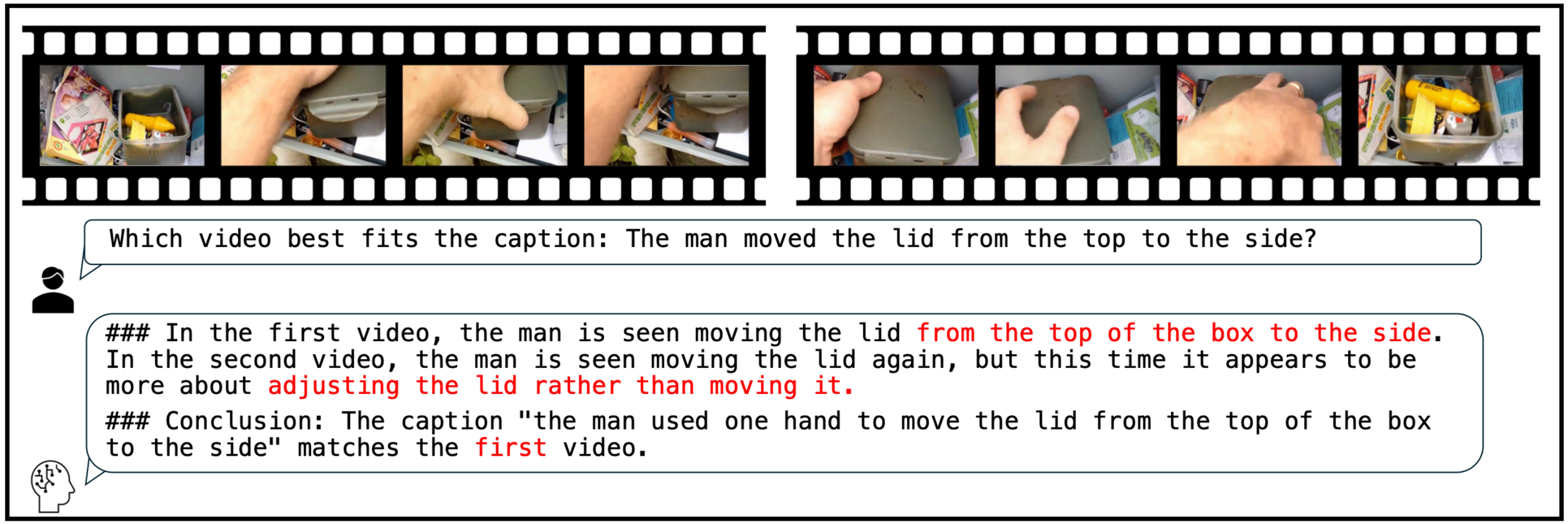
Copyright Issues

- Image generation models *require* the work of artists, including illustrators and photographers, for training.
- LAION-5B contains many copyrighted images, and LAION claims no ownership of them. The team claims that the images are under their original creators' copyright.

- A class-action lawsuit was filed against StabilityAI (developer of Stable Diffusion), DeviantArt, and Midjourney, alleging copyright infringement, terms-of-service violations, unfair competition, among other damages.

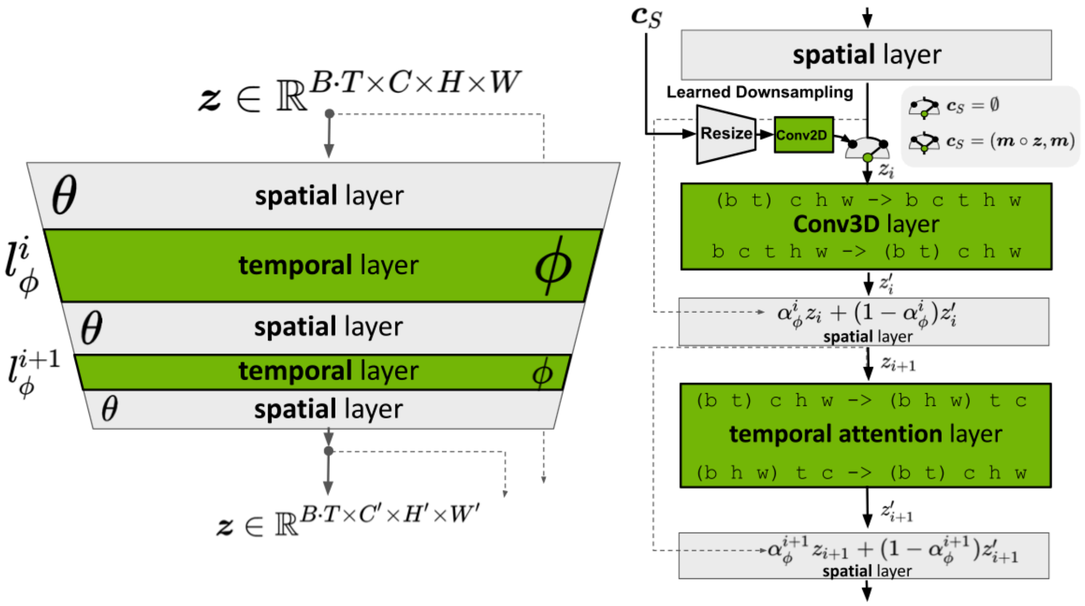
Frontiers in Multimodal NLP

Video–Language Models

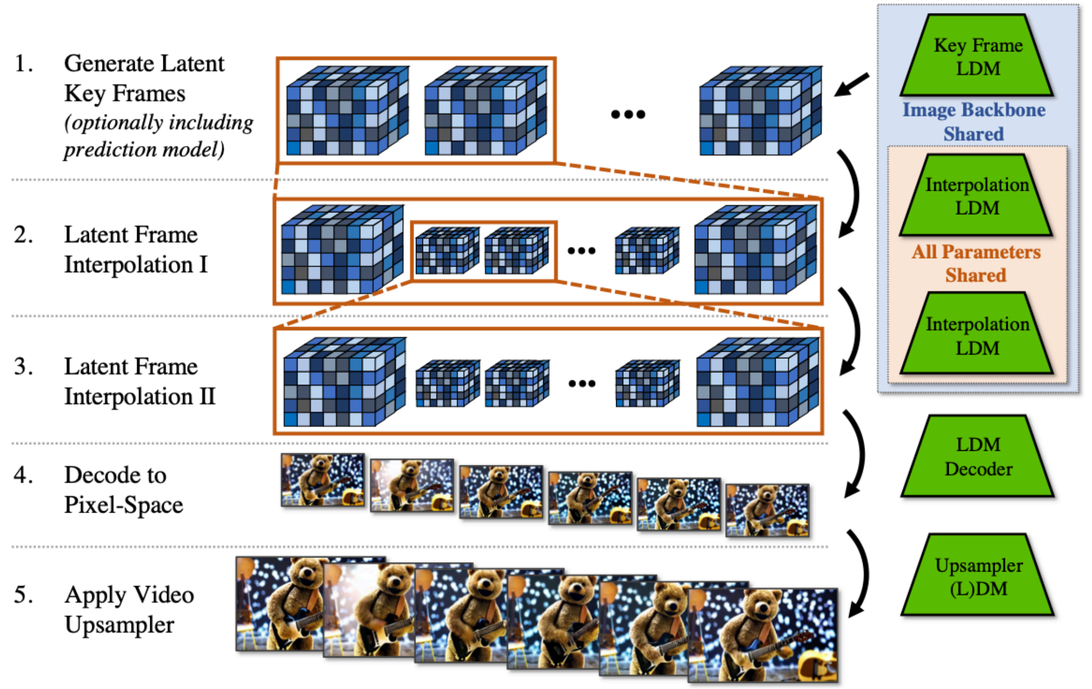


- How can we instill a sense of temporality in the image space?
- Data is relatively scarce
- Image-and-text models are strong backbones; can adapt them for video. See LLaVa-OneVision

Text-to-Video Models



(a) **Additional temporal layer**. A pre-trained LDM is turned into a video generator by inserting temporal layers that learn to align frames into temporally consistent sequences. During optimization, the image backbone θ remains fixed and only the parameters ϕ of the temporal layers l^i_ϕ are trained.



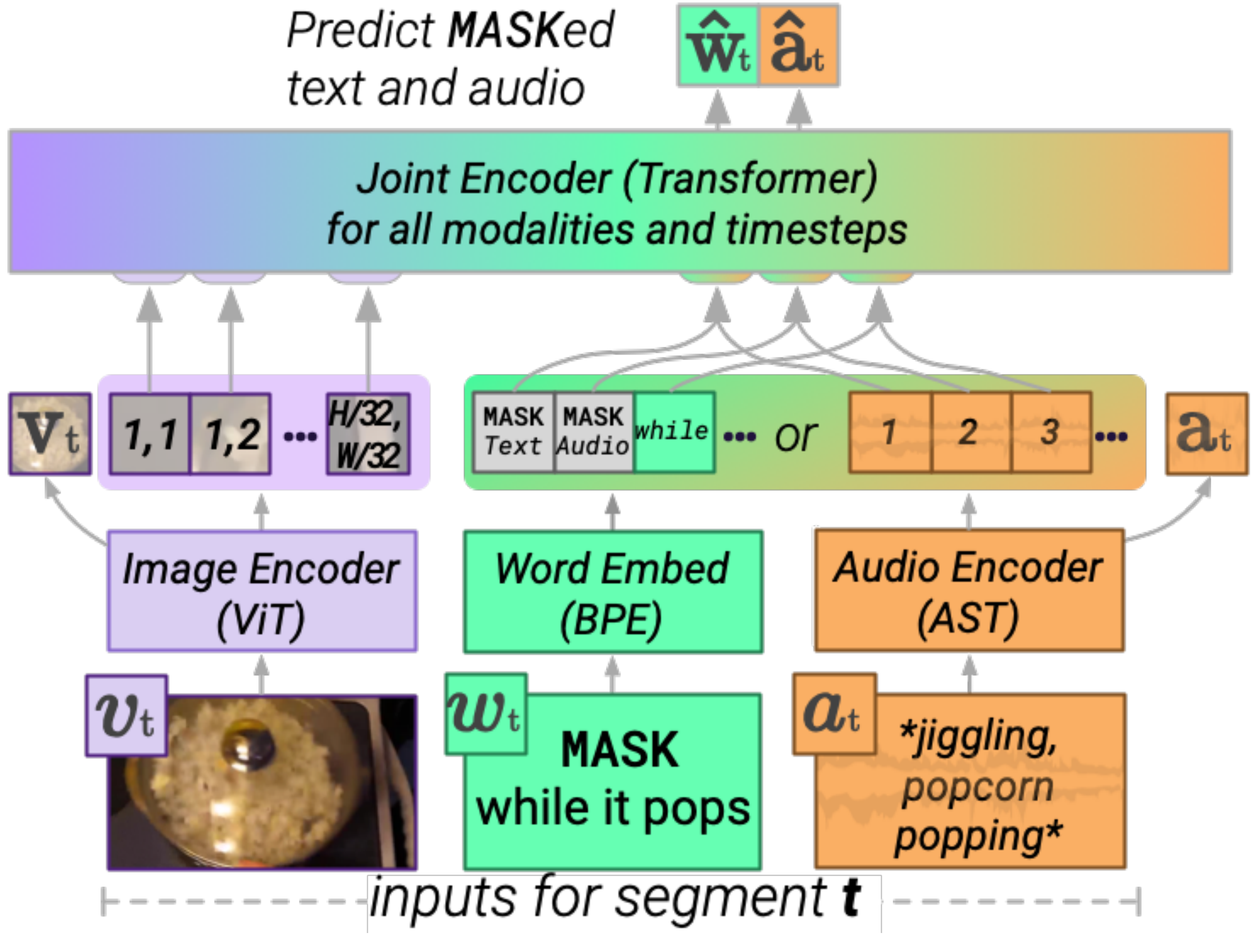
(b) **Video LDM stack.** Video LDM first generates sparse key frames and then temporally interpolates twice with the same latent diffusion models to achieve a high frame rate. Finally, the latent video is decoded to pixel space, and optionally, a video upsampler diffusion model is applied.

Systems like Sora are based on a mixture of Transformers and diffusion.

Frontiers in Multimodal NLP

Many-modal Models

- MERLOT Reserve: a vision-audio-language model
- Gets great performance on TVQA and VCR (video understanding tasks)
- Audio signals believed to help with physical dynamics and social reasoning

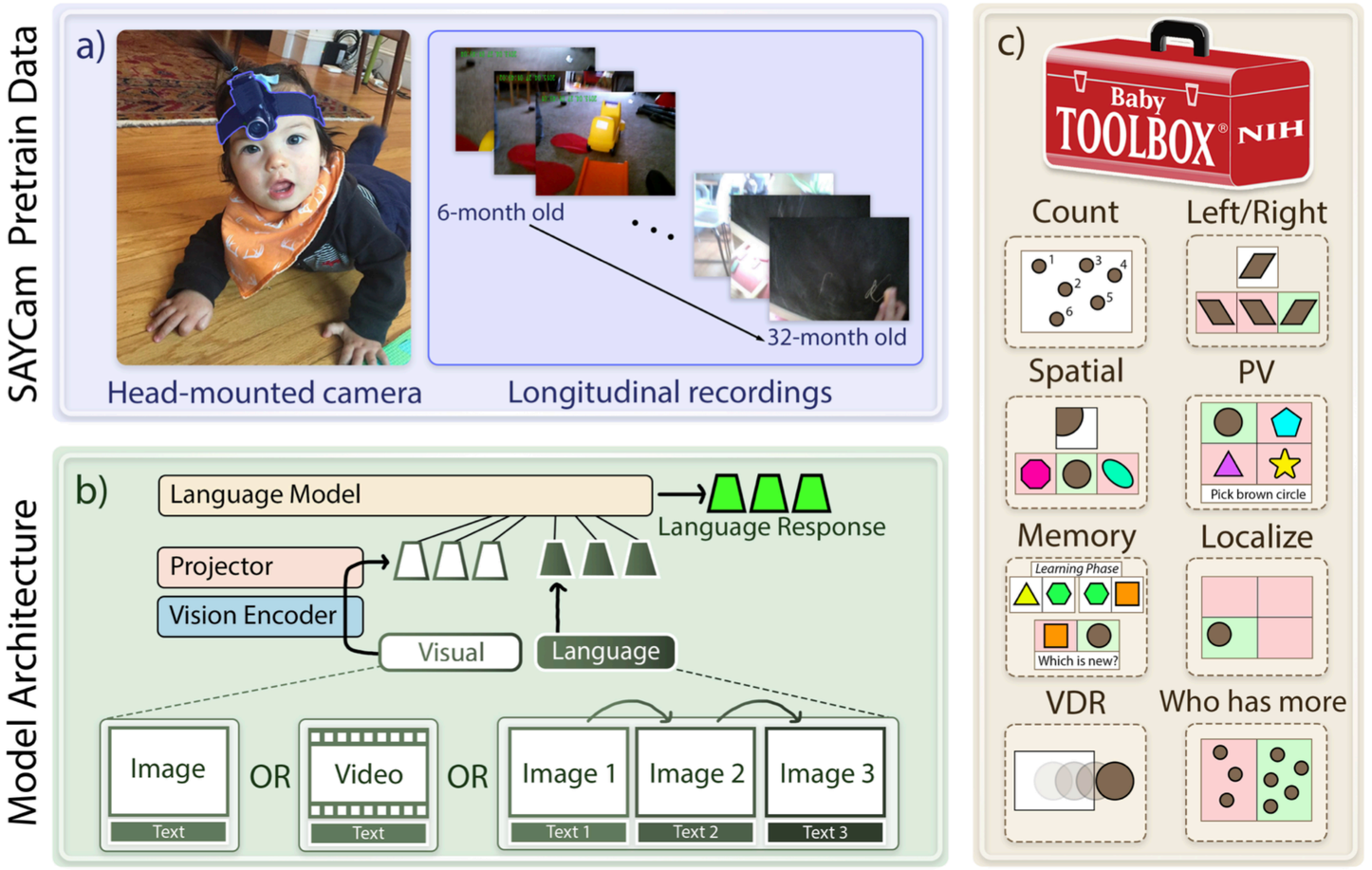


Frontiers in Multimodal NLP

BabyVLM

- SAYCam: a dataset of video data from headcams attached to children, with transcribed speech

- BabyVLM(-v2): A video/vision-language model trained on developmentally plausible data
    - Developed here at BU!



Multimodal ML at BU!

If you’re interested in this topic, we have a whole course dedicated to multimodal machine learning at BU!

CS598

We have many great computer vision faculty here. Do take their courses and learn from them!

Next Time

- Human language processing
- How human-like are pre-trained language models in how they process language?
- Can we train better LMs on more human-like data?
- How can we train LMs to be better models of human cognition?