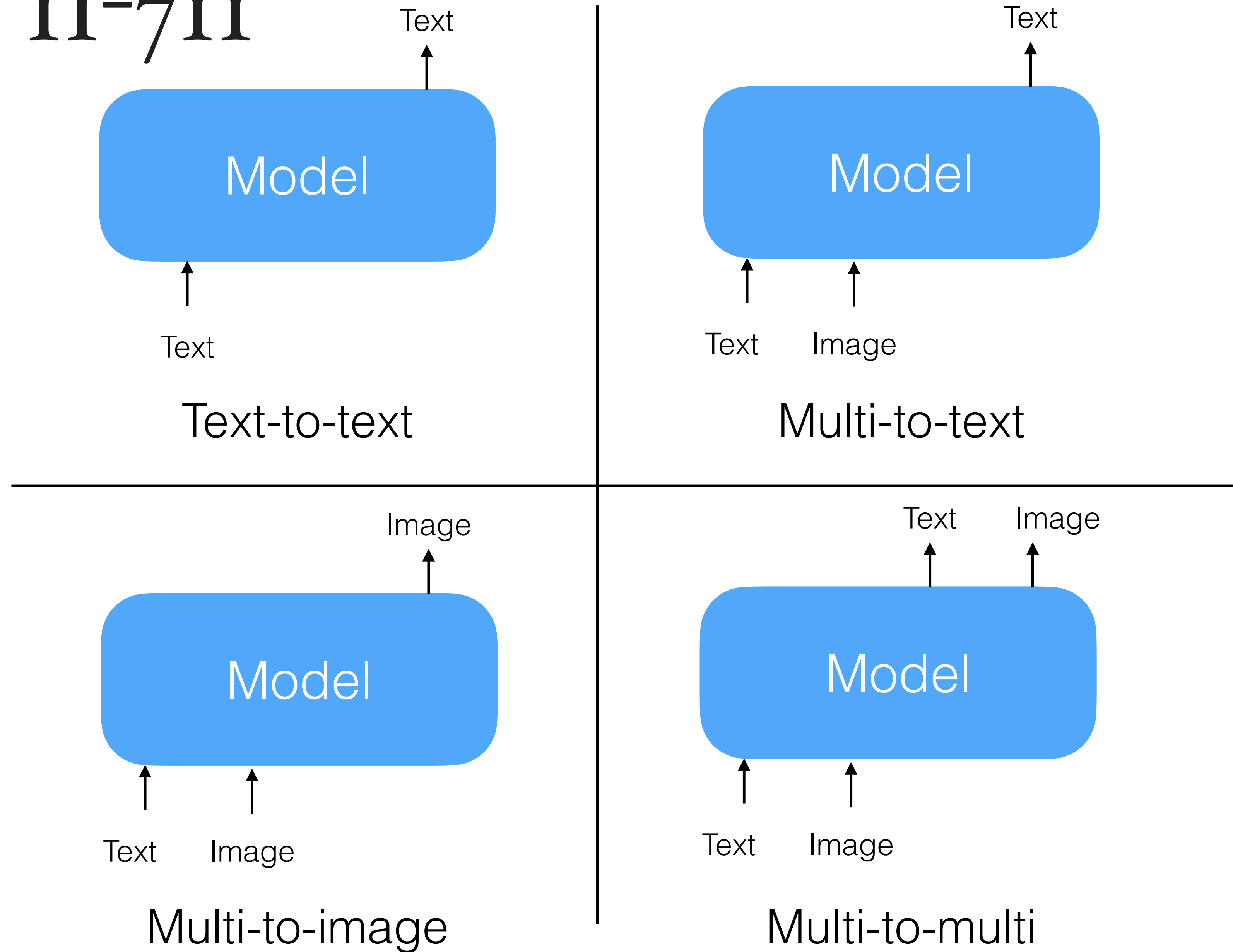


Large Language Model Applications

Image generation and conversing about images

From II-7II



Where language and vision intersect

Image captioning



→ A tabby and white cat is captured in a mid-stretch or liquid-like slump, draping its front half out of a gray and black pet carrier mounted to the front handlebars of an Aventon electric cargo bike. The bike, which is a light sage green color, is parked on a sun-dappled concrete driveway next to a wooden fence and a small garden bed with red flowers. The cat's head hangs low toward the front tire, showing off its white paws and tabby-patterned back. The bicycle is heavily outfitted for transport, featuring a large black pannier bag on the rear rack, a smaller black bag tucked under the frame, and a stack of blue and white plastic bins secured to the back with blue bungee cords. A black rain barrel stands in the background behind a section of chain-link fence, and the ground is scattered with fallen autumn leaves, suggesting a crisp, sunny day. The bike's kickstand is deployed, and the front wheel is turned slightly, while the "AVENTON" branding is clearly visible on the lower frame tube.

Where language and vision intersect

Image question answering

Q: What is strapped to the bike?



The Front Load

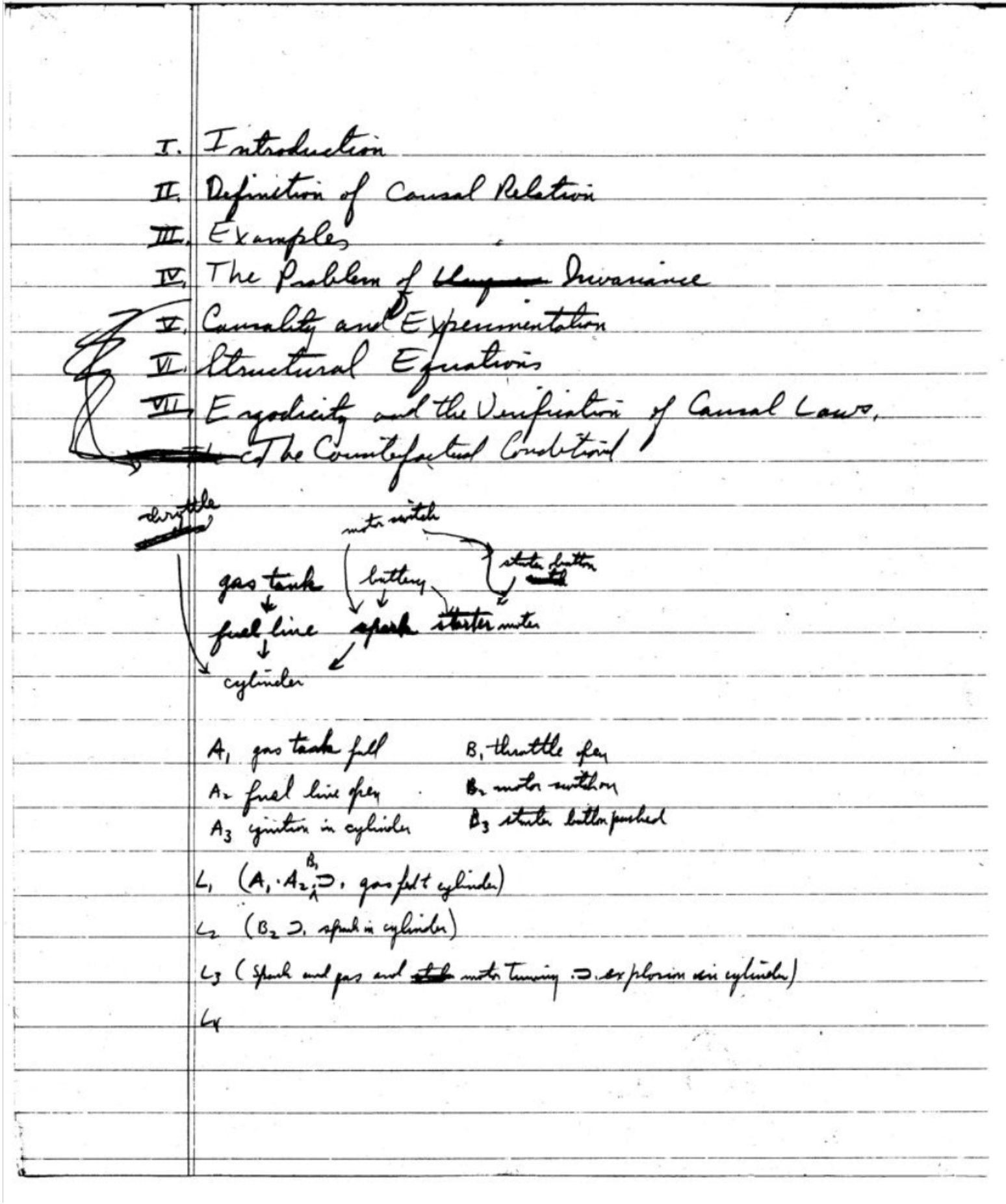
- **A Tabby Cat:** The most prominent "cargo," currently performing a dramatic liquid-like exit from its carrier.
- **Pet Carrier:** A grey and black mesh pet travel bag sitting inside the front basket.
- **Front Basket:** A sturdy black metal rack/basket attached to the frame.

The Rear Load

- **Plastic Bins:** A white plastic bin on the bottom with a blue plastic bin nested on top.
- **Bungee Cords:** Blue elastic cords crisscrossing to secure the bins to the rear rack.
- **Large Pannier Bags:** Heavy-duty black storage bags draped over both sides of the rear wheel.
- **Small Accessory Bag:** A smaller black bag tucked just behind the seat post.

Where language and vision intersect

Optical character recognition (OCR)



I. Introduction

II. Definition of Causal Relation

III. Examples

IV. The Problem of Invariance

V. Causality and Experimentation

VI. Structural Equations

VII. Ergodicity and the Verification of Causal Laws, and the Counterfactual Conditional

<diagram>A causal flowchart diagram. Components include 'throttle', 'gas tank', 'fuel line', 'cylinder', 'motor switch', 'battery', 'starter button', 'starter motor', and 'spark'. Arrows represent causal links: 'throttle' points to 'cylinder'; 'gas tank' points to 'fuel line' which points to 'cylinder'; 'motor switch' points to 'starter button'; 'battery' points to both 'starter button' and 'spark'; 'starter button' points to 'starter motor'; 'starter motor' points to 'spark'; and 'spark' points to 'cylinder'.</diagram>

A₁ gas tank full

A₂ fuel line open

A₃ ignition in cylinder

B₁ throttle open

B₂ motor switch on

B₃ starter button pushed

$L_1 (A_1, A_2, \overset{B_1}{\supset} gas \parallel fuel \parallel to \parallel cylinder)$

$L_2 (B_2 \supset spark \parallel in \parallel cylinder)$

$L_3 (spark \parallel and \parallel gas \parallel and \parallel motor \parallel turning \supset explosion \parallel in \parallel cylinder)$

L_4

Where language and vision intersect

Image generation

A tabby and white cat is captured in a mid-stretch or liquid-like slump, draping its front half out of a gray and black pet carrier mounted to the front handlebars of an Aventon electric cargo bike. The bike, which is a light sage green color, is parked on a sun-dappled concrete driveway next to a wooden fence and a small garden bed with red flowers. The cat's head hangs low toward the front tire, showing off its white paws and tabby-patterned back. The bicycle is heavily outfitted for transport, featuring a large black pannier bag on the rear rack, a smaller black bag tucked under the frame, and a stack of blue and white plastic bins secured to the back with blue bungee cords. A black rain barrel stands in the background behind a section of chain-link fence, and the ground is scattered with fallen autumn leaves, suggesting a crisp, sunny day. The bike's kickstand is deployed, and the front wheel is turned slightly, while the "AVENTON" branding is clearly visible on the lower frame tube.



Where language and vision intersect

Image generation

A tabby and white cat trying to
escape a bicycle



Where language and vision intersect

Retrieval and clustering

Search my Google photos for any photos or illustrations of a dog in a laboratory



Where language and vision intersect

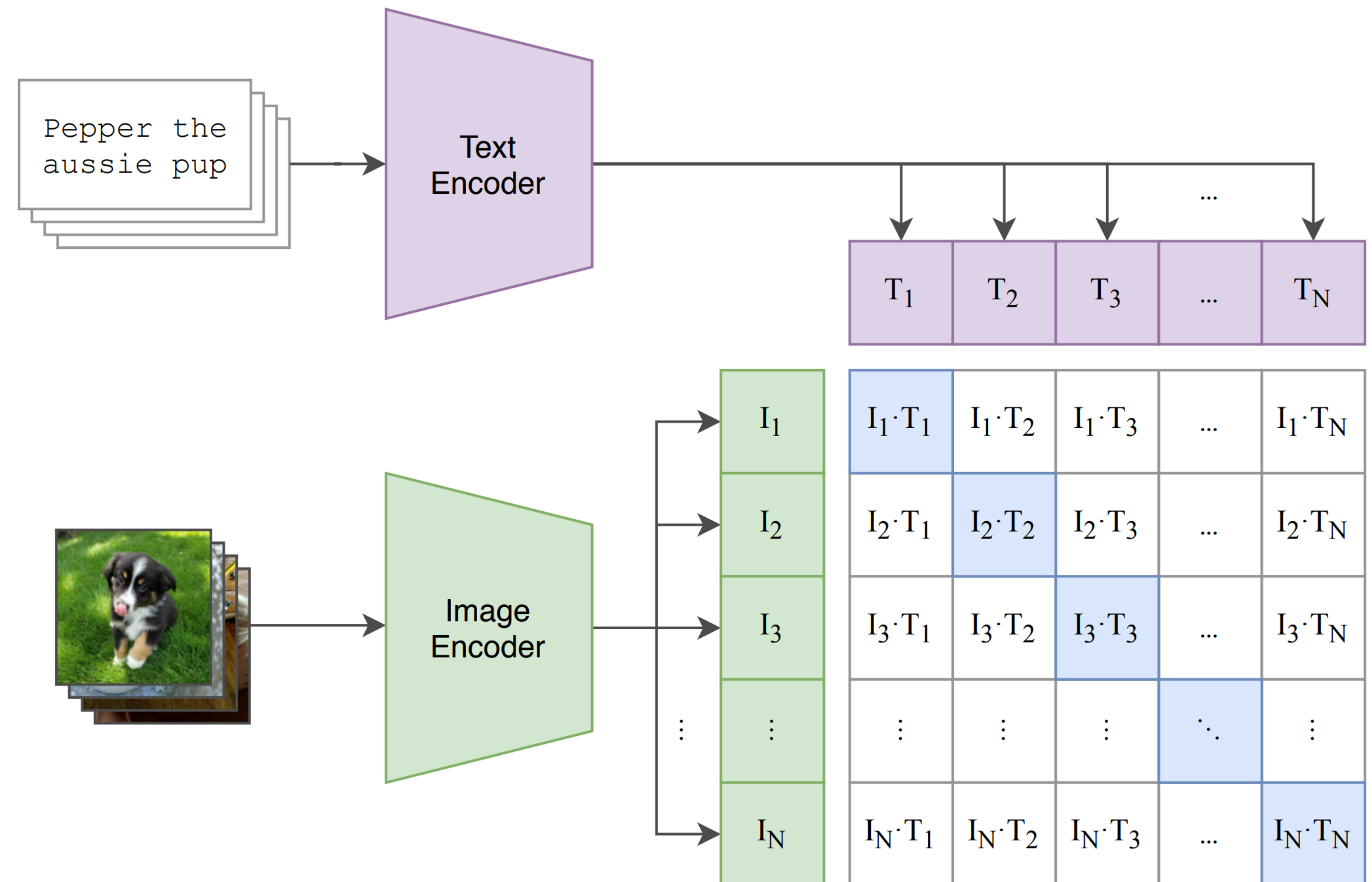
Possible objectives

- Models that map images and text to a joint embedding space
 - useful for retrieval, clustering, building classifiers
- Models that input an image and maybe text, and output text
 - useful for image captioning and question answering
- Models that input text and output an image
 - uses are self-explanatory
- Models that input a sequence of text and images and can continue the sequence
 - useful for multimodal chatbots and iterative image revision

Text+image models

CLIP - joint text+image embeddings circa 2021

- At train time, take a batch of n images along with their captions
- Training objective: for image caption pair $\langle i, c \rangle$, $\text{image_embedding}(i)$ should be close to $\text{caption_embedding}(c)$



Which of these is NOT a good use of CLIP?

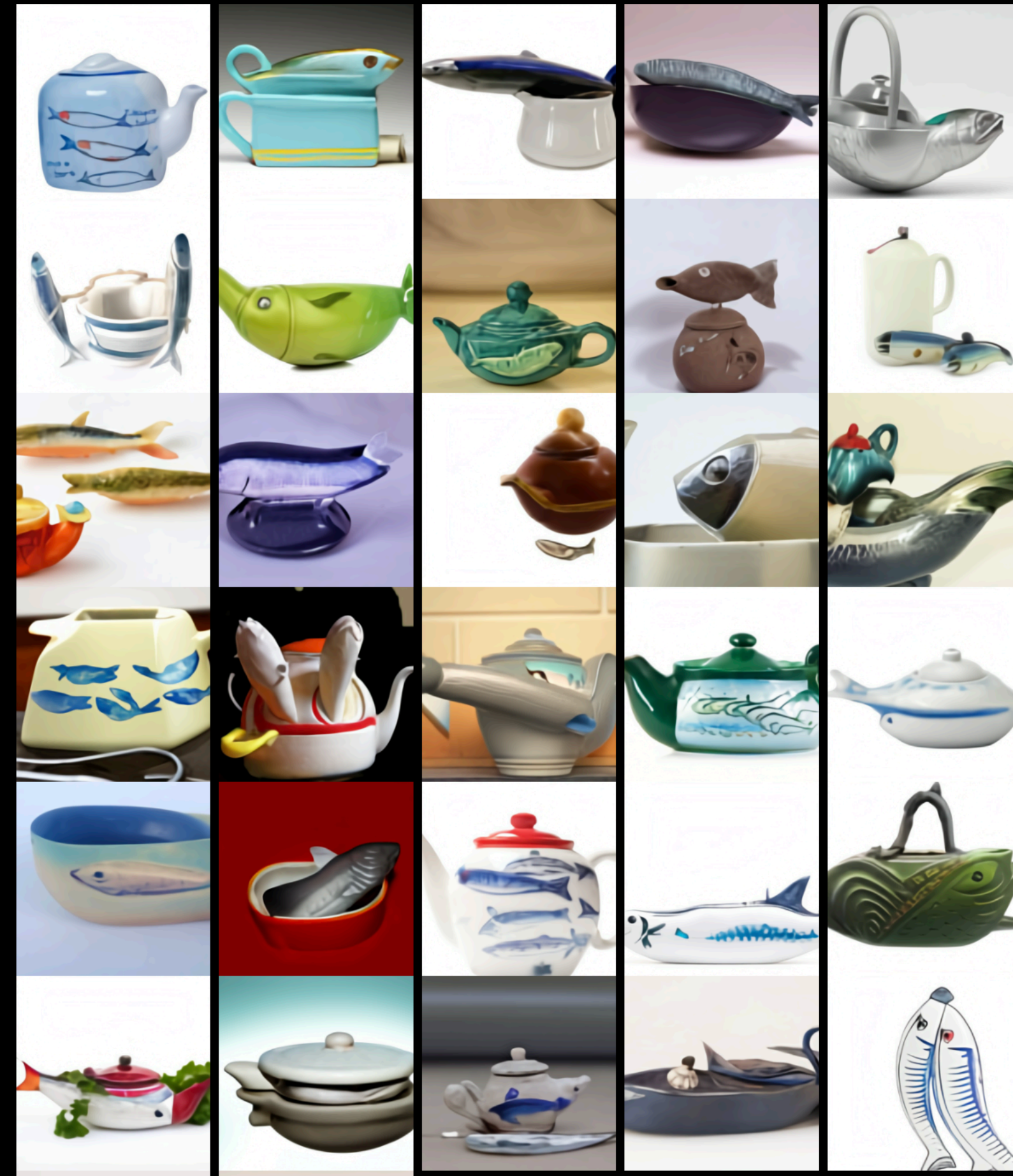
- A. Building an image classifier
- B. Finding images similar to a text query like "a dog in a lab"
- C. Placing a bounding box around the dog in a photo
- D. Scoring how well a generated image matches a text prompt
- E. Recommending visually similar products in an e-commerce store

DALL-E: image generation circa 2021

a teapot in the shape of a strawberry. a teapot imitating a strawberry.



a teapot in the shape of sardines. a teapot imitating sardines.



DALL-E

Main idea

- Goal: text and image tokens as a single stream of data
- Text is easy, but how do images get represented as token sequences?

DALL-E

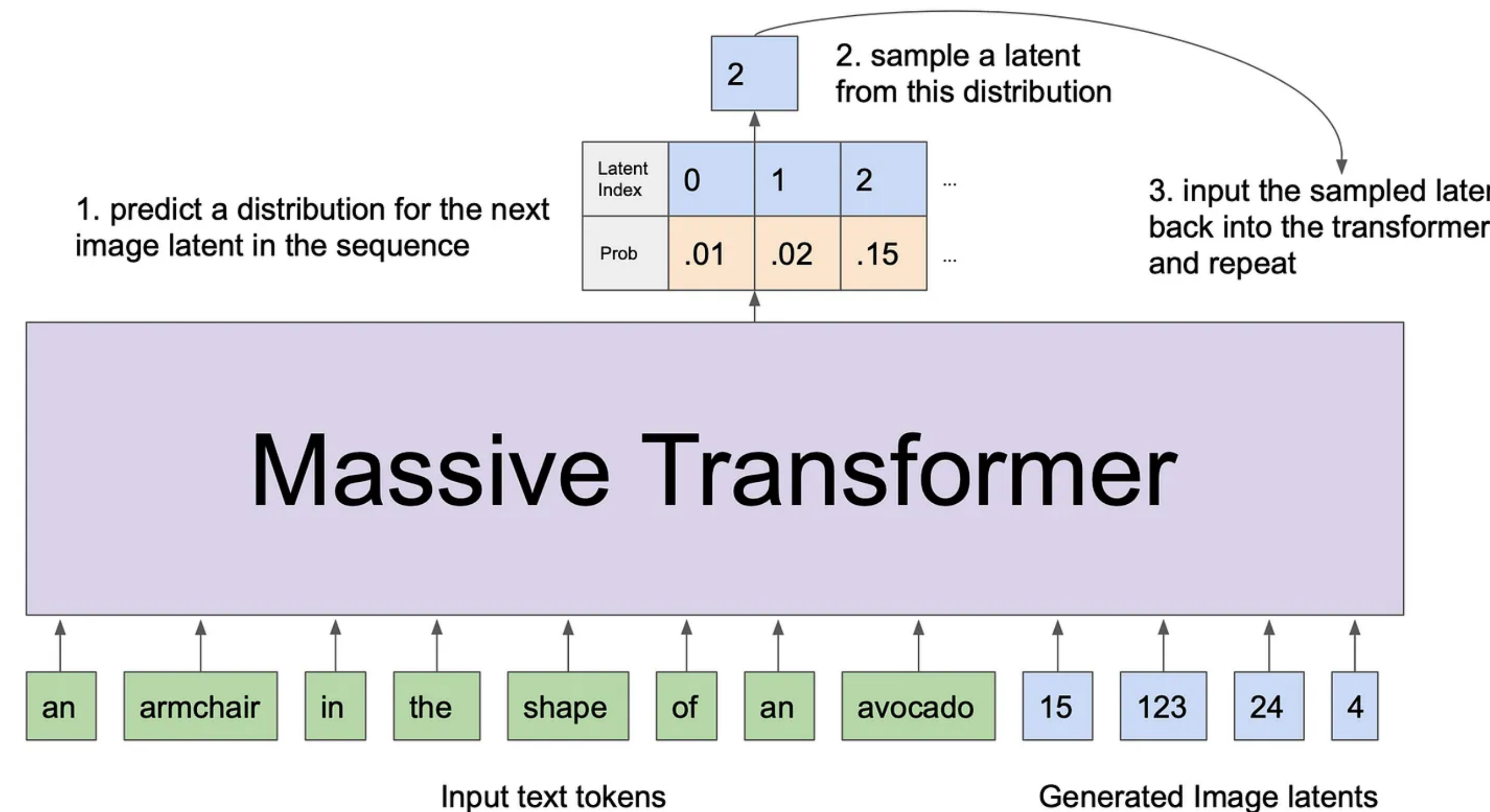
Main idea

- Goal: text and image tokens as a single stream of data
- Text is easy, but how do images get represented as token sequences?
- Start with 256x256 image (where each pixel has a 24 bit value).
- DALL-E's proposal: compress this down to 32x32 “image” where each pixel has only 8,192 possible values.

DALL-E

Main idea

- Goal: text and image tokens as a single stream of data
- Text is easy, but how do images get represented as token sequences?
- Start with 256x256 image (where each pixel has a 24 bit value).
- DALL-E's proposal: compress this down to 32x32 “image” where each pixel has only 8,192 possible values.



1024 image “tokens”

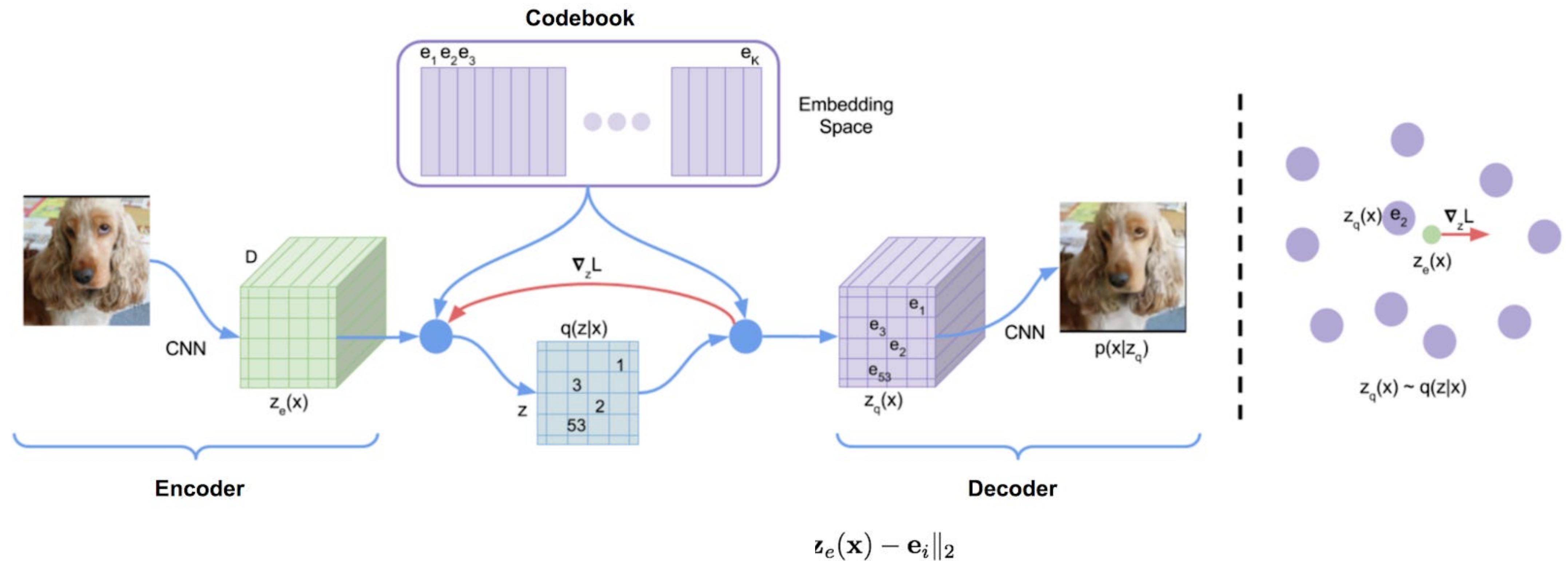
DALL-E

How do we get image tokens?

- Goal: compress an image down to a sequence of 1024 tokens.
 - Same as in language tokens, images tokens map to embeddings stored in an embedding matrix.
 - Dalle-E used an image vocabulary size (aka **codebook**) of 8,192 (which means 8k possible embeddings)
- Training stage 1:
 - Train a variational autoencoder that encodes images into sequence of embeddings that are nearby to embeddings in the codebook
 - no text involved in this stage
- Training stage 2:
 - Train an autoregressive transformer to model the joint distribution over the text and image tokens
 - Transformer sees as input 256 text tokens followed by the 1024 image tokens

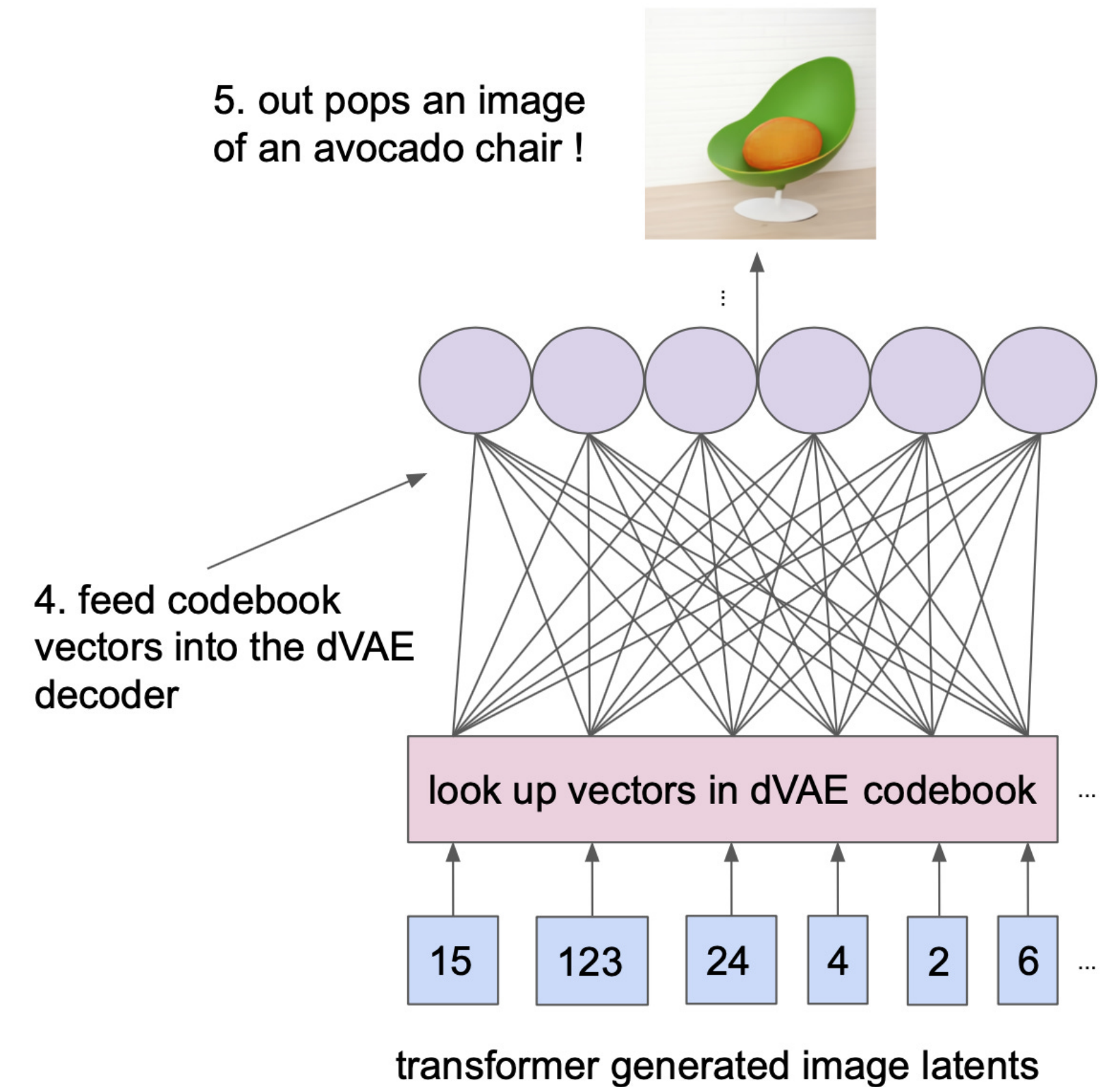
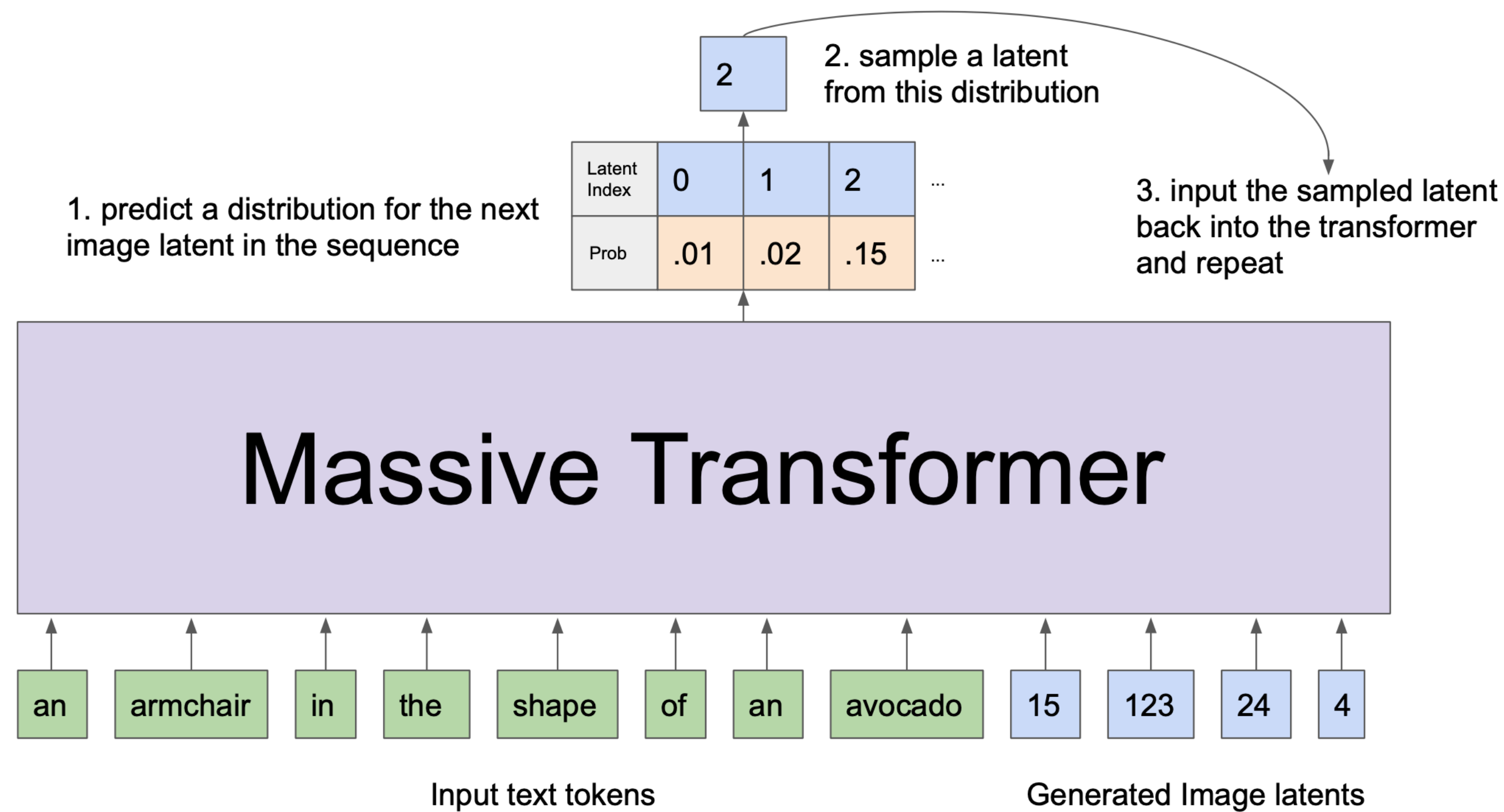
DALL-E

How do we get image tokens?



DALL-E

Inference time





Suppose your boss has tasked you with building an autoregressive transformer that generates images by predicting one embedding at a time, without using a codebook or any discretization step. What is the most fundamental problem with this approach?

- A. The model cannot use a softmax output head to define a probability distribution over a continuous space, making it unclear how to train or sample from the model at each step.
- B. The model will train too slowly because the image embeddings will have higher dimensionality than the text token embeddings.
- C. Without a codebook, the model has no way to encode the input image before generating, so it cannot condition on image prompts.
- D. The embedding sequences for the images will be too long for the Transformer architecture to efficiently handle.



Suppose your boss has tasked you with building an autoregressive transformer that generates images by predicting one embedding at a time, without using a codebook or any discretization step. What is the most fundamental problem with this approach?

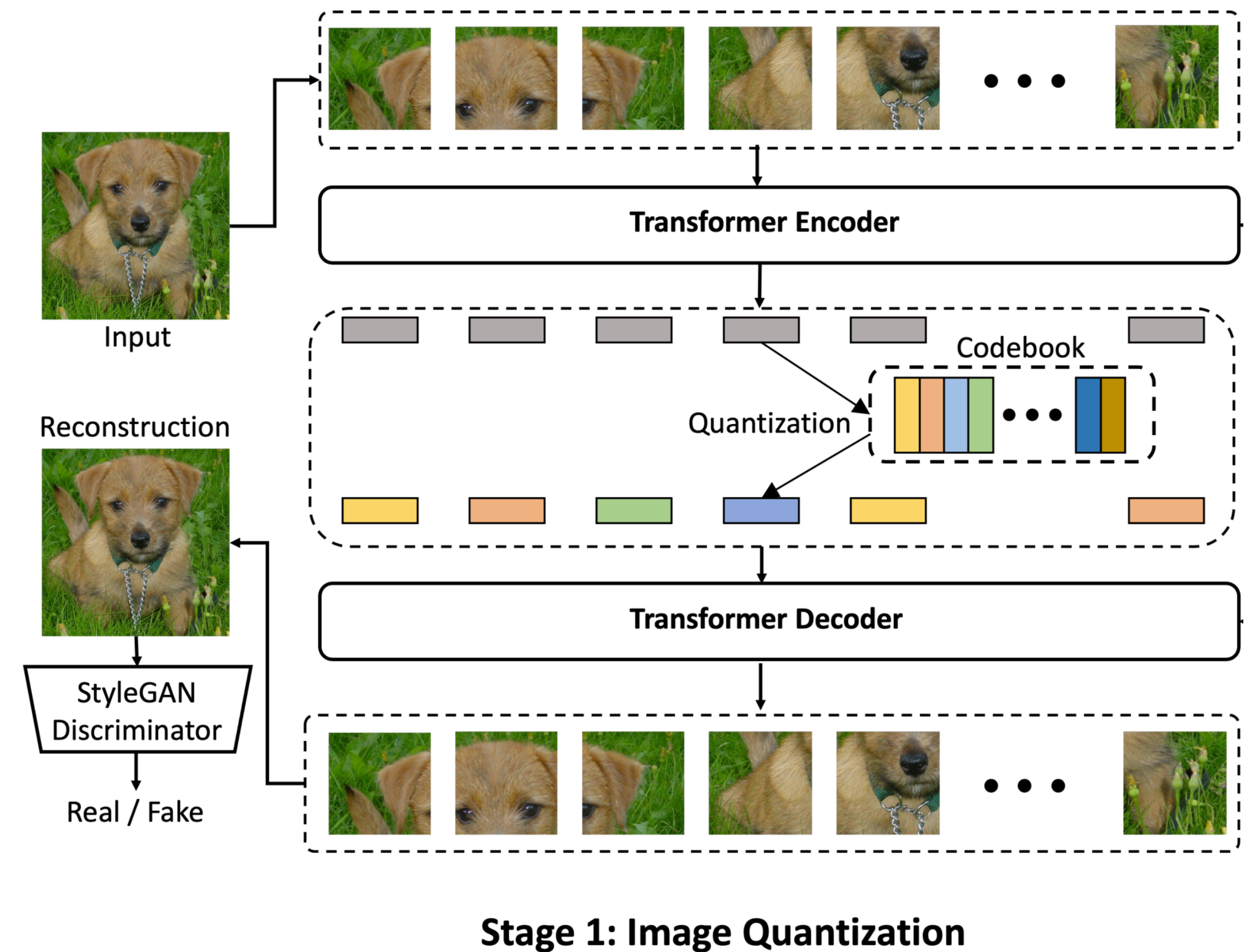
- A. The model cannot use a softmax output head to define a probability distribution over a continuous space, making it unclear how to train or sample from the model at each step.
- B. The model will train too slowly because the image embeddings will have higher dimensionality than the text token embeddings.
- C. Without a codebook, the model has no way to encode the input image before generating, so it cannot condition on image prompts.
- D. The embedding sequences for the images will be too long for the Transformer architecture to efficiently handle.

Other image-generation advancements

- DALL-E (2021): introduces idea of a transformer that autoregressively models the text and image tokens as a single stream of data

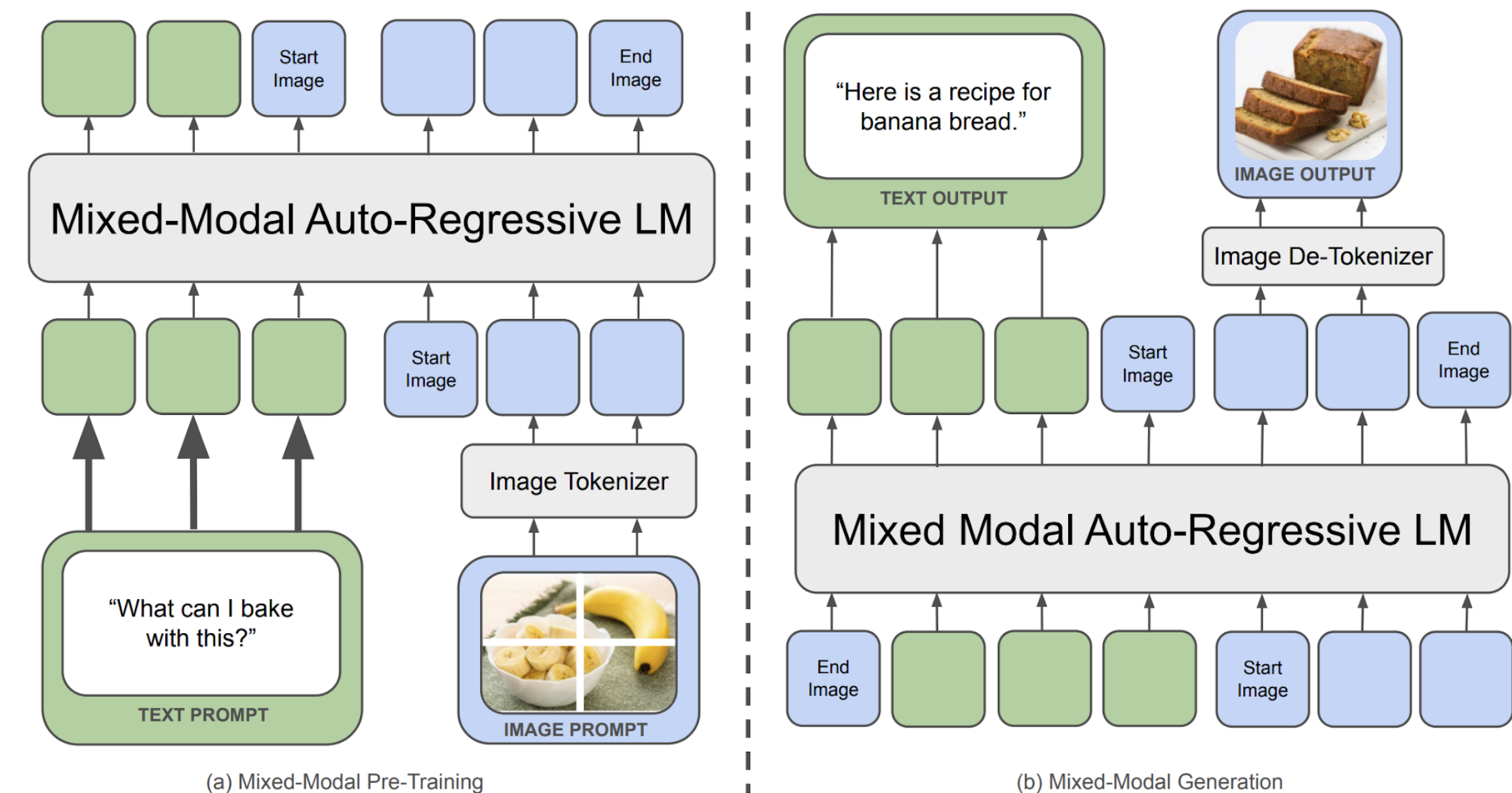
Other image-generation advancements

- DALL-E (2021): introduces idea of a transformer that autoregressively models the text and image tokens as a single stream of data
- Parti (2022): improved image tokenizer, ViT-VQGAN



Other image-generation advancements

- DALL-E (2021): introduces idea of a transformer that autoregressively models the text and image tokens as a single stream of data
- Parti (2022): improved image tokenizer, ViT-VQGAN
- CM₃ (2022): masked sequence modeling training objective on sequence containing text interspersed with images



Why are we trying to generate
images autoregressively?

Image diffusion models

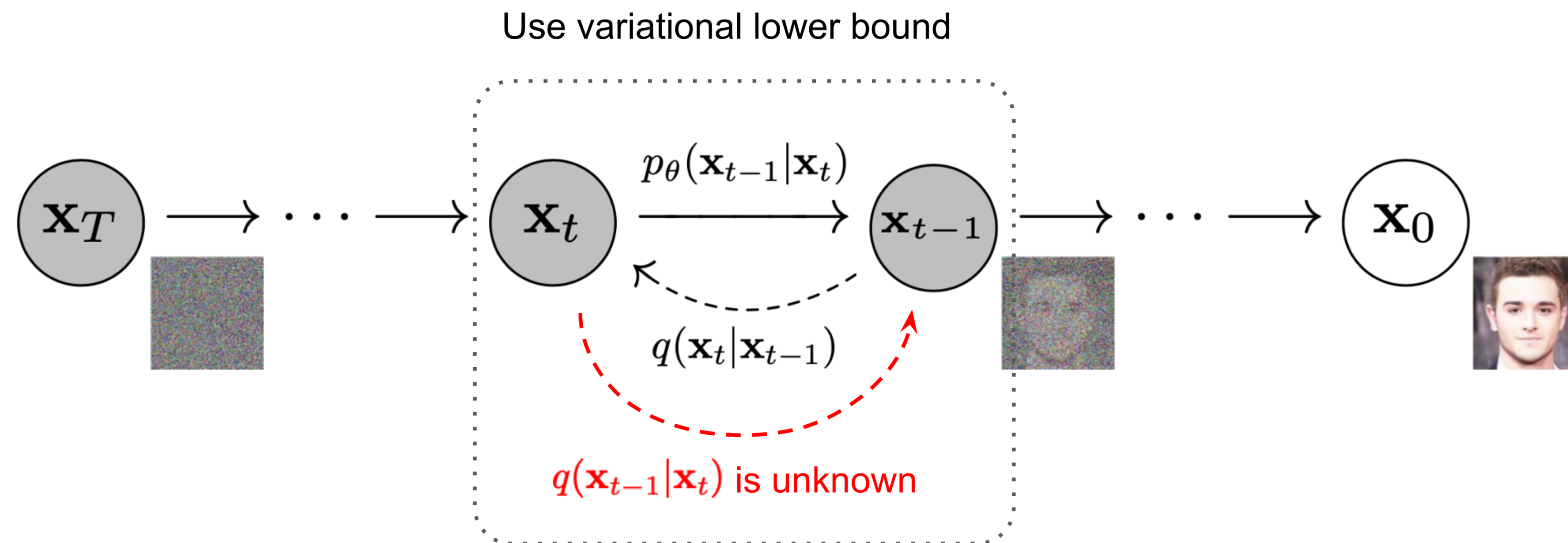


Image understanding advancements

- LLaVA (2023)
- PaLI (2022)

Image understanding advancements

LLaVA

- LLaVA was built by connecting a CLIP encoder with an chatbot-trained LLM (Vicuna), and fine-tuning end-to-end on instructional vision-language data

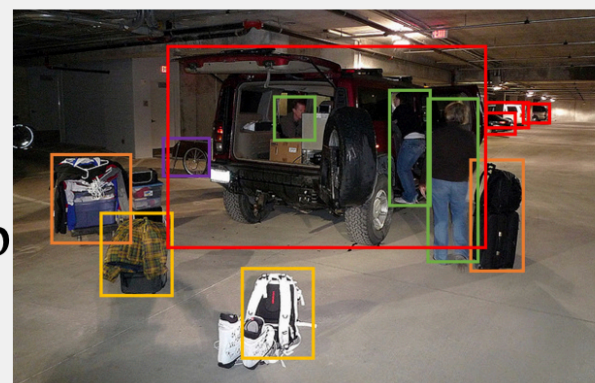
Image understanding advancements

LLaVA - training data

Existing datasets of image-text pairs

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>



Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Synthesized
conversational data

Image understanding advancements

LLaVA

- LLaVA was built by connecting a CLIP encoder with an chatbot-trained LLM (Vicuna), and fine-tuning end-to-end on instructional vision-language data
- Stage 1 training:
 - Keep the visual encoder and LLM weights frozen; train a projection matrix to align the CLIP image features with the LLM word embedding features.
- Stage 2 training:
 - Keep the CLIP encoder frozen, do instruction-tuning of the LLM using a standard auto-regressive training objective

Image understanding advancements

PaLI

- Instead of starting from decoder-only LM, starts with T5 (an encoder-decoder language model)
- More complex method for building joint image/language embedding space
- Trains on more traditional vision-question-answering datasets

Vision+language models today

Trends today

- A. Frontier VLMs are trained on multi-turn conversations containing both language tokens and image tokens.
 - At inference time, when users ask questions that involve images, the images are turned into sequences of tokens and included in the sequence.
 - At inference time, when users request images to be generated, this is typically treated as a tool call to a specialized image generation model.
- B. Increasing emphasis on training VLMs that use non-language representations to answer questions
- C. Image-language joint embedding models are still important and somewhat separate from VLMs

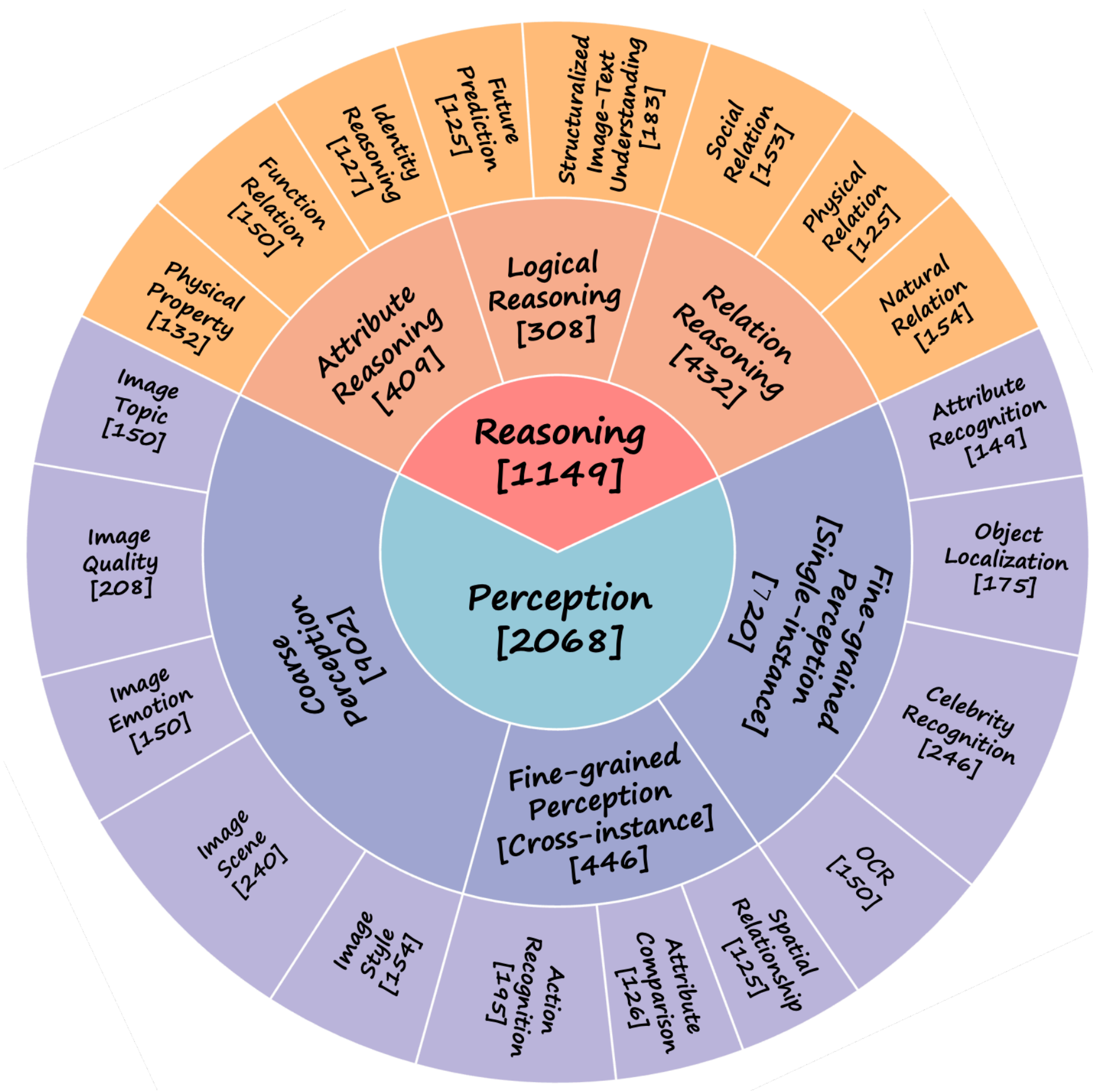
Open-weight vision-language models

Circa 2026

- Molmo
- GLM-4.6V
- Gemma 3
- Qwen3-VL

Evaluating vision language models

MMBench



Published Model					
	Method	Time	Parameters	Language Model	Vision Model
Overall					
1	SenseNova-V6-Pro Verified SenseTime	2025/05/05			
2	Qwen2.5-VL-72B Verified Alibaba	2025/02/02	73.4B	Qwen2.5-72B	QwenViT
3	InternVL2.5-78B-MPO Verified Shanghai AI Laboratory & Tsinghua University	2024/12/29	78B	Qwen-2.5-72B	InternViT-6B-v2.5
4	CongRong-v2.0 Verified CloudWalk	2025/05/20			
5	GLM-4v-Plus-20250111 Verified Zhipu AI	2025/01/25			

Evaluating vision language models

MMBench



Two magnets are placed as shown. Will these magnets attract or repel each other?

A: Repel.

B: Attract.



What will happen next?"

A: the car is gonna crash into the house

B: the car is gonna fly

C: the car is gonna drive backwards

D: both A,B, and C



Identify the emotion expressed in this image.

A: happiness

B: sadness

C: anger

D: loneliness

Evaluating vision language models

MMMU

- 11.5K multimodal questions from college exams, quizzes, and textbooks
- Intended to be a multimodal version of MMLU

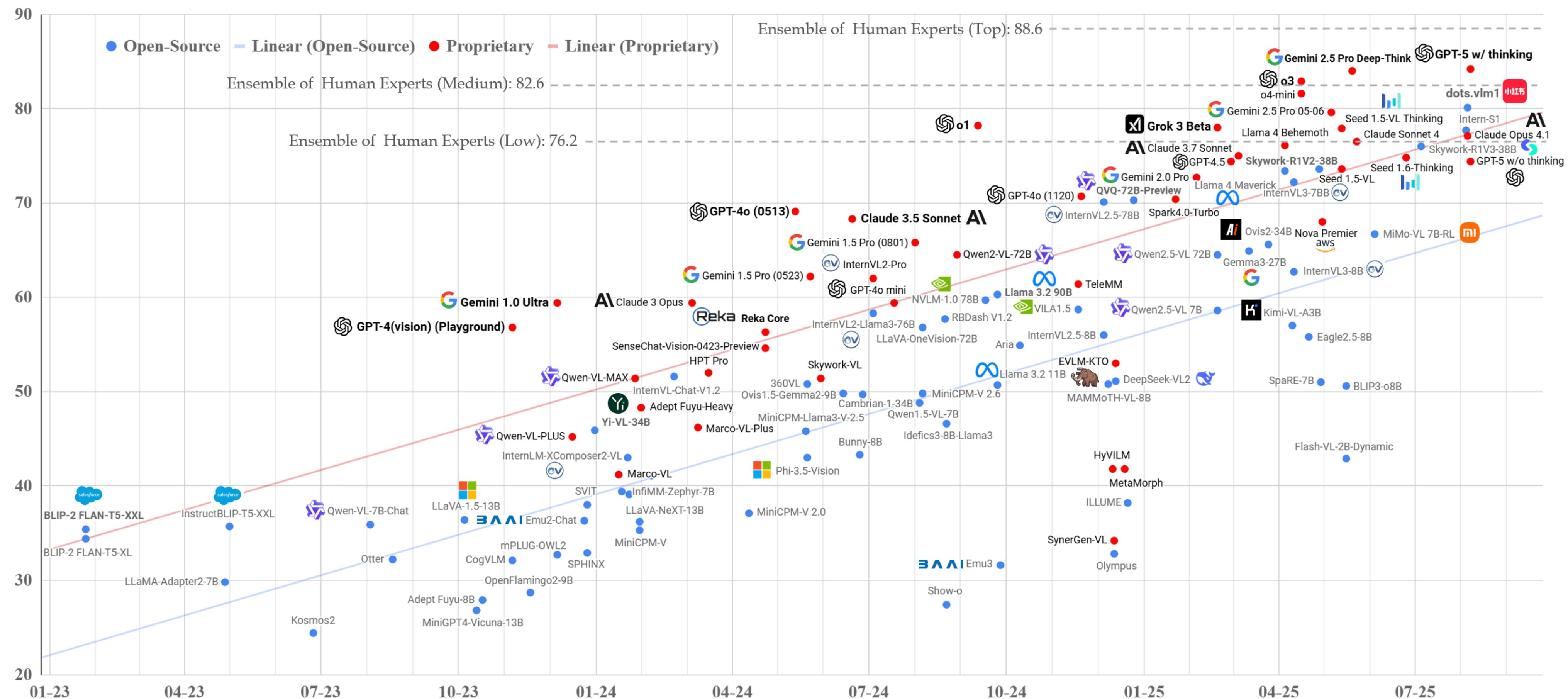
Evaluating vision language models

MMMU

Art & Design	Business	Science
Question: Among the following harmonic intervals, which one is constructed incorrectly? Options: (A) Major third <i><image 1></i> (B) Diminished fifth <i><image 2></i> <u>(C) Minor seventh <i><image 3></i></u> (D) Diminished sixth <i><image 4></i>	Question: ...The graph shown is compiled from data collected by Gallup <i><image 1></i>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82? Options: (A) 0 (B) 0.2142 <u>(C) 0.3571</u> (D) 0.5	Question: <i><image 1></i> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R. Options: <u>(A) $\int_0^{1.5} [f(x) - g(x)] dx$</u> (B) $\int_0^{1.5} [g(x) - f(x)] dx$ (C) $\int_0^2 [f(x) - g(x)] dx$ (D) $\int_0^2 [g(x) - x(x)] dx$
Subject: Music; Subfield: Music; Image Type: Sheet Music; Difficulty: Medium	Subject: Marketing; Subfield: Market Research; Image Type: Plots and Charts; Difficulty: Medium	Subject: Math; Subfield: Calculus; Image Type: Mathematical Notations; Difficulty: Easy
Health & Medicine	Humanities & Social Science	Tech & Engineering
Question: You are shown subtraction <i><image 1></i>, T2 weighted <i><image 2></i> and T1 weighted axial <i><image 3></i> from a screening breast MRI. What is the etiology of the finding in the left breast? Options: (A) Susceptibility artifact (B) Hematoma <u>(C) Fat necrosis</u> (D) Silicone granuloma	Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? <i><image 1></i> Option: (A) Oppressor (B) Imperialist <u>(C) Savior</u> (D) Isolationist	Question: Find the VCE for the circuit shown in <i><image 1></i>. Neglect VBE Answer: 3.75 Explanation: ...$I_E = [(V_{EE}) / (R_E)] = [(5 \text{ V}) / (4 \text{ k-ohm})] = 1.25 \text{ mA}$; $V_{CE} = V_{CC} - I_E R_L = 10 \text{ V} - (1.25 \text{ mA}) 5 \text{ k-ohm}$; $V_{CE} = 10 \text{ V} - 6.25 \text{ V} = 3.75 \text{ V}$
Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans: MRI, CT.; Difficulty: Hard	Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons; Difficulty: Easy	Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams; Difficulty: Hard

MMMU

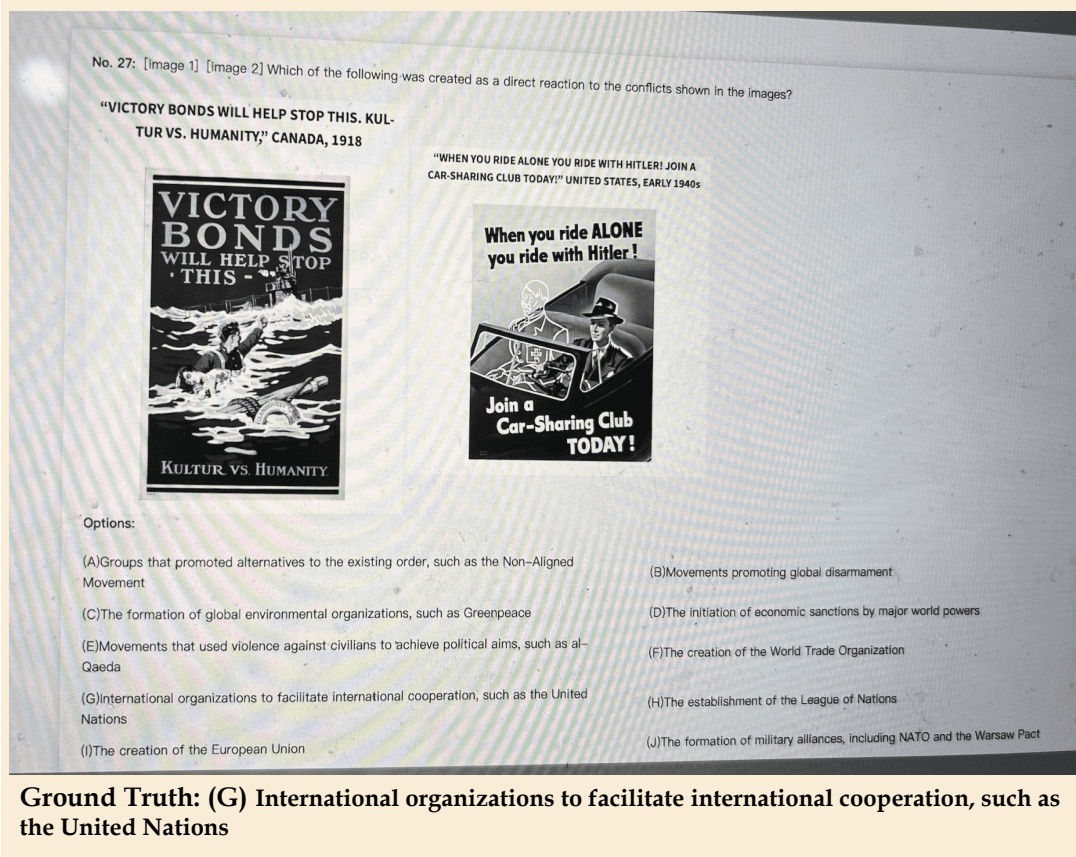
- Performance is saturated.



Evaluating vision language models

MMMU Pro

- Removed questions that were solvable without the images
- Increased number of candidate options from 4 to 10, decreasing the efficacy of random guessing
- Vision-only setting



Reset			MMMU-Pro
Name	Size	Date	Overall ↓
Human Expert (High)	-	2024-01-31	85.4
Human Expert (Medium)	-	2024-01-31	80.8
GPT-5 w/ thinking	-	2025-08-07	78.4*
o3	-	2025-04-16	76.4*
GPT-5.1	-	2025-11-13	76.0*
Human Expert (Low)	-	2024-01-31	73.0
dots.vlm1	37B	2025-08-05	70.1*
Qwen3-VL 235B-A22B	22B	2025-09-22	68.1*
Gemini 2.5 Pro 05-06	-	2025-05-06	68.0*
Seed 1.5-VL Thinking	20B	2025-05-13	67.6*
Seed 1.6-Thinking	20B	2025-06-25	66.4*
GLM-4.5V w/ Thinking	12B	2025-08-11	65.2*
GPT-5 w/o thinking	-	2025-08-07	62.7*
Seed 1.5-VL	20B	2025-05-13	59.9*
GLM-4.1V w/ Thinking	9B	2025-07-01	57.1*
Skywork-R1V3-38B	38B	2025-07-05	55.4*

Suppose you would like to build an app where you can take photos of your houseplant and get suggestions on how to take care of it better.



How would you go about picking a VLM to use?

- A. Take whichever model is currently at the top of the MMMU leaderboard
- B. Take whichever model is currently at the top of the MMBench leaderboard
- C. Manually annotate 20 houseplant images, and pick the model whose outputs best match your annotations
- D. Pick whichever model was released most recently
- E. Just grab the VLM API endpoint from your OpenAI/Anthropic/Google AI of choice
- F. something else?

Suppose you would like to build an app where you can take photos of your houseplant and get suggestions on how to take care of it better.



How would you go about deciding whether or not you need to collect a custom dataset and finetune?

Open-weight image generation models

Circa 2026

- Stable Diffusion
- SDXL

Evaluating image generation models

CLIPScore

- CLIPScore: cosine similarity between the CLIP embeddings of an image and a text prompt
-

VLMs using non-language representations

- Predicting bounding boxes or locations in the image
- Counting objects
- Identifying layout properties

Training data



Suppose you are tasked with training a VLM from scratch. How would you go about sourcing pre-training data?

LAION

curated listed of images found in CommonCrawl



Large-scale Artificial Intelligence Open Network

TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

LAION, as a non-profit organization, provides datasets, tools and models to liberate machine learning research. By doing so, we encourage open public education and a more environment-friendly use of resources by reusing existing datasets and models.

[Re-LAION 5B release \(30.08.2024\)](#)

LAION-400M

An open dataset containing 400 million English image-text pairs.

LAION-5B

A dataset consisting of 5.85 billion multilingual CLIP-filtered image-text pairs.

Clip H/14

The largest CLIP (Contrastive Language-Image Pre-training) vision transformer model.

LAION-Aesthetics

A subset of LAION-5B filtered by a model trained to score aesthetically pleasing images.

FAQ

Does LAION datasets respect copyright laws?

LAION datasets are simply indexes to the internet, i.e. lists of URLs to the original images together with the ALT texts found linked to those images. While we downloaded and calculated CLIP embeddings of the pictures to compute similarity scores between pictures and texts, we subsequently discarded all the photos. Any researcher using the datasets must reconstruct the images data by downloading the subset they are interested in. For this purpose, we suggest the [img2dataset](#) tool.

Links in your dataset show to my copyrighted data. I would like to have them removed.

LAION is a non-profit research organization that studies how learning algorithms work. Therefore, TDM exemptions are valid for LAION according to Art. 3 EU TDM exemption and §60d UrhG of German law, TDM exemption, for research. LAION is thus permitted to use any copyrighted material as data for conducting research on learning algorithms and foundation models as training outcome.

Kneschke v. LAION (Germany)

- German photographer Robert Kneschke sued non-profit LAION e.V. for scraping one of his watermarked stock photos to build the LAION-5B dataset
- [Sept 2024] The court ruled LAION did not breach copyright, finding the scraping qualified as lawful text and data mining (TDM) for scientific research under German law
 - Had LAION been developed by a for-profit company, courts may have ruled differently.
- [Dec 2025] Kneschke's appeal failed; higher level court confirmed LAION could rely on TDM copyright exceptions, and additionally ruled that Kneschke's opt-out was invalid because it was not machine-readable

Getty Images v. Stability AI (UK)

- Getty Images sued Stability AI, alleging that Stable Diffusion was trained on over 12 million Getty-owned images scraped without consent
 - Did Stability AI's distribution of its text-image diffusion model count as importing, selling, possessing, or distributing unlawful content?
- The court rejected secondary copyright infringement, finding that Stable Diffusion's model weights do not store or reproduce copies of Getty's images — they are the product of learned patterns, not reproductions

PixMo-Cap

- 717k images with very detailed captions
- web images across ~70 diverse topics (e.g., street signs, memes, food, drawings, websites, blurry photos, etc.)
- For each image, 1-2 annotators were asked to speak about the image for 60+ seconds
 - What is the image at first glance?
 - What are the objects and their counts?
 - What does the text say?
 - What are the positions of the objects?
 - What subtle details are noticeable?
 - What is in the background?
 - What is the style and color?
- Final high-quality image caption then created by prompting a language-only LLM to summarize the raw transcript(s), removing spoken artifacts and normalizing style

PixMo-Cap



This photograph, taken from a first-person perspective using a head-mounted camera, captures someone laying on a cream-colored couch covered with a gray blanket, **reading a white book**. Their right hand holds the book open with their thumb, while their left hand is curled around the spine. The person has bare arms and a light complexion. The scene is set in a spacious room with very tall ceilings that reach at least two stories high. The walls have a gradient, with the top portion a darker white and the bottom a brighter white. To the left, a large window with dark frames allows a soft blue light to filter in, illuminating the room. There are cream-colored window shades with a sheer white centerpiece hanging on the bottom half. Straight ahead, another large window is situated behind the couch. On the right side of the image, a brown couch cushion is visible. In front of the couch, a round coffee table adorned with a red tablecloth or centerpiece cloth, holds two potted plants – one with white roses and another with greenery. Additionally, there is a silver indoor clothes hanger with black shorts draped over it. The room is **bathed in daylight**, creating a welcoming and relaxed atmosphere.

PixMo-Cap



This detailed close-up image showcases a USPS receipt prominently positioned against what appears to be a white and beige marble countertop. The background on either side of the receipt features a gray and white pattern, adding a contrasting frame to the document. The receipt itself is printed in black ink and starts with the sentence: "We offer all the services you need to keep your business going." Directly below this statement is a partially visible barcode followed by a series of numbers: "001 00008 (022) Priority Mail." The receipt also features a long tracking number, "9405510200973700889896," which is circled in pen. The financial details indicate a total cost of "\$10.45," which is consistently listed as the subtotal, total, and the amount charged to a Visa card. The receipt displays the last four digits of the account number as "4929" and includes an APPR code. The document appears to have been signed, marking the transaction as complete. This comprehensive receipt captures the essential details of a USPS transaction for sending a package.

Detection of AI-Generated Content

Copyright & AI generations