

RBE474X/595-B01-ST: Deep Learning For Perception

Class 12: Advanced Generative Models++:
Multi-modal Generative Deep Learning

Prof. Wei Xiao



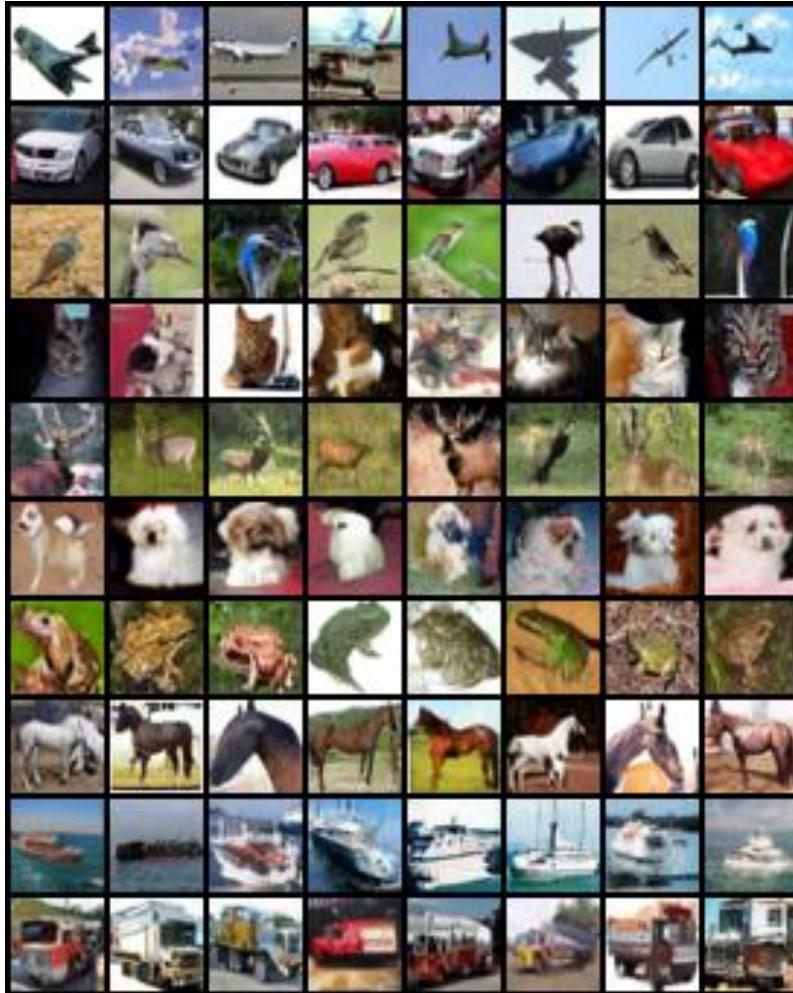
Recall Data Is The Key!

Harness The Power Of GenAI!

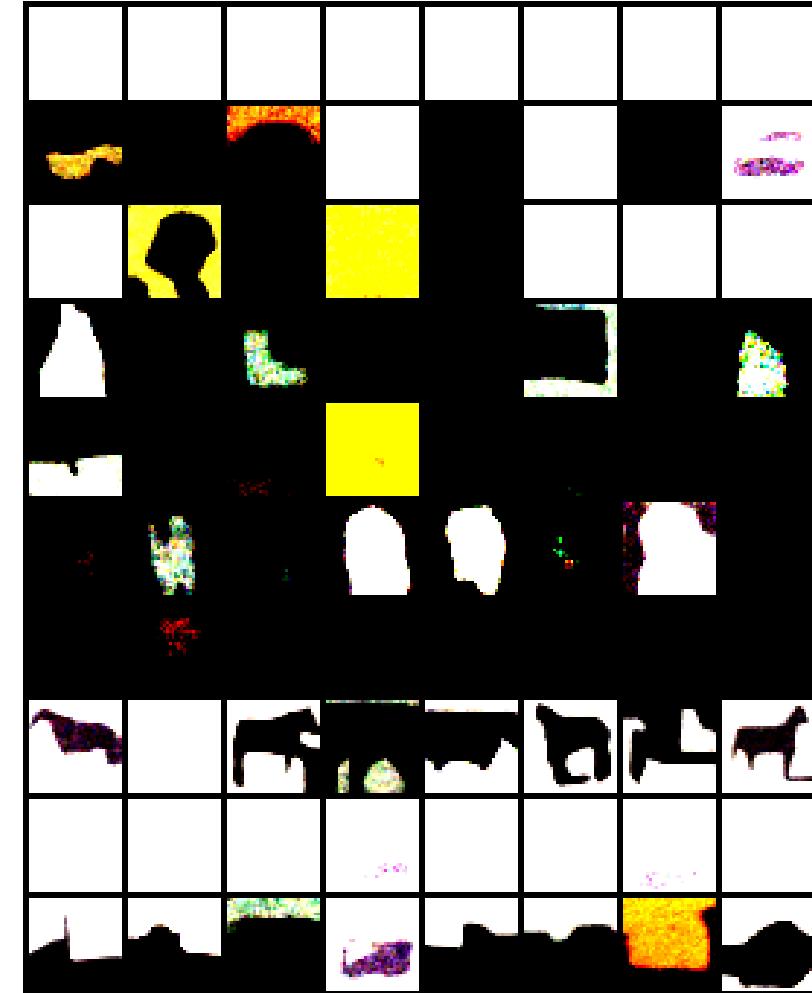
	Actual Data										Generated Data									
airplane																				
automobile																				
bird																				
cat																				
deer																				
dog																				
frog																				
horse																				
ship																				
truck																				

Generation Quality Varies!

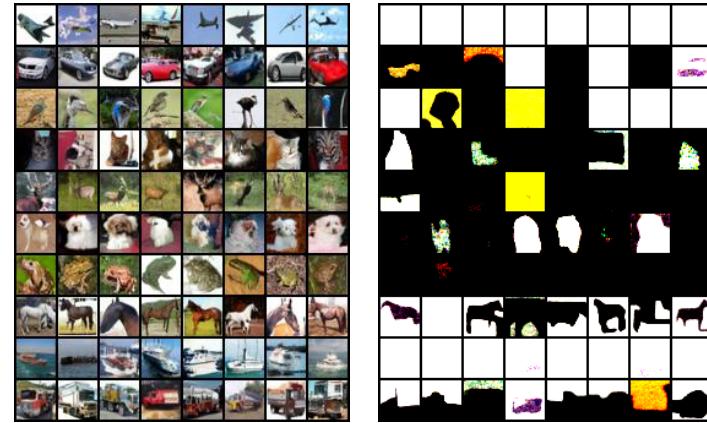
“Good” Hyperparameters



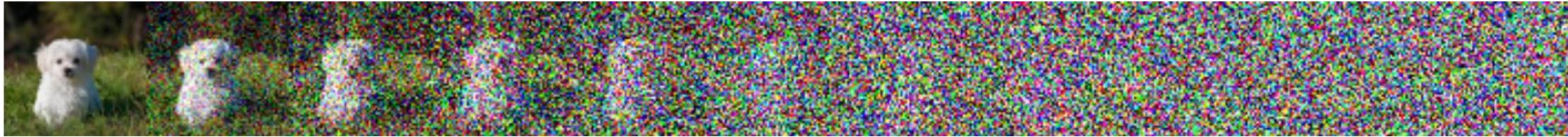
“Bad” Hyperparameters



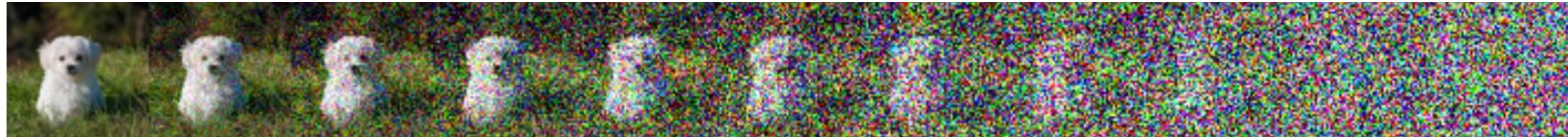
Generation Quality Varies!



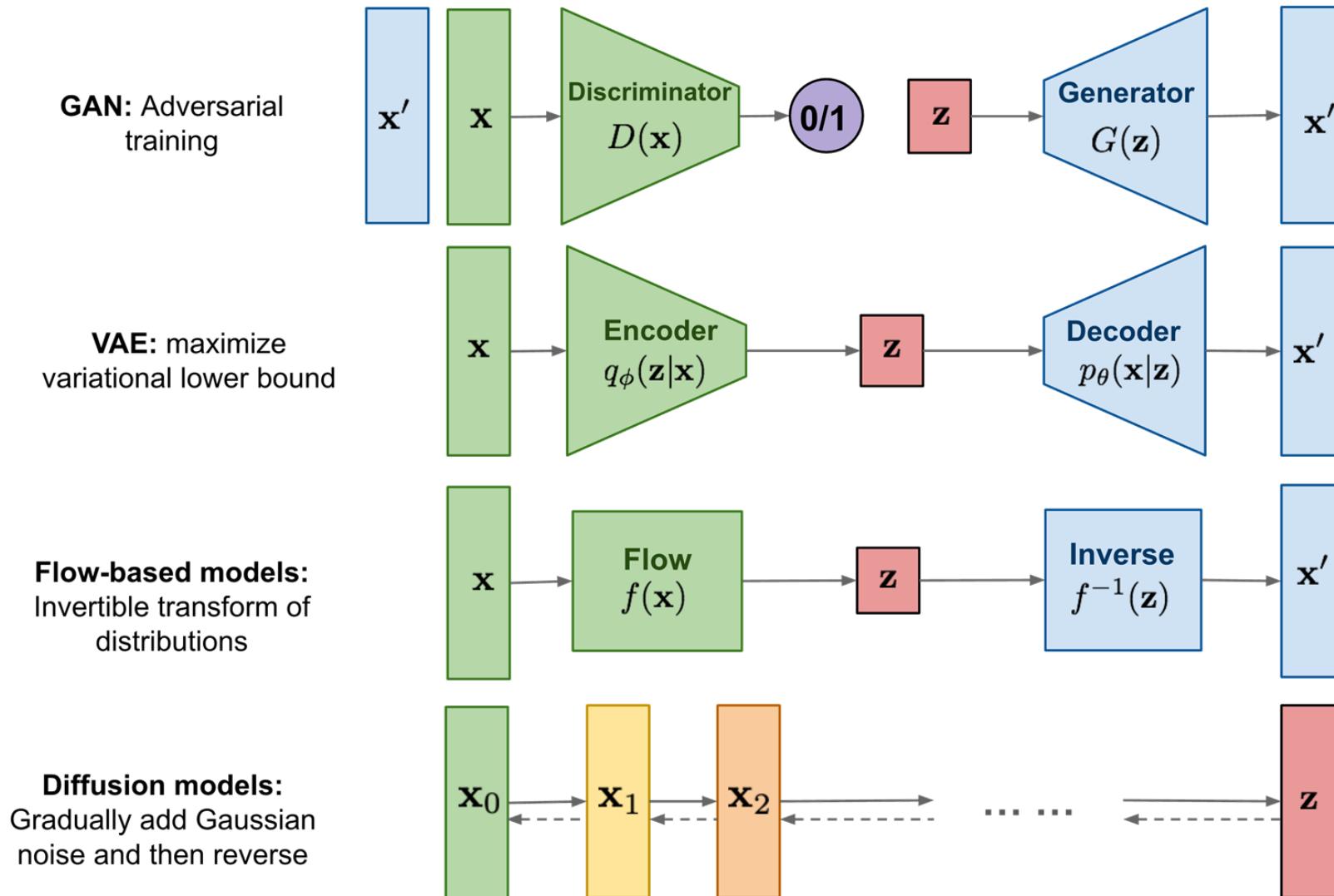
β_t follows a **Linear schedule**



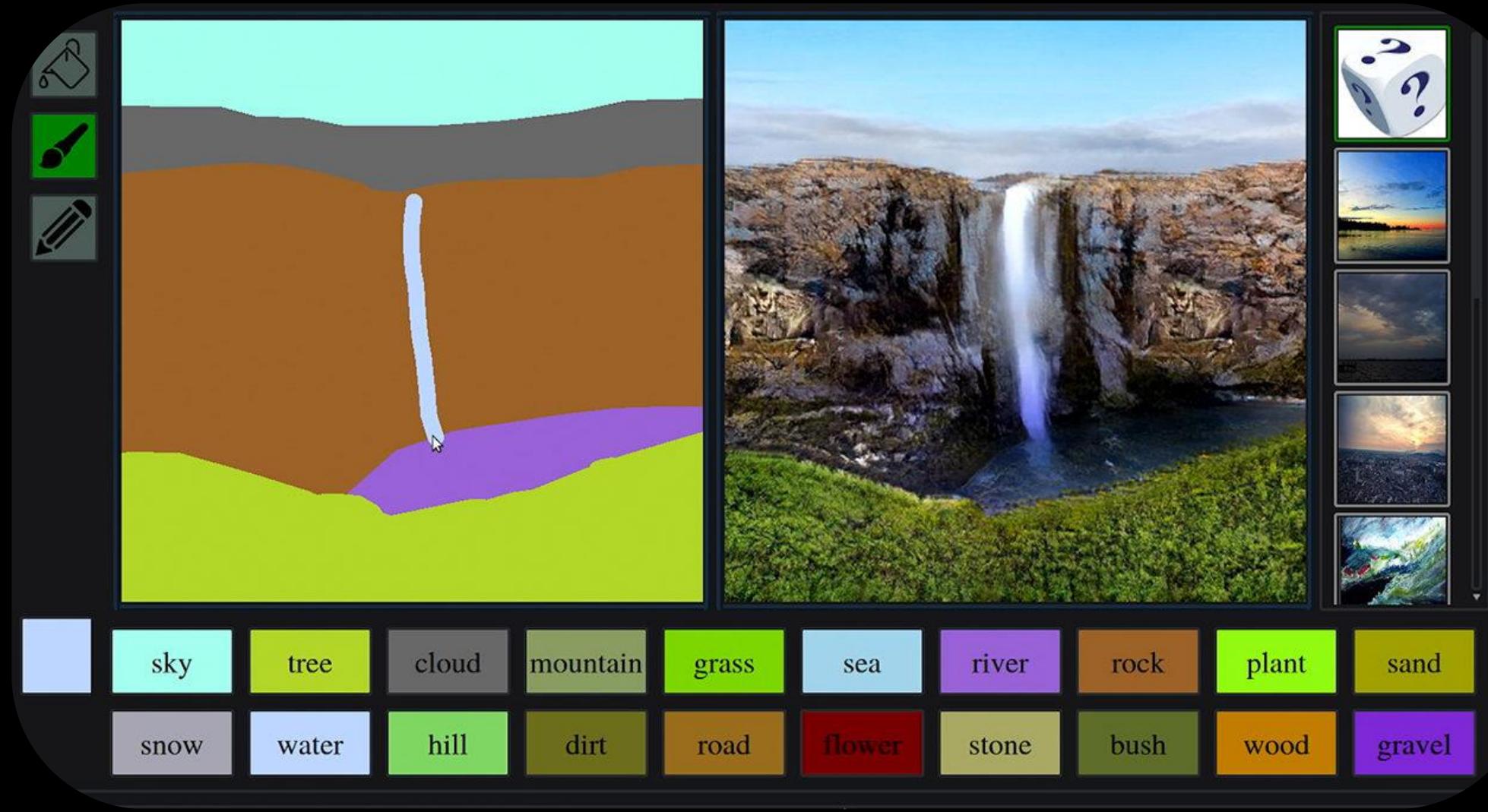
β_t follows a better **Cosine schedule**



GenAI Is Cool!



We've Seen This!



Generating Images Is Hard!

Not intuitive!

Everyone is not artistic!



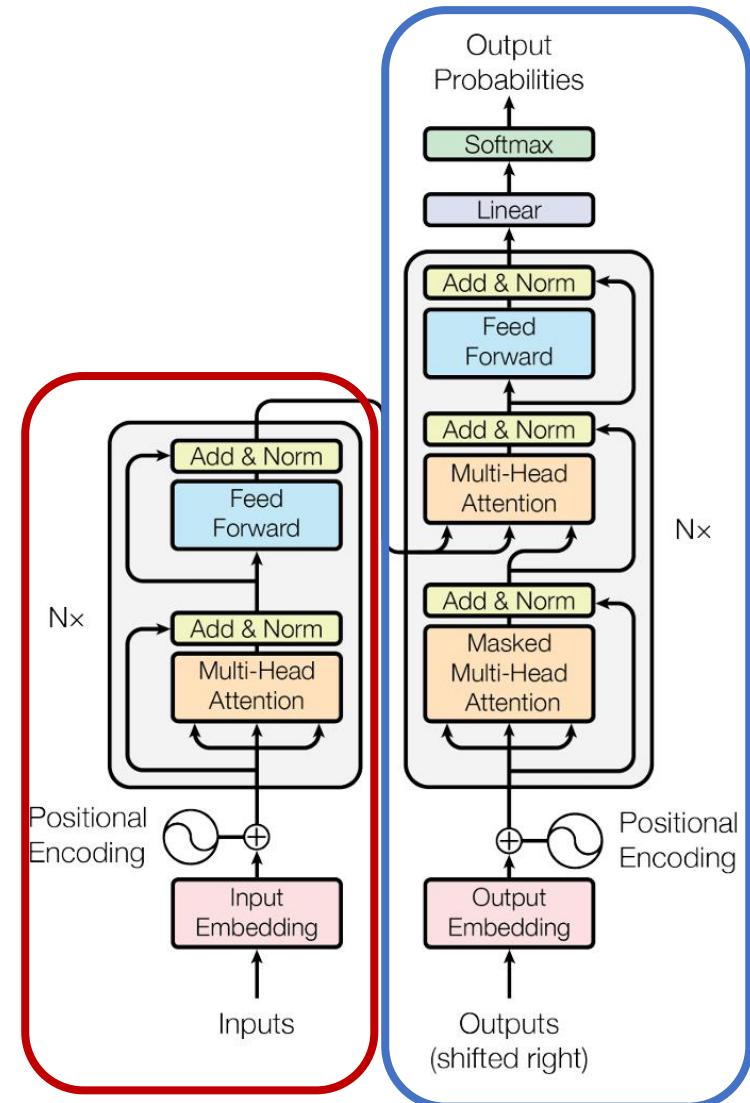
Recall Transformers

Encoder and Decoder

Learn compressed representation Decompress the representation

Transformers can be

- **Encoder only:** Used for learning representations BERT
- **Decoder only:** Used for generation GPT-3
- Encoder-Decoder: Used for sequence-to-sequence Language Translation



Recall Transformers

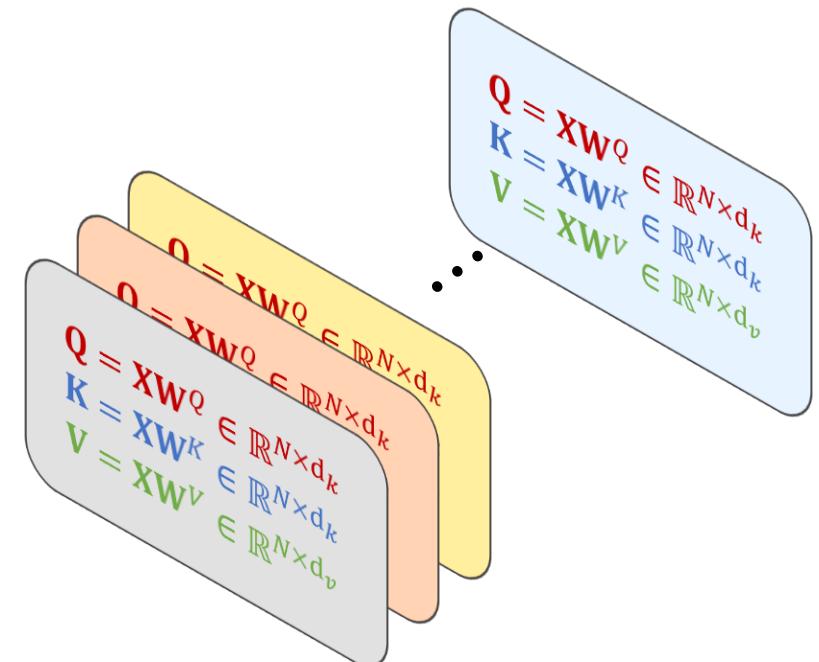
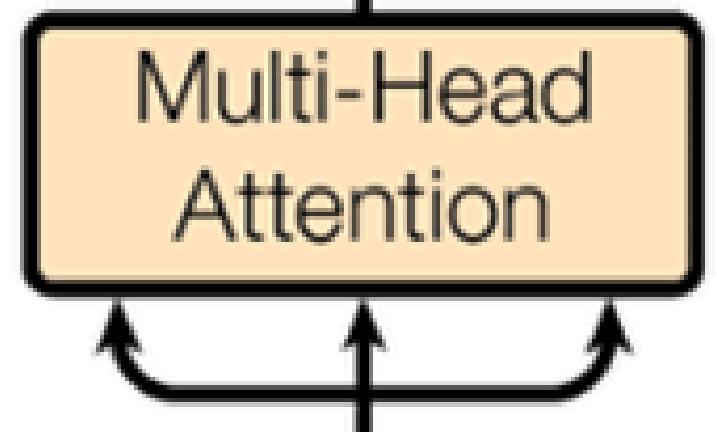
Multi-Head Attention

For $h = 1 \dots H$: **In parallel!**

$$\begin{aligned} Q &= XW^Q \in \mathbb{R}^{N \times d_k} \\ K &= XW^K \in \mathbb{R}^{N \times d_k} \\ V &= XW^V \in \mathbb{R}^{N \times d_v} \end{aligned}$$

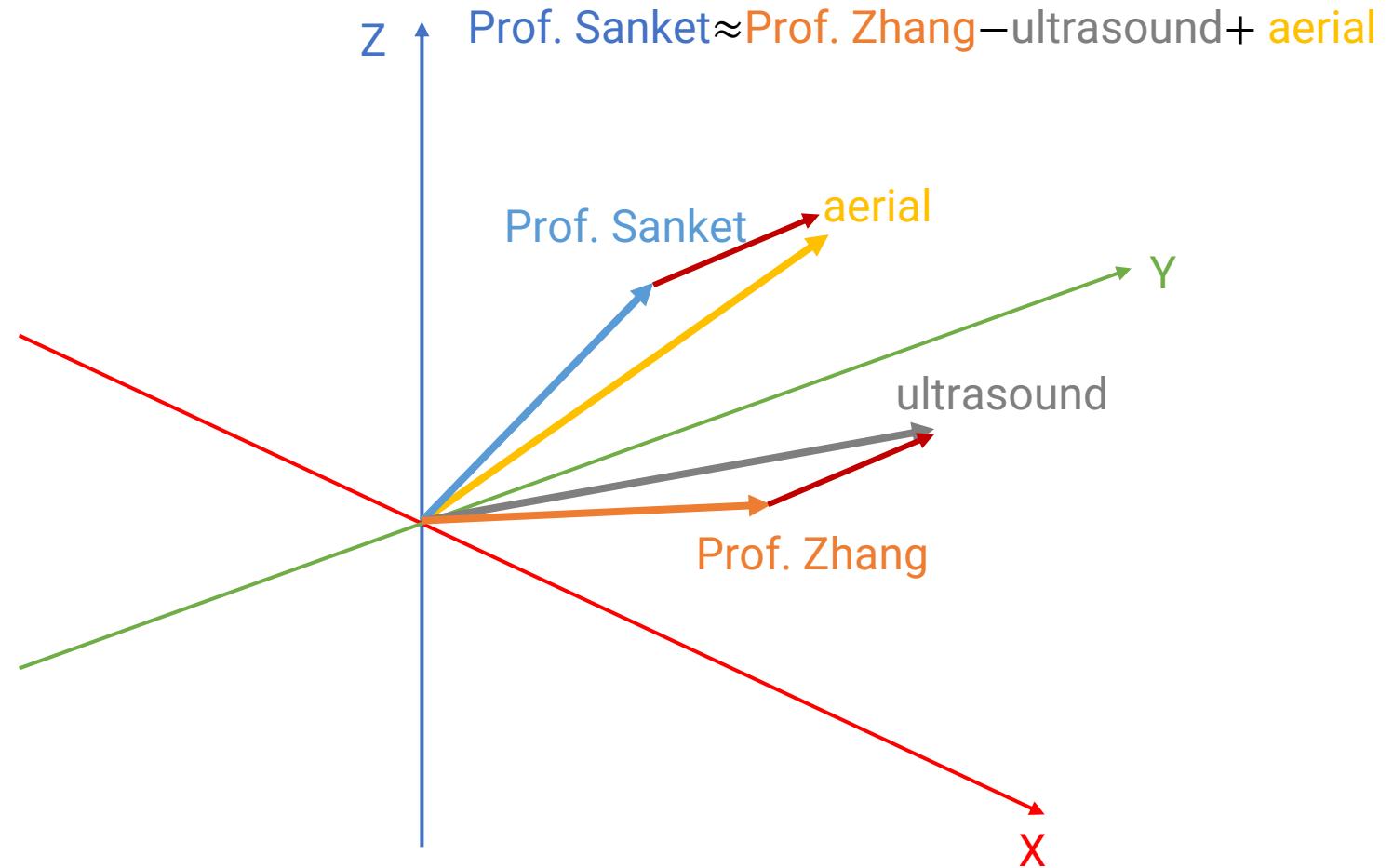
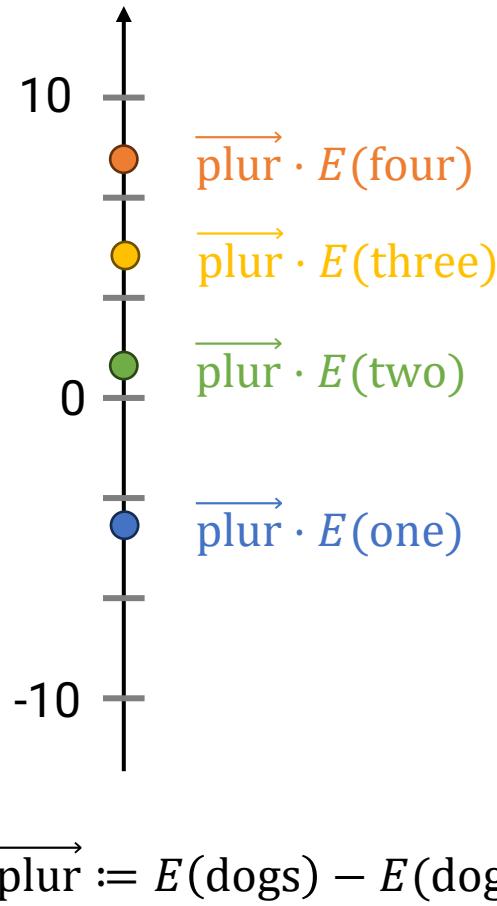
$$\text{head}_h = \text{Softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^T}{\sqrt{d_k}}\right) \mathbf{V}_h$$

$$\text{MultiHead}(\mathbf{X}) = \text{Conact}(\text{head}_1 \dots \text{head}_h) \mathbf{W}^O$$

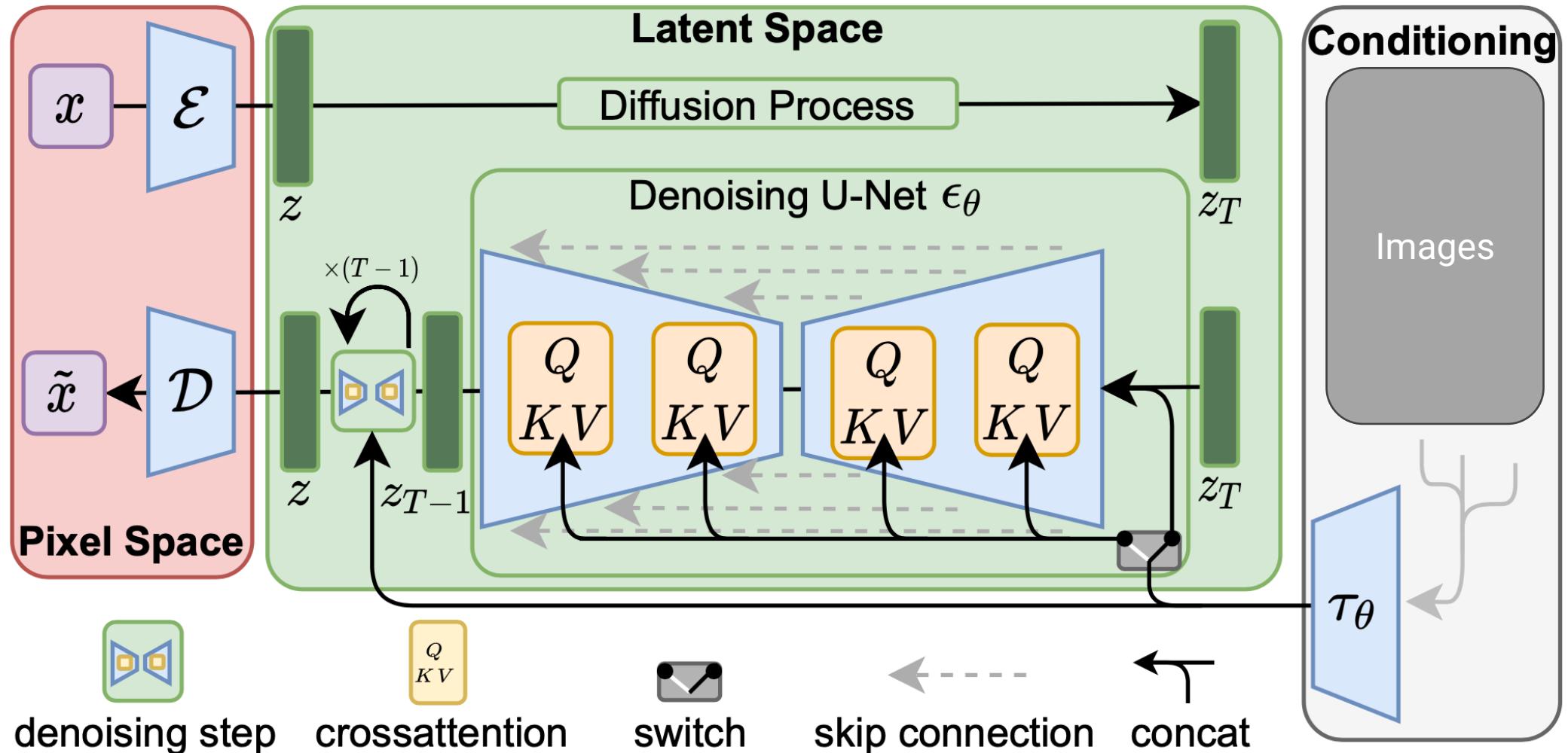


Recall Embeddings

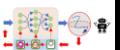
AKA Latent Space



Recall Diffusion Models



Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.





Describe This Image!

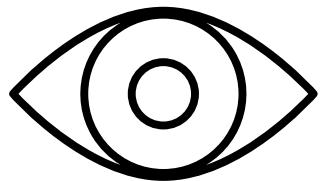


What We Have Been Doing!

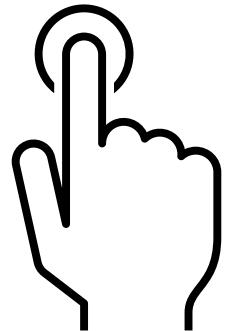
We Don't Work Like This!

We Convey Information In Many Ways!

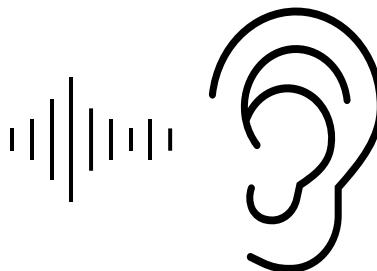
Perceive Too



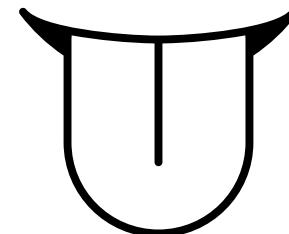
Vision



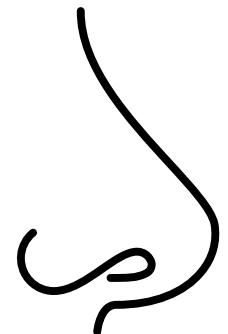
Touch



Sound



Taste

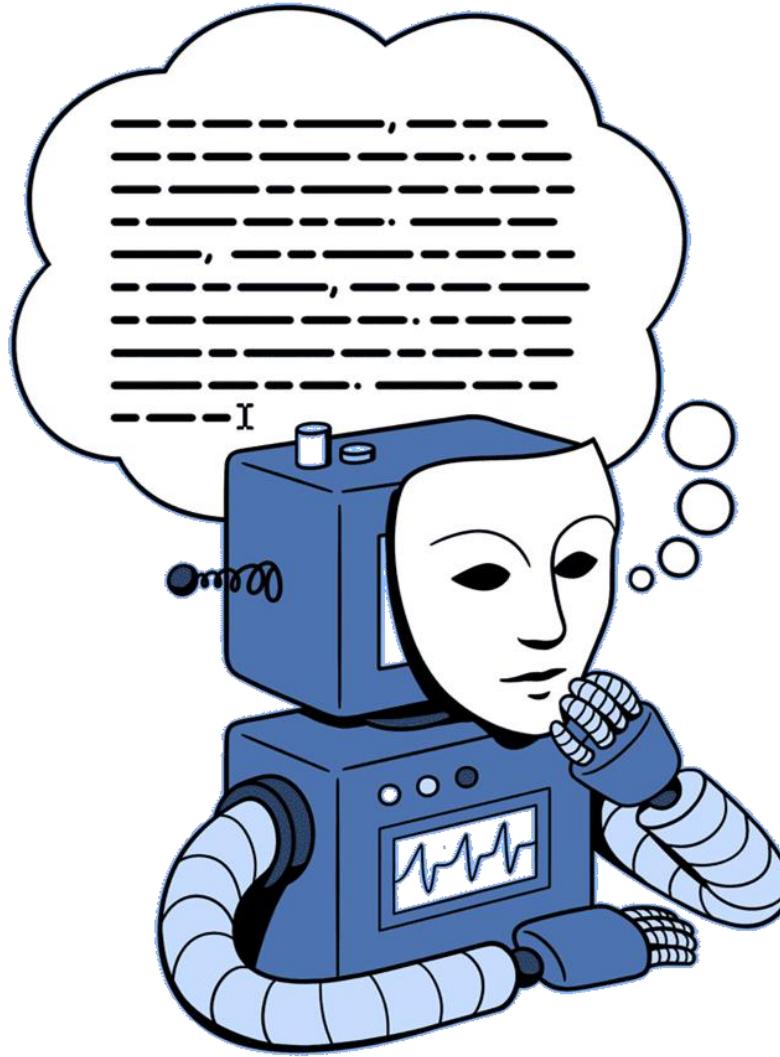


Smell



Proprioception

Large Language Models



- Summarization
- Translation
- Prediction
- Generation

We Learn And Use Concepts Across Modalities!



A scale or balance is a device used to measure weight or mass. These are also known as mass scales, weight scales, mass balances, massometers, and weight balances.

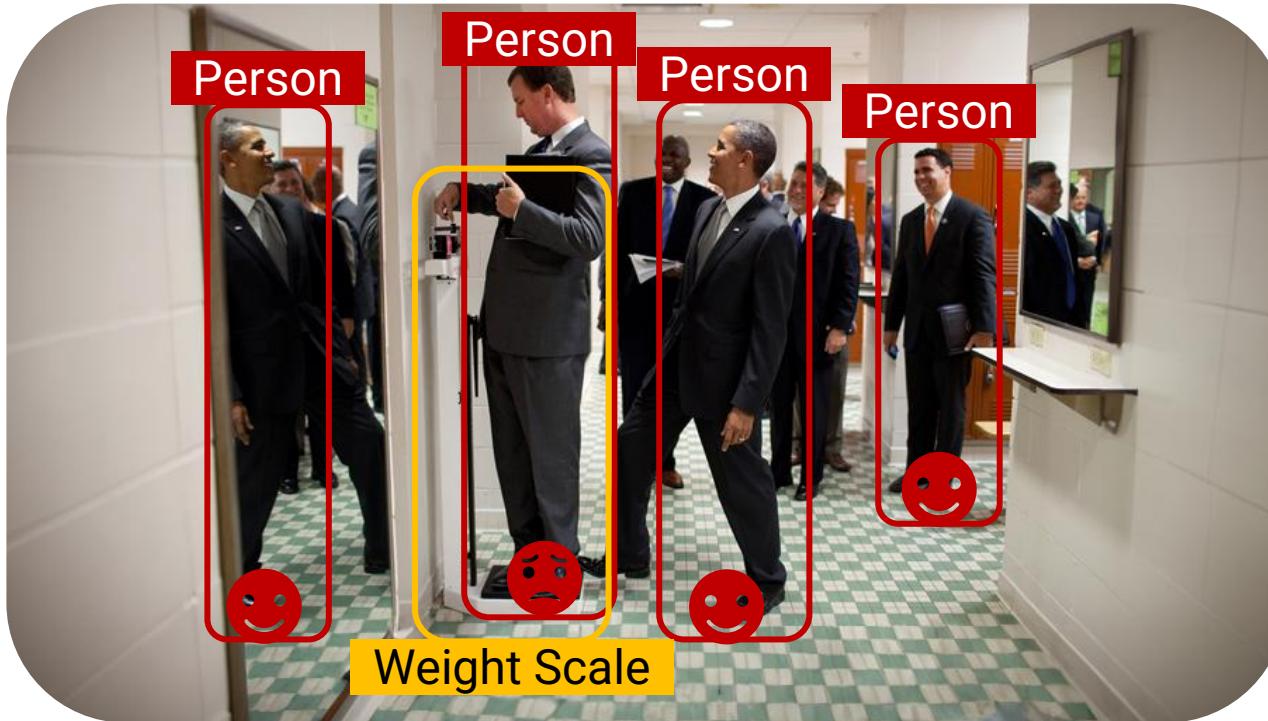
The traditional scale consists of two plates or bowls suspended at equal distances from a fulcrum. One plate holds an object of unknown mass (or weight), while objects of known mass or weight, called weights, are added to the other plate until mechanical equilibrium is achieved and the plates level off, which happens when the masses on the two plates are equal.

For some people, gaining weight can cause further stress. Understanding the relationship between stress and weight gain can help people recognize and manage these aspects of their health.

Prank: a practical joke or mischievous act.

Psychologists have studied pranks for years. Humor, in general, is good for us. Neuropsychology research has shown that laughing improves well-being.

What Did We Learn?



A scale or balance is a device used to measure weight or mass. These are also known as mass scales, weight scales, mass balances, massometers, and weight balances.

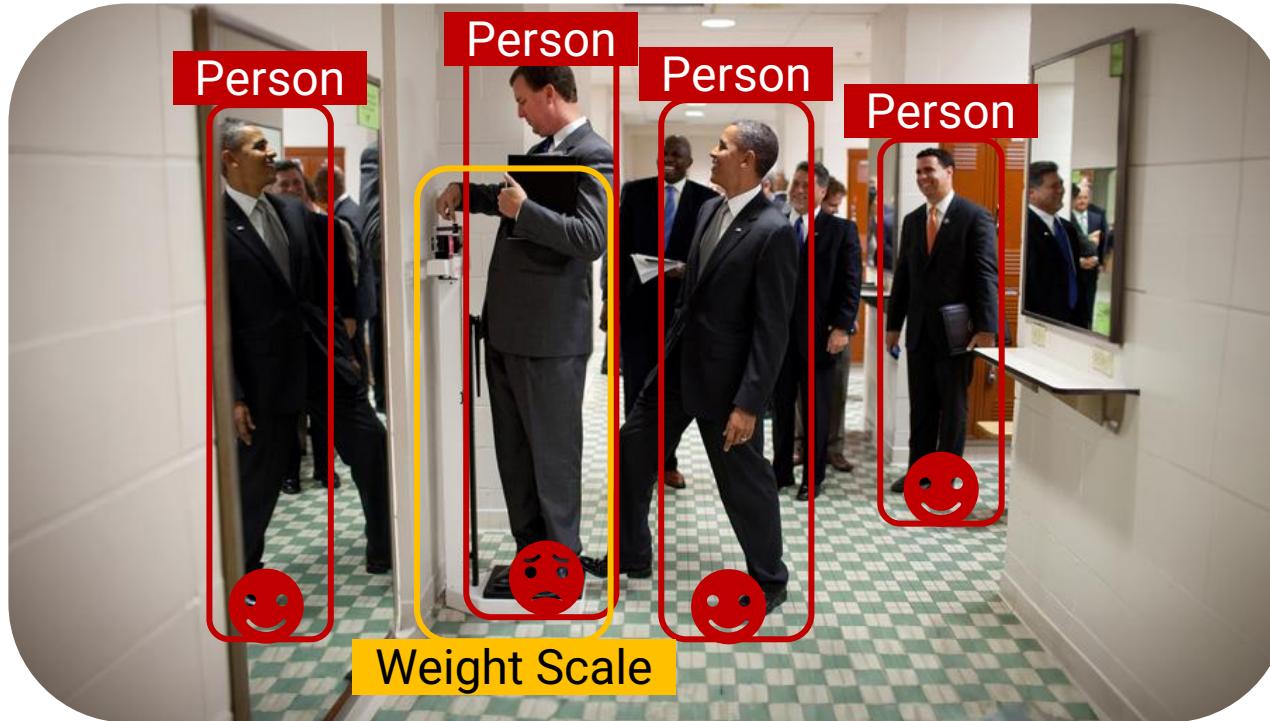
The traditional scale consists of two plates or bowls suspended at equal distances from a fulcrum. One plate holds an object of unknown mass (or weight), while objects of known mass or weight, called weights, are added to the other plate until mechanical equilibrium is achieved and the plates level off, which happens when the masses on the two plates are equal.

For some people, gaining weight can cause further stress. Understanding the relationship between stress and weight gain can help people recognize and manage these aspects of their health.

Prank: a practical joke or mischievous act.

Psychologists have studied pranks for years. Humor, in general, is good for us. Neuropsychology research has shown that laughing improves well-being.

What Did We Learn?



A person is checking their weight and another person is playing a prank by placing their foot on the weighing scale.

A scale or balance is a device used to measure weight or mass. These are also known as mass scales, weight scales, mass balances, massometers, and weight balances.

The traditional scale consists of two plates or bowls suspended at equal distances from a fulcrum. One plate holds an object of unknown mass (or weight), while objects of known mass or weight, called weights, are added to the other plate until mechanical equilibrium is achieved and the plates level off, which happens when the masses on the two plates are equal.

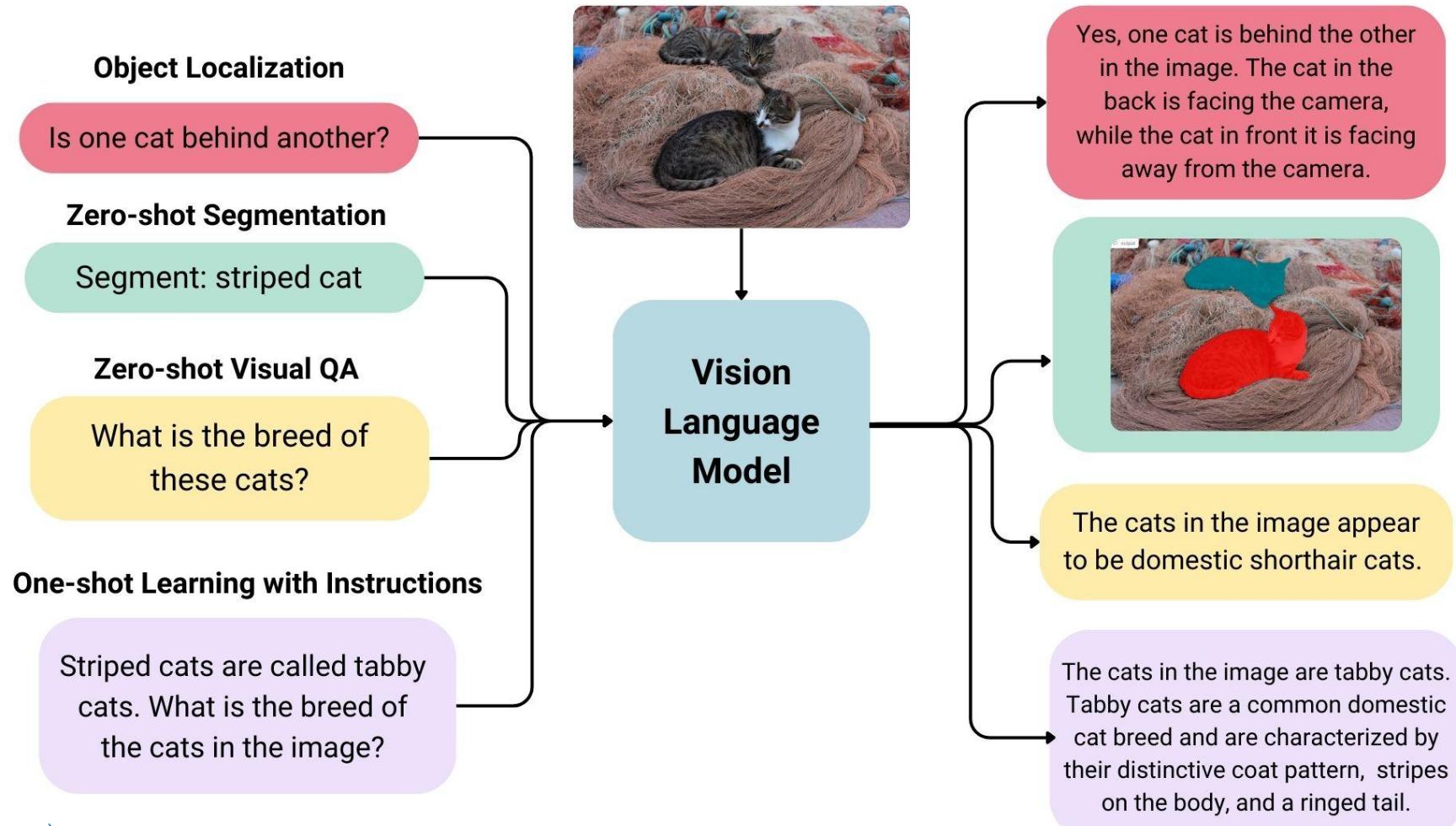
For some people, gaining weight can cause further stress. Understanding the relationship between stress and weight gain can help people recognize and manage these aspects of their health.

Prank: a practical joke or mischievous act.

Psychologists have studied pranks for years. Humor, in general, is good for us. Neuropsychology research has shown that laughing improves well-being.

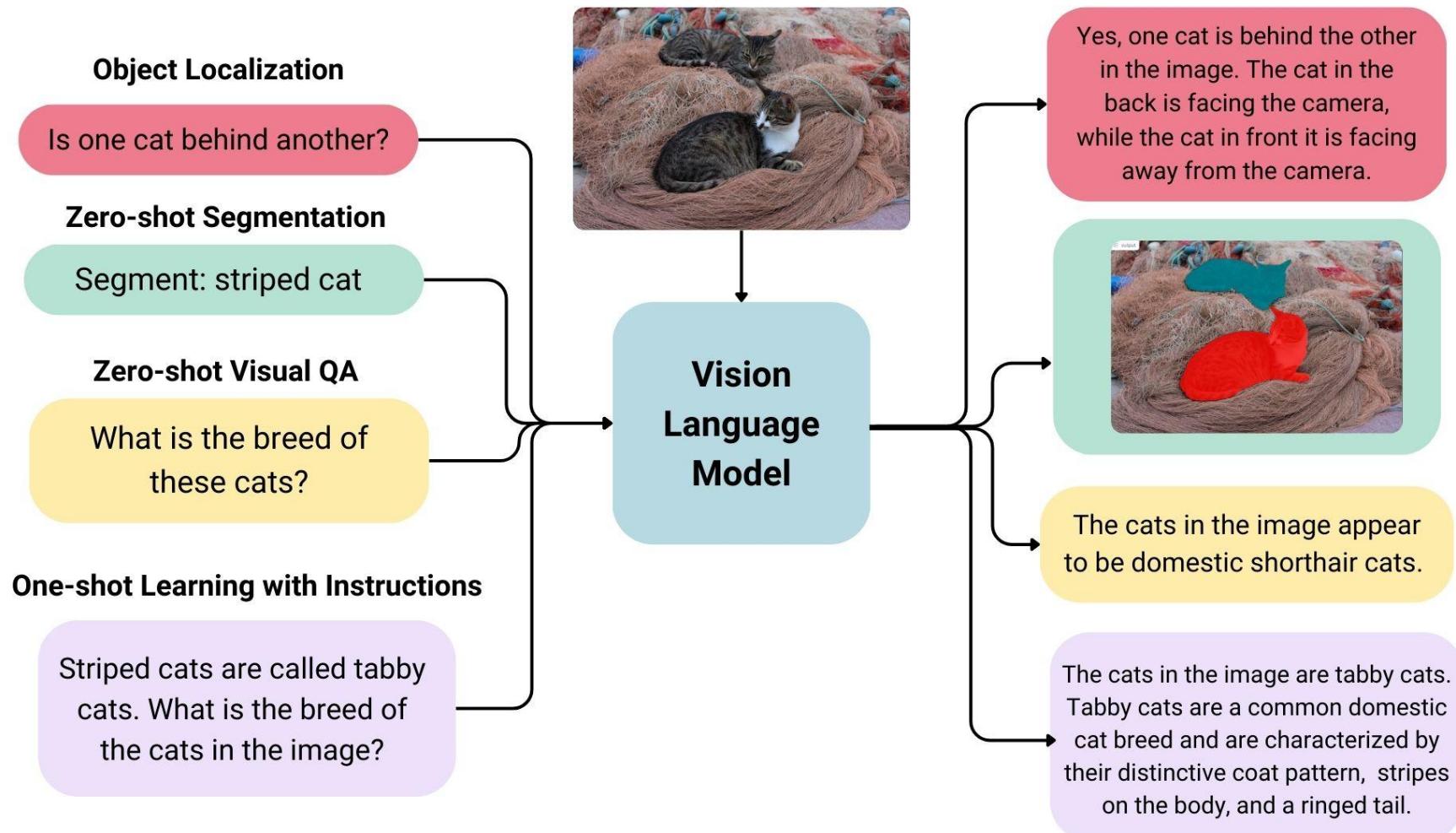
Vision Large Language Models

Inspired From Visual Question Answering Literature



Vision Large Language Models

Inspired From Visual Question Answering Literature



Images are hard to get
For all cases!

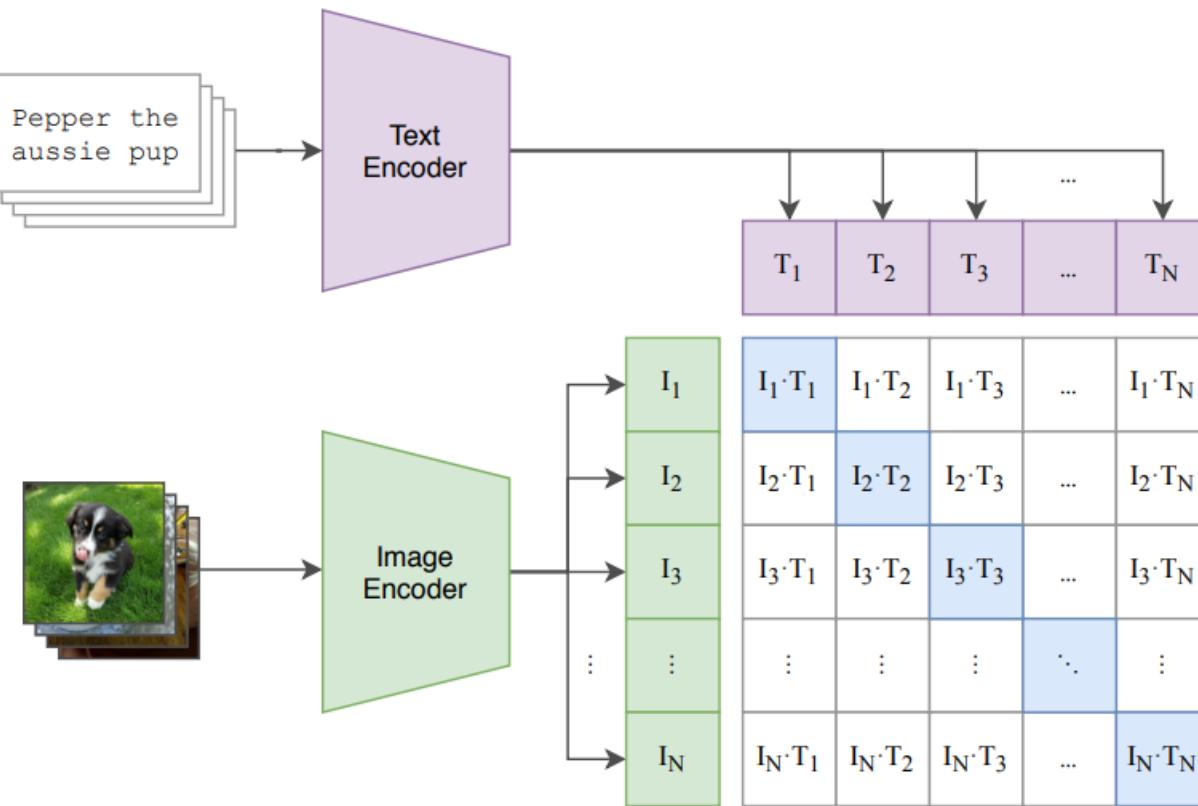
But text is generally easier

Learn from all contexts

How Do You Learn Vision \leftrightarrow Language?

Our Best Friend Embeddings

(1) Contrastive pre-training

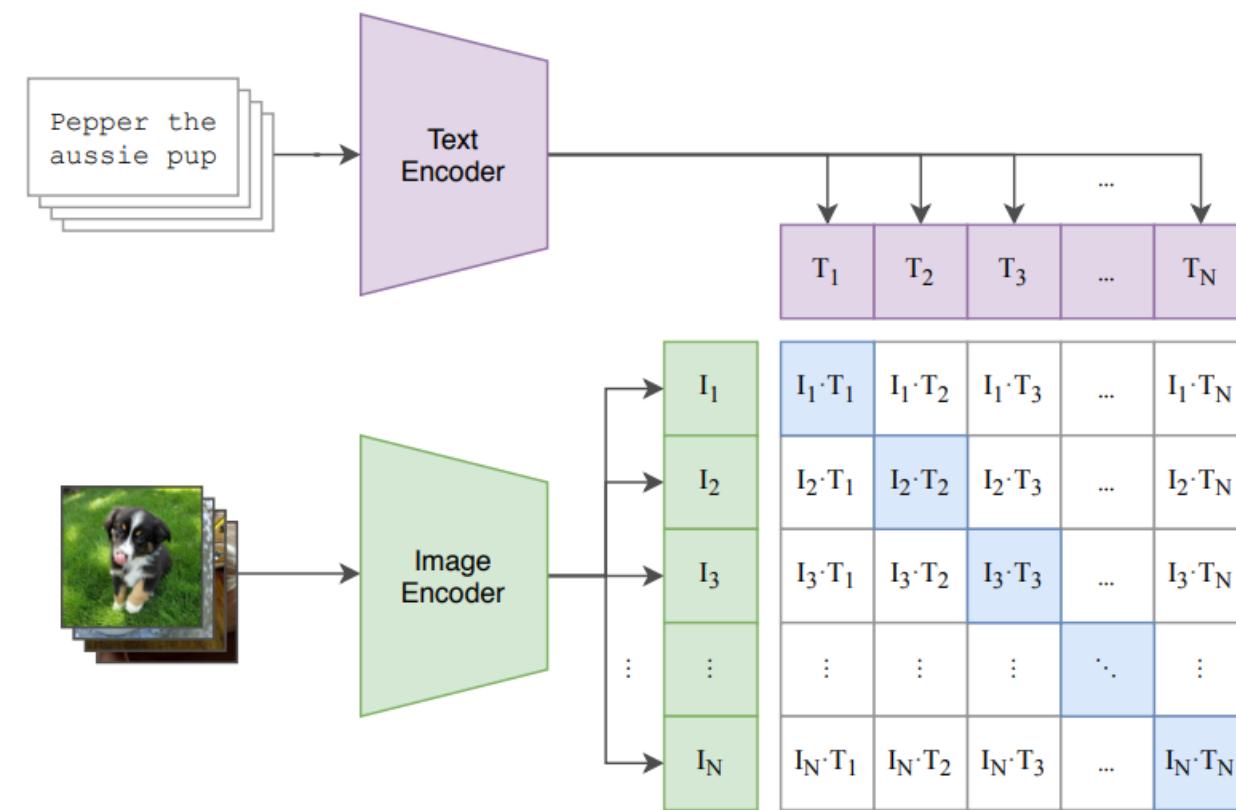


Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

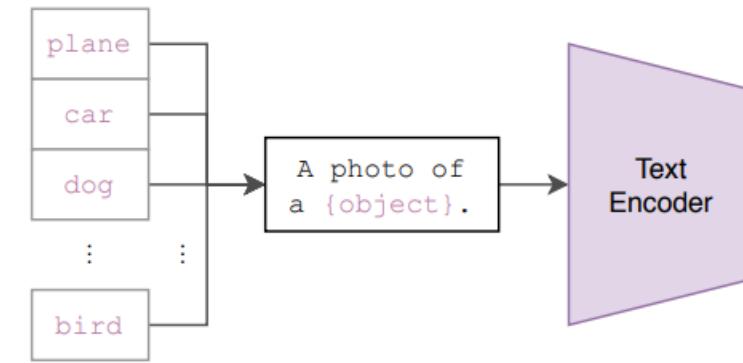
How Do You Learn Vision \leftrightarrow Language?

Our Best Friend Embeddings

(1) Contrastive pre-training



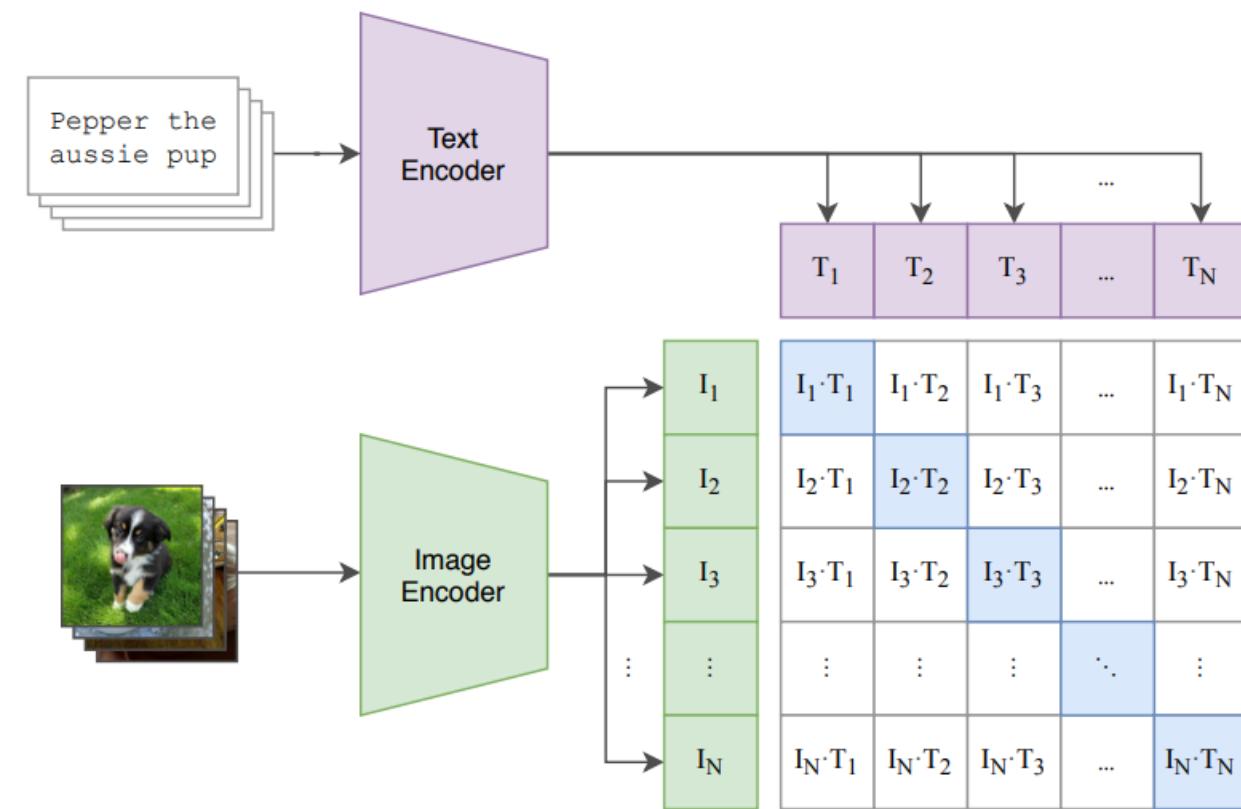
(2) Create dataset classifier from label text



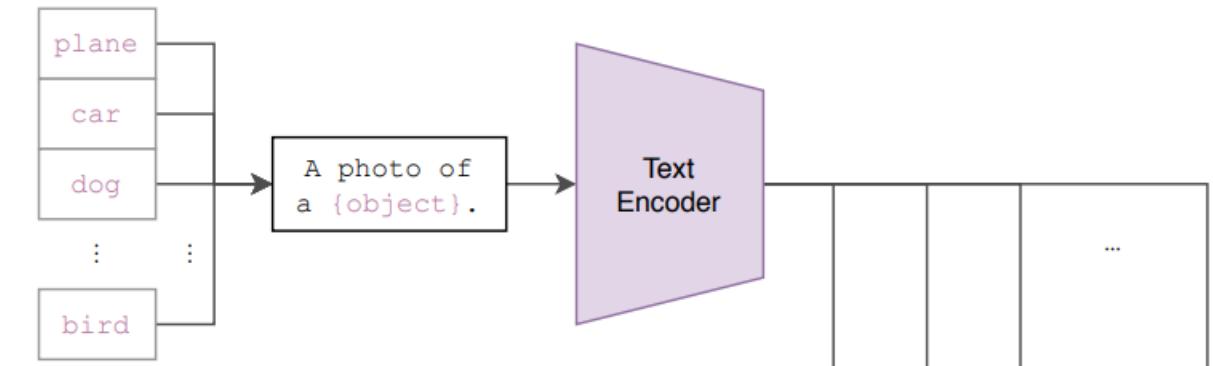
How Do You Learn Vision \leftrightarrow Language?

Our Best Friend Embeddings

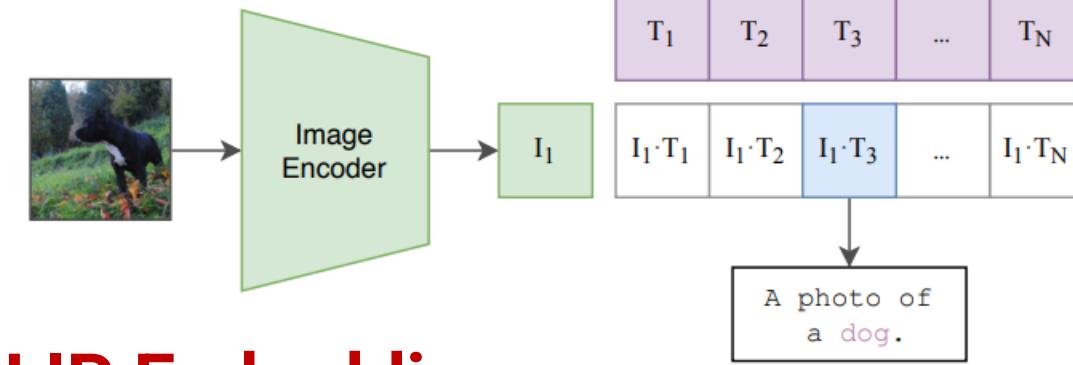
(1) Contrastive pre-training



(2) Create dataset classifier from label text

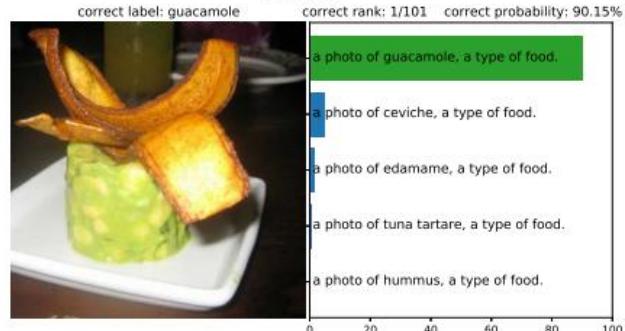
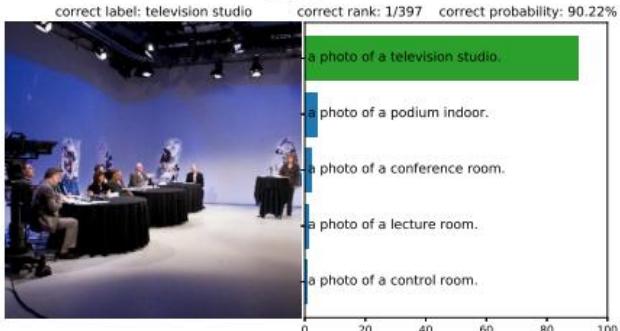
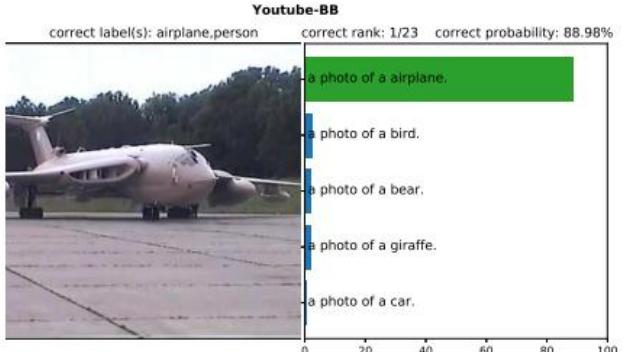
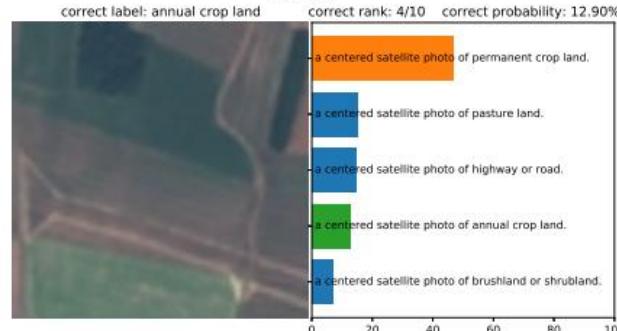
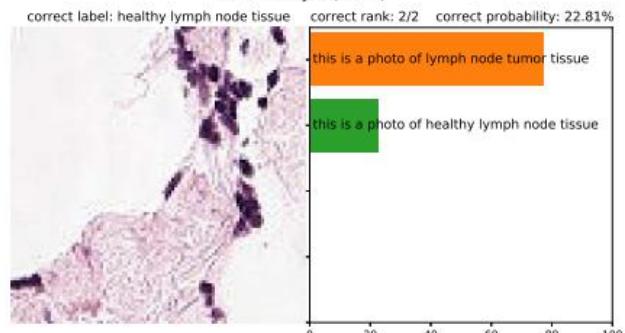
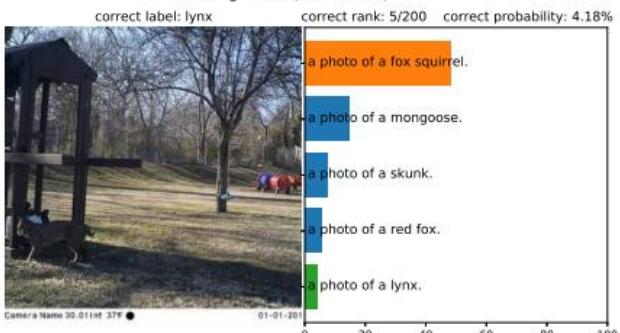
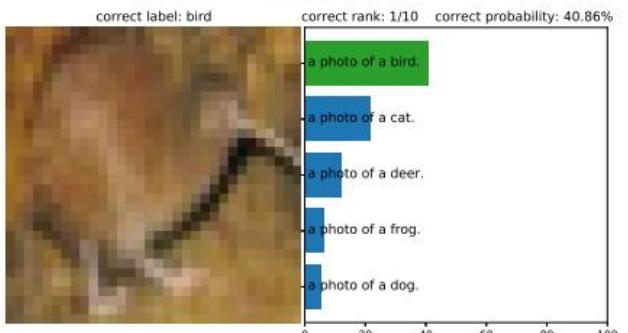
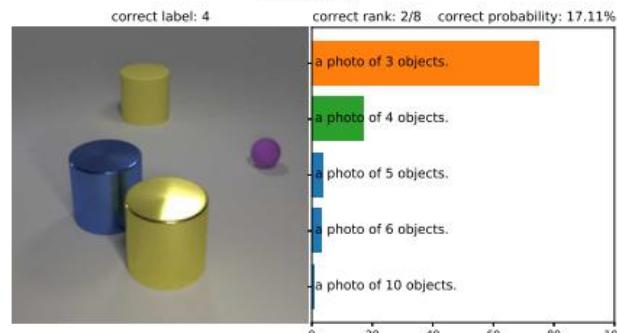
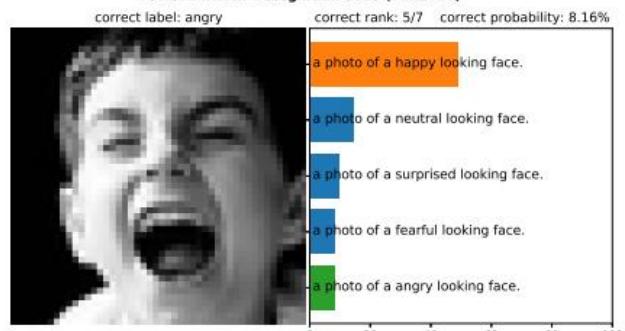
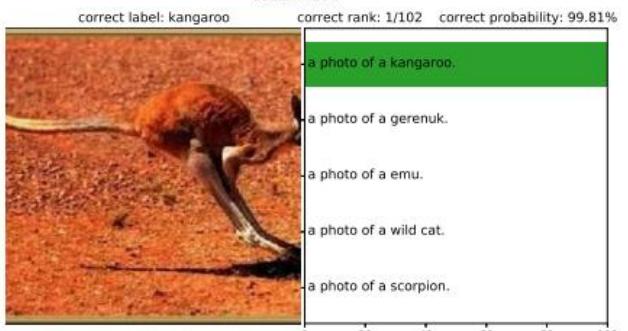
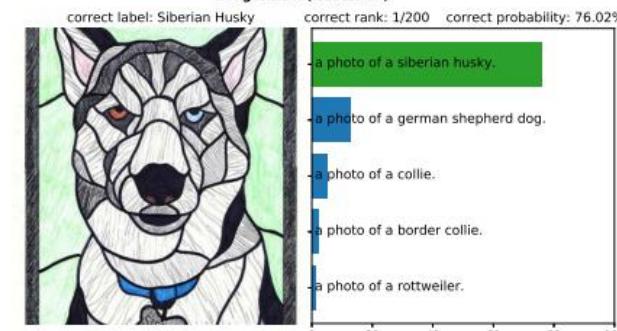


(3) Use for zero-shot prediction



CLIP Embeddings

Works Very Well!

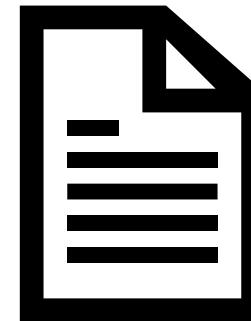
Food101**SUN397****Youtube-BB****EuroSAT****PatchCamelyon (PCam)****ImageNet-A (Adversarial)****CIFAR-10****CLEVR Count****Facial Emotion Recognition 2013 (FER2013)****UCF101****Caltech-101****ImageNet-R (Rendition)**

We Pass Information In Many Ways!

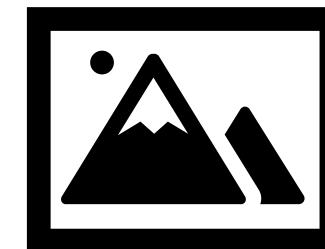
We Want AI To Take Similar Inputs



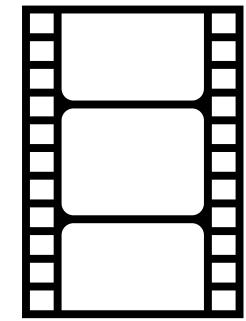
Audio



Text



Images



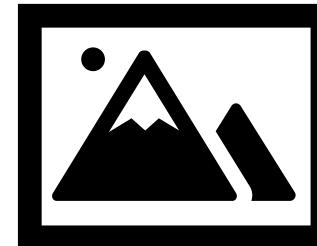
Video

We Pass Information In Many Ways!

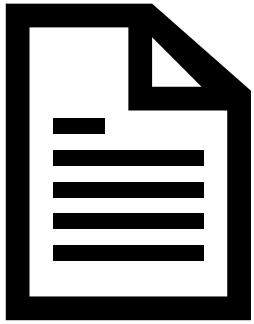
We Want AI To Take Similar Inputs



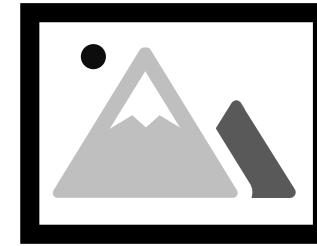
Audio



Images



Text



Depth

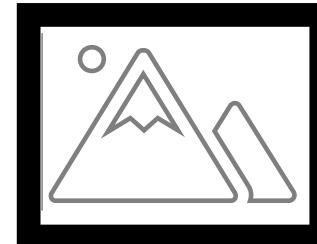


Style

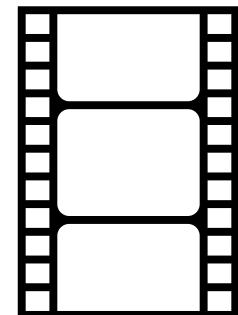


Semantics

•
•
•

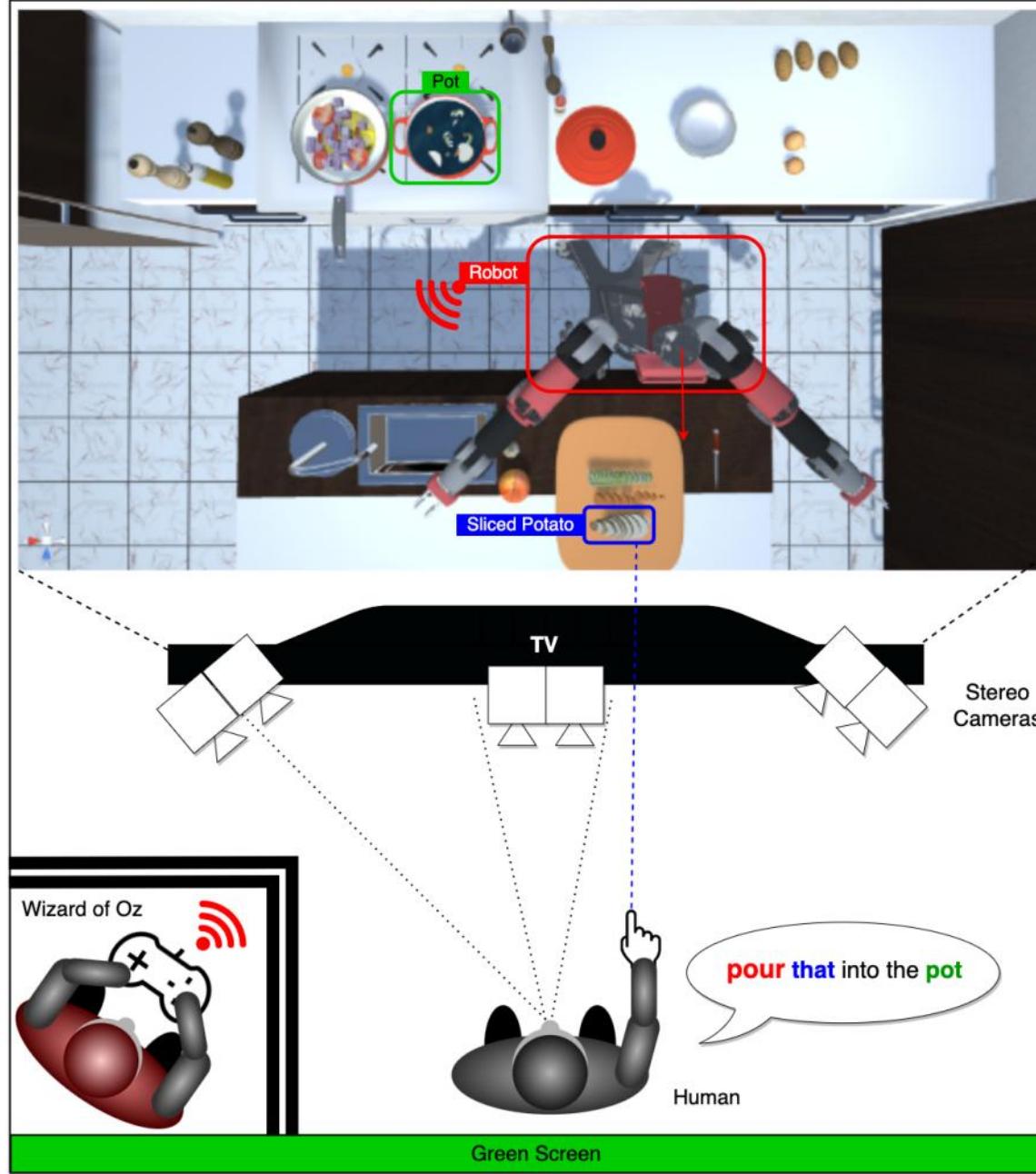


Sketch



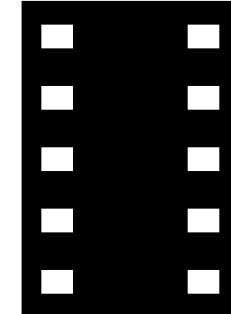
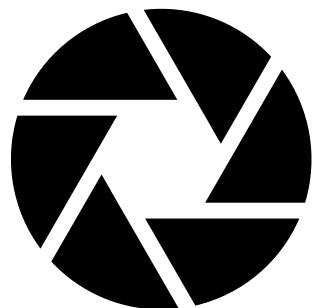
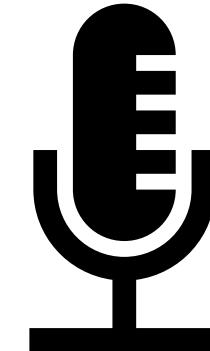
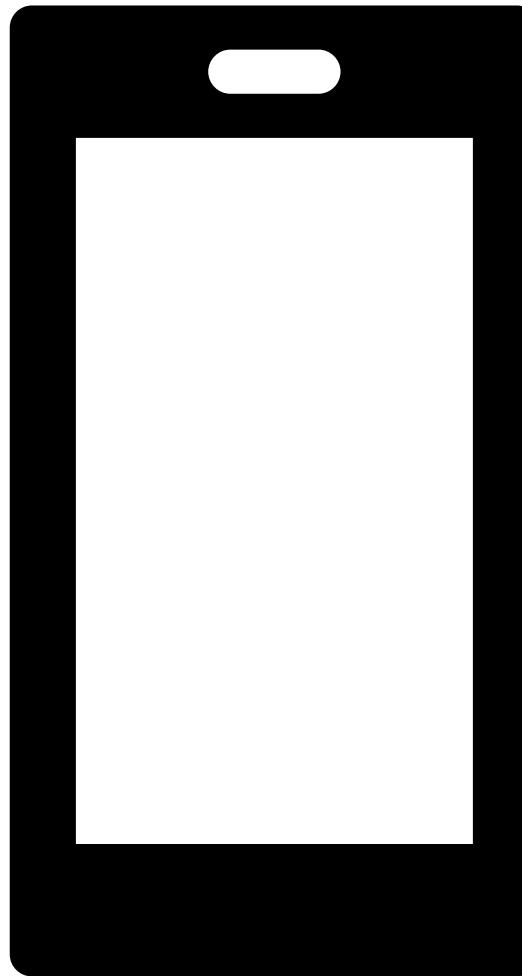
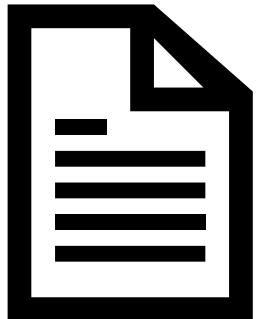
Video

Fun Story!



Shrestha, Snehesh, Ge Gao, Cornelia Fermüller, Yiannis. "Considerations for Minimizing Data Collection Biases for Eliciting Natural Behavior in Human-Robot Interaction." HRI Workshop on Emerging Test Methods & Metrics for Accessible HRI (2023).

Smartphones Are Multimodal!



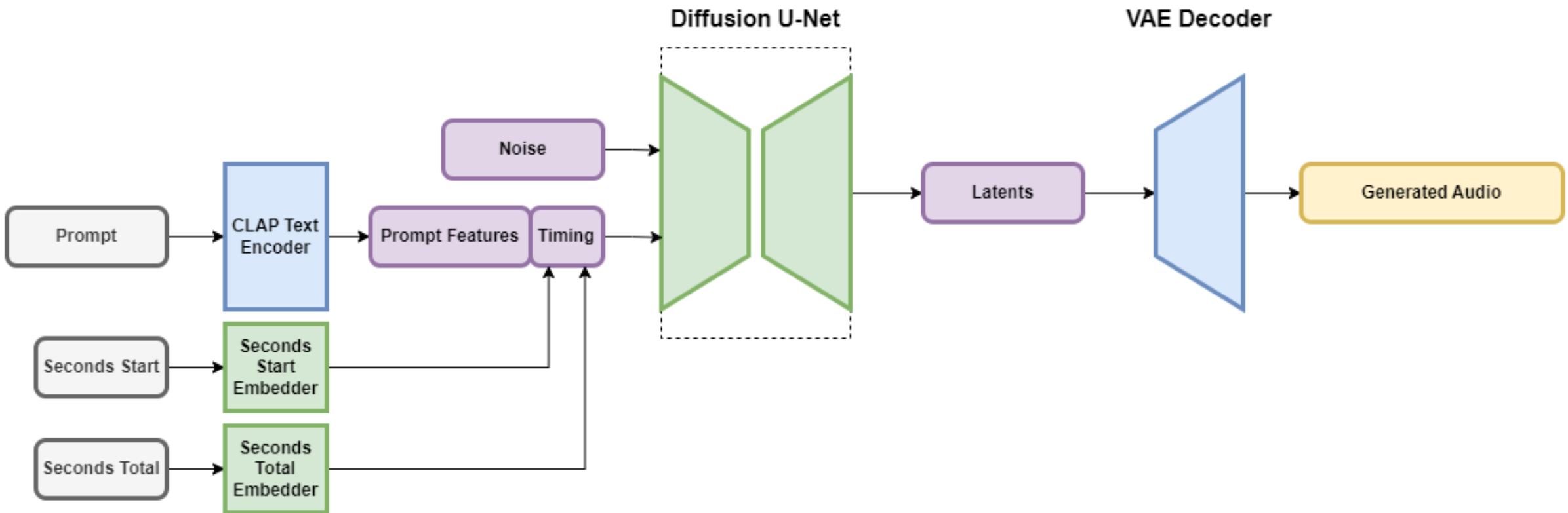


Audio Generation

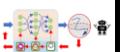
The screenshot shows the Stable Audio interface. On the left, the 'Prompt' section contains the text: 'Disco, Driving Drum Machine, Synthesizer, Bass, Piano, Guitars, Instrumental, Clubby, Euphoric, Chicago, New York, 115 BPM'. Below it are settings for 'Prompt Library' (None), 'Model' (Stable Audio AudioSpax 2.0), 'Duration (max. 3m 0s)' (3m 0s), 'Input audio (3m 0s remaining)' (Add audio), and 'Add extras'. A 'Generate' button is visible. On the right, a large green circular progress bar indicates the generation process. Below the bar, the text reads: 'Disco, Driving Drum Machine, Synthesizer, Bass, Piano, Guitar...'. Further down, it says 'Stable Audio AI music creation' and 'stable-audio-audiosparx-v2-0 Seed: 23266'. At the bottom, there's a waveform preview from 0:00 to 3:00, with various control icons like play, stop, and download.

The screenshot shows the generated audio file in the 'History' tab. The file is titled 'Disco, Driving Drum Machine, Synthesizer, Bass, Piano, Guitars, Instrumental, Clubby, Euphoric, Chicago, New York, 115 BPM' and has a duration of '3:00 / Less than a minute ago'. It includes download and other file management icons.

Speech Generation



[Stable Audio: Fast Timing-Conditioned Latent Audio Diffusion – Stability AI](#)



Speech Generation

Keeping Your Heroes Alive Forever

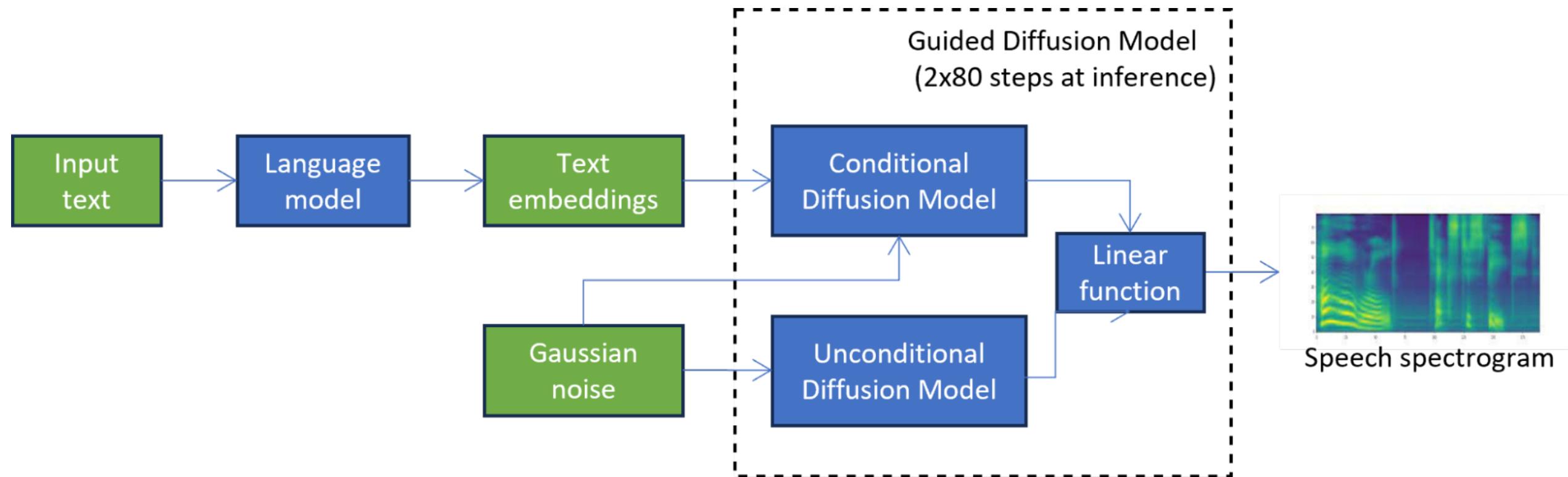


Sir David Attenborough

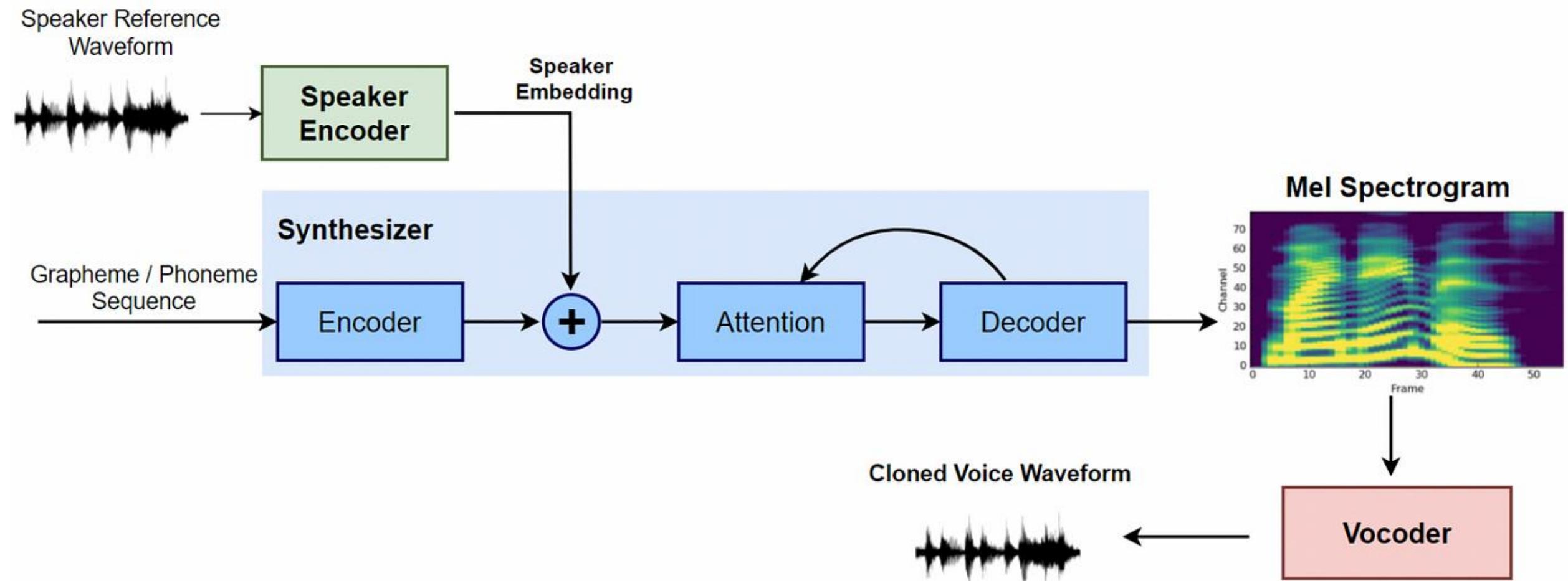


Morgan Freeman

Speech Generation



Speech Cloning



Video Generation

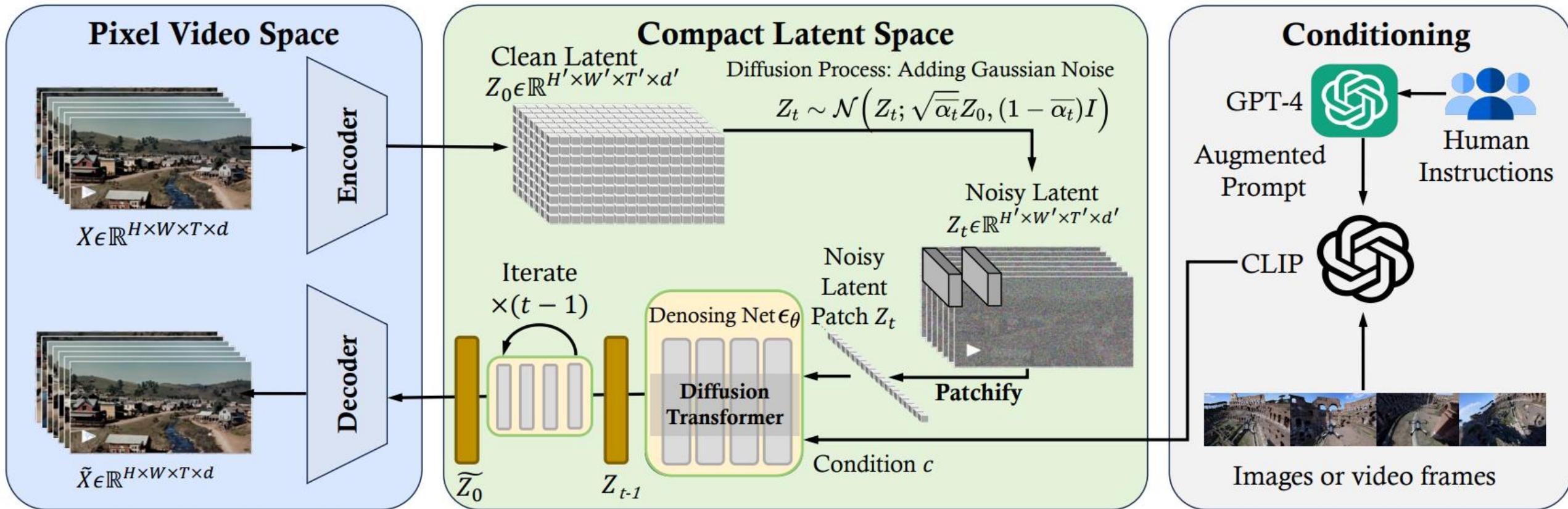


Prompt: This close-up shot of a Victoria crowned pigeon showcases its striking blue plumage and red chest. Its crest is made of delicate, lacy feathers, while its eye is a striking red color. The bird's head is tilted slightly to the side, giving the impression of it looking regal and majestic. The background is blurred, drawing attention to the bird's striking appearance.

Prompt: Photorealistic closeup video of two pirate ships battling each other as they sail inside a cup of coffee.



OpenAI Sora

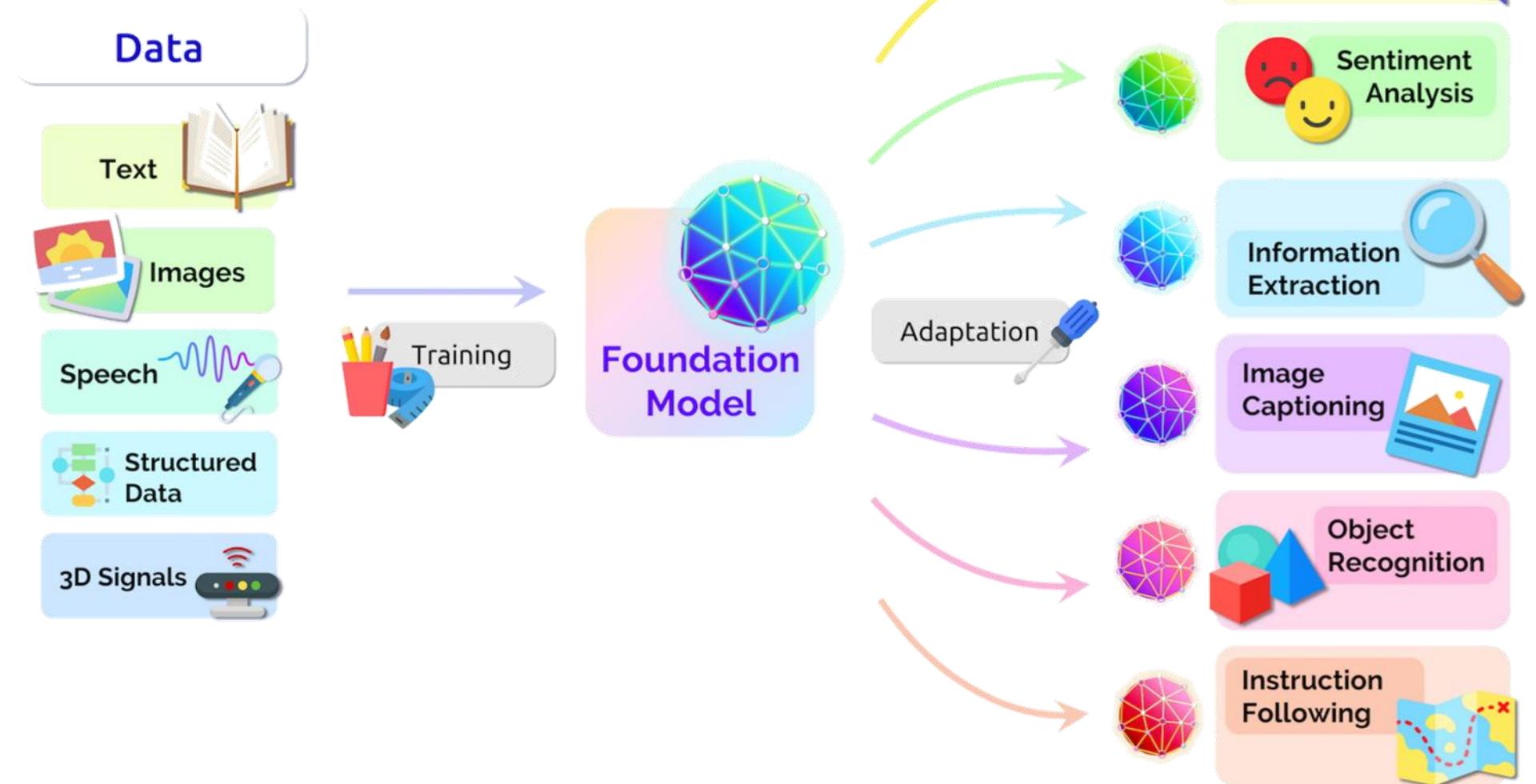


Sora | OpenAI

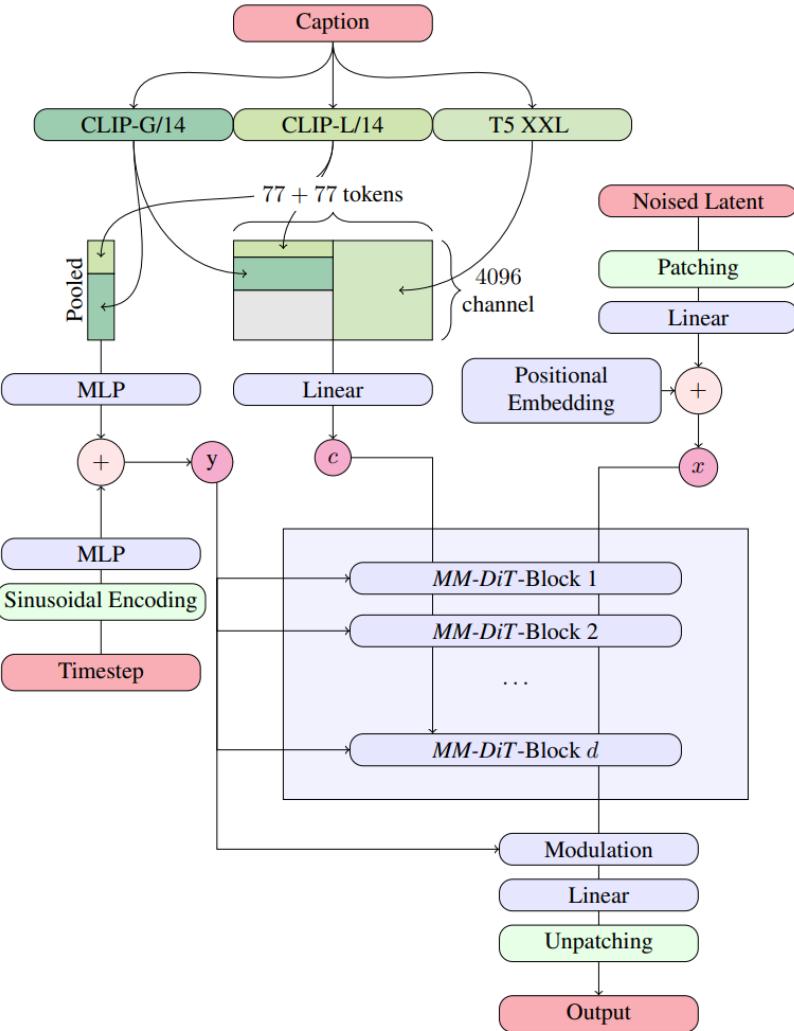
Liu, Yixin, et al. "Sora: A review on background, technology, limitations, and opportunities of large vision models." arXiv preprint arXiv:2402.17177 (2024).

Foundational Models

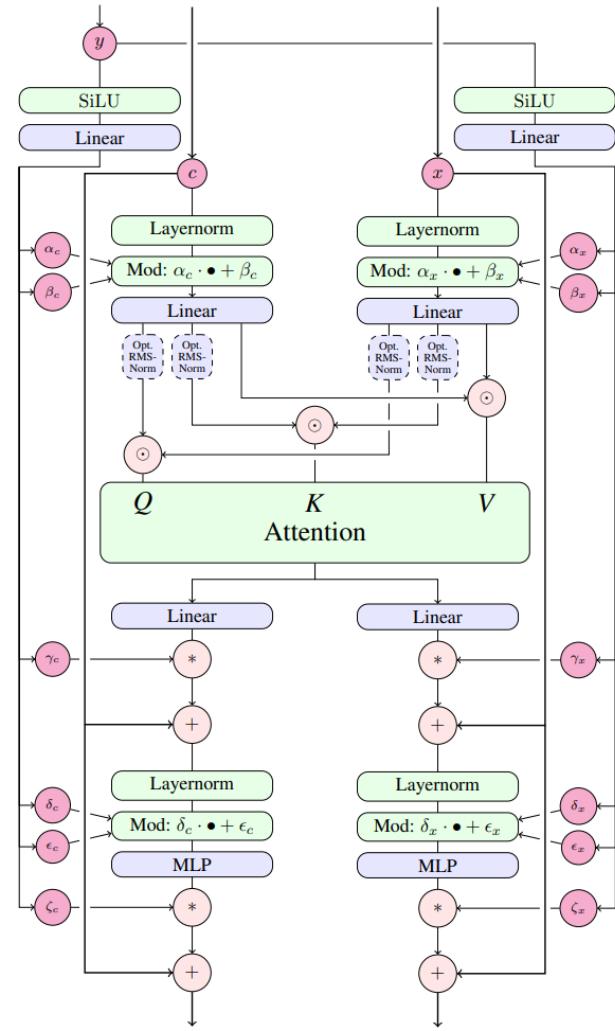
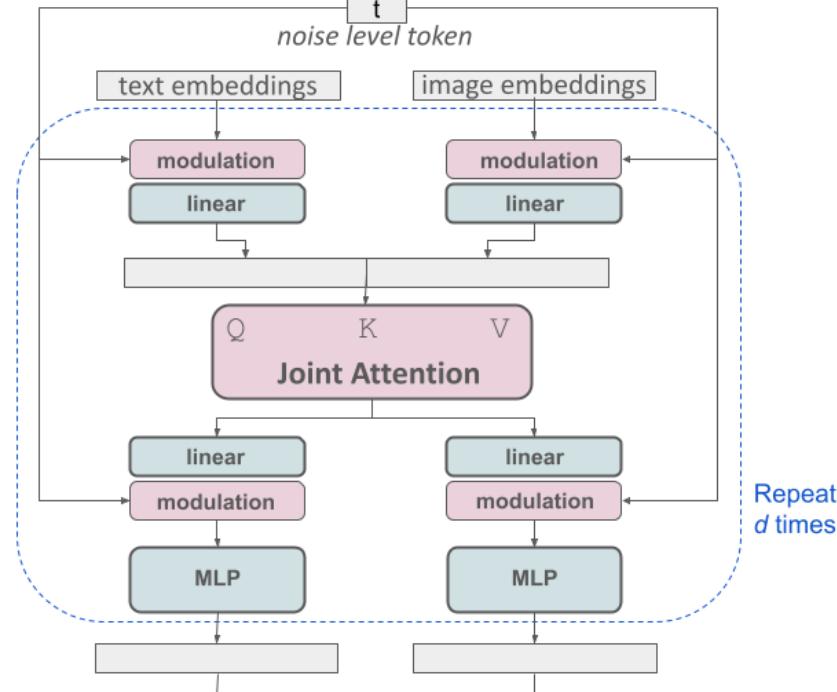
Generally Trained On Multi-modal Datasets



Stable Diffusion 3



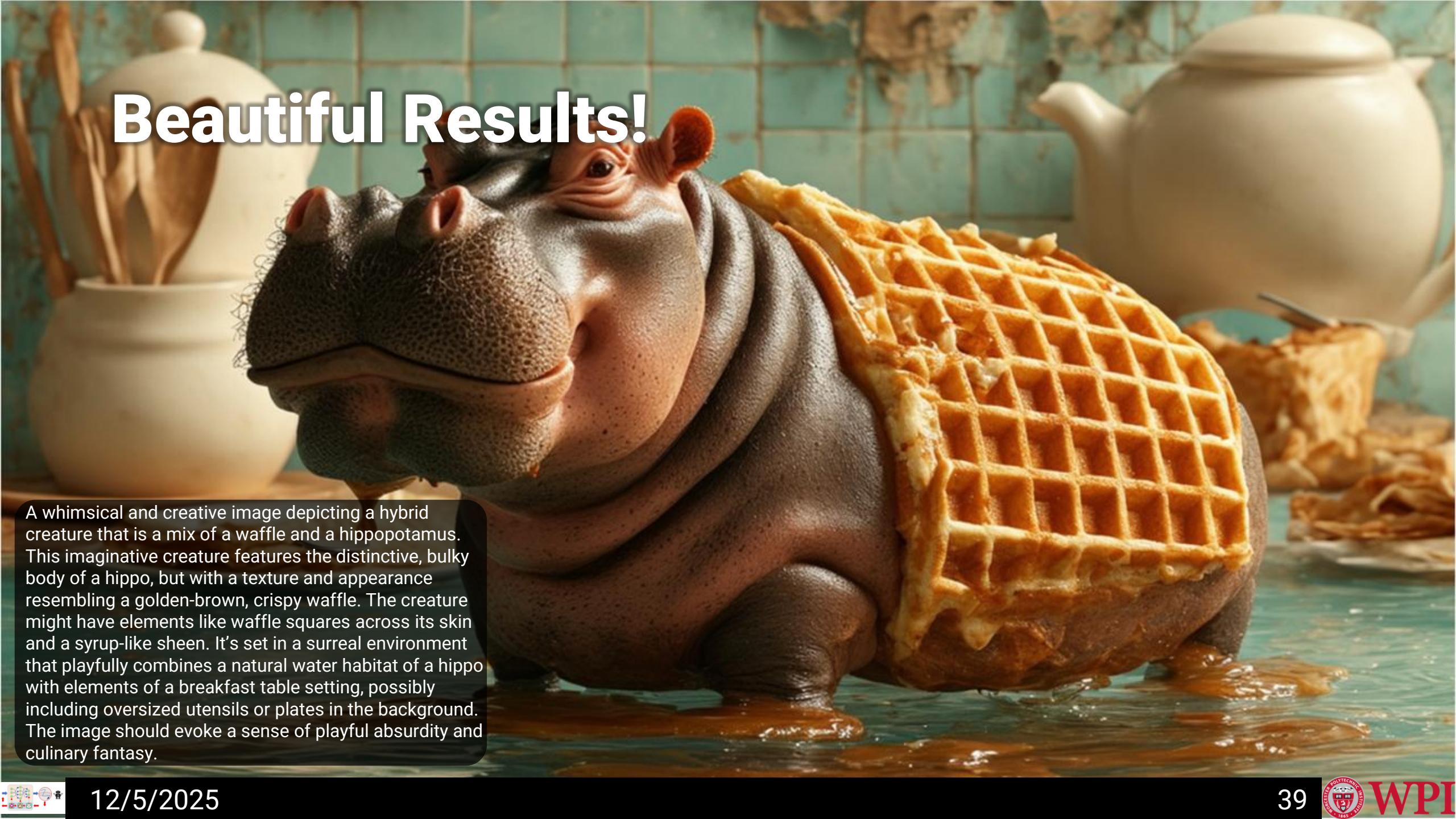
(a) Overview of all components.

(b) One **MM-DiT** block

Esser, Patrick, et al. "Scaling rectified flow transformers for high-resolution image synthesis." Forty-first International Conference on Machine Learning. 2024.

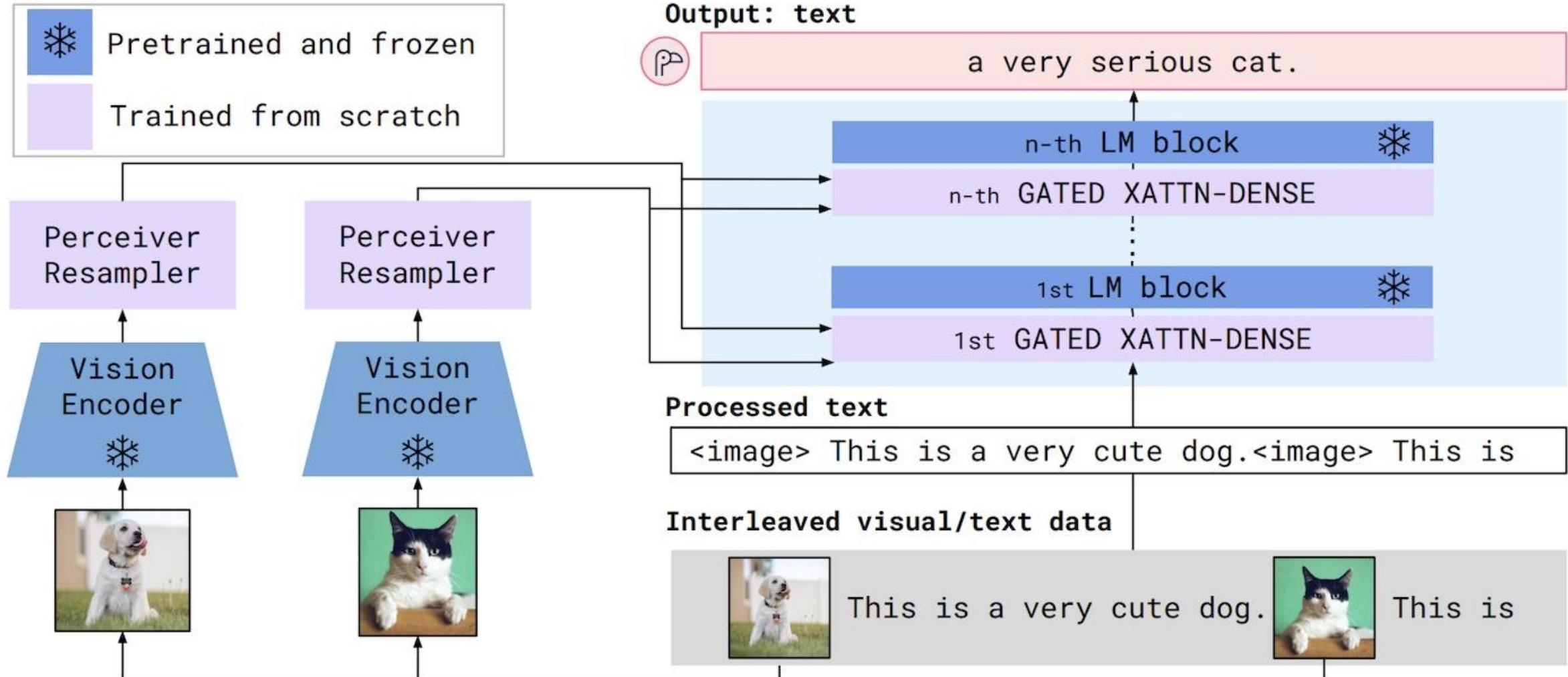
Beautiful Results!

A whimsical and creative image depicting a hybrid creature that is a mix of a waffle and a hippopotamus. This imaginative creature features the distinctive, bulky body of a hippo, but with a texture and appearance resembling a golden-brown, crispy waffle. The creature might have elements like waffle squares across its skin and a syrup-like sheen. It's set in a surreal environment that playfully combines a natural water habitat of a hippo with elements of a breakfast table setting, possibly including oversized utensils or plates in the background. The image should evoke a sense of playful absurdity and culinary fantasy.





Embedding Magic!



Works Well!



P This is a picture of two teddy bears on the moon.

What are they doing?

P They are having a conversation.

What object are they using?

P It looks like a computer.

Is this surprising?

P Yes, it is surprising.

Why is this picture surprising to you?

P I think it is surprising because teddy bears are not usually found on the moon.

Input Prompt



This is a chinchilla. They are mainly found in Chile.



This is a shiba. They are very popular in Japan.

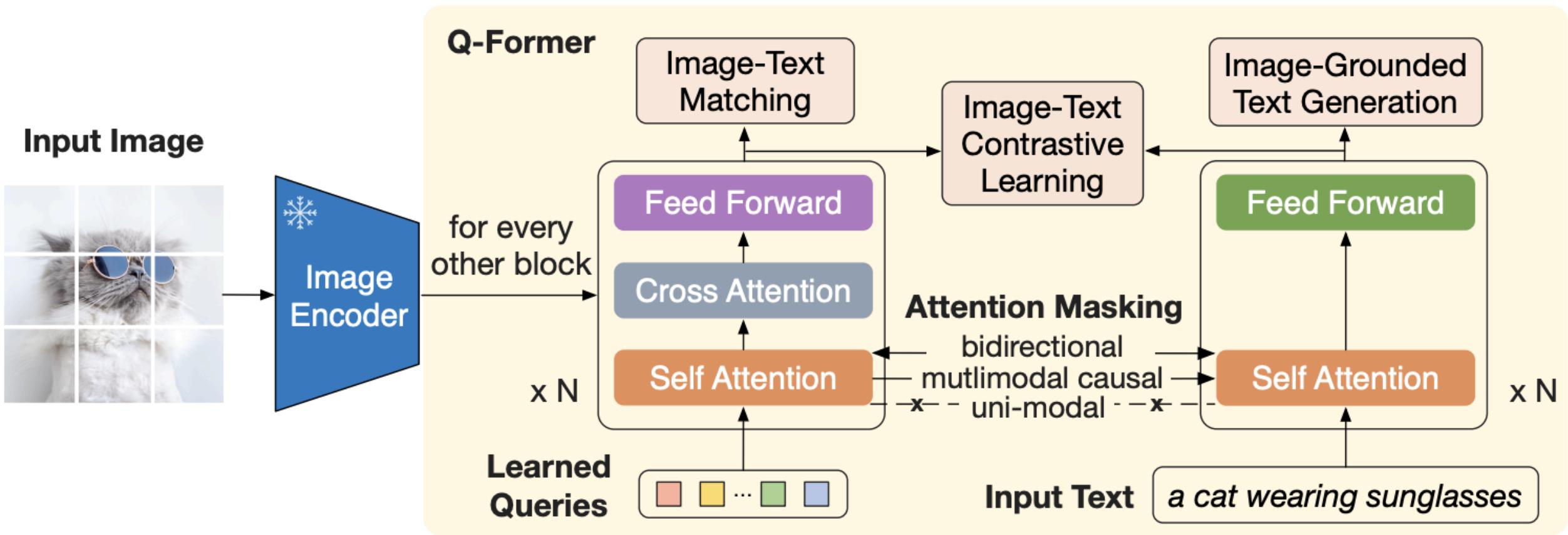


This is

Completion

a flamingo.
They are found in the Caribbean and South America.

Embedding Magic!



Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." International conference on machine learning. PMLR, 2023.

Very Cool!



8

Explain the advantages of this product.

8

The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.



8

Tell me something about the history of this place.

8

The great wall of china was built in 221 bc by emperor qin shihuang to protect the emperor's capital from invaders from the north.



8

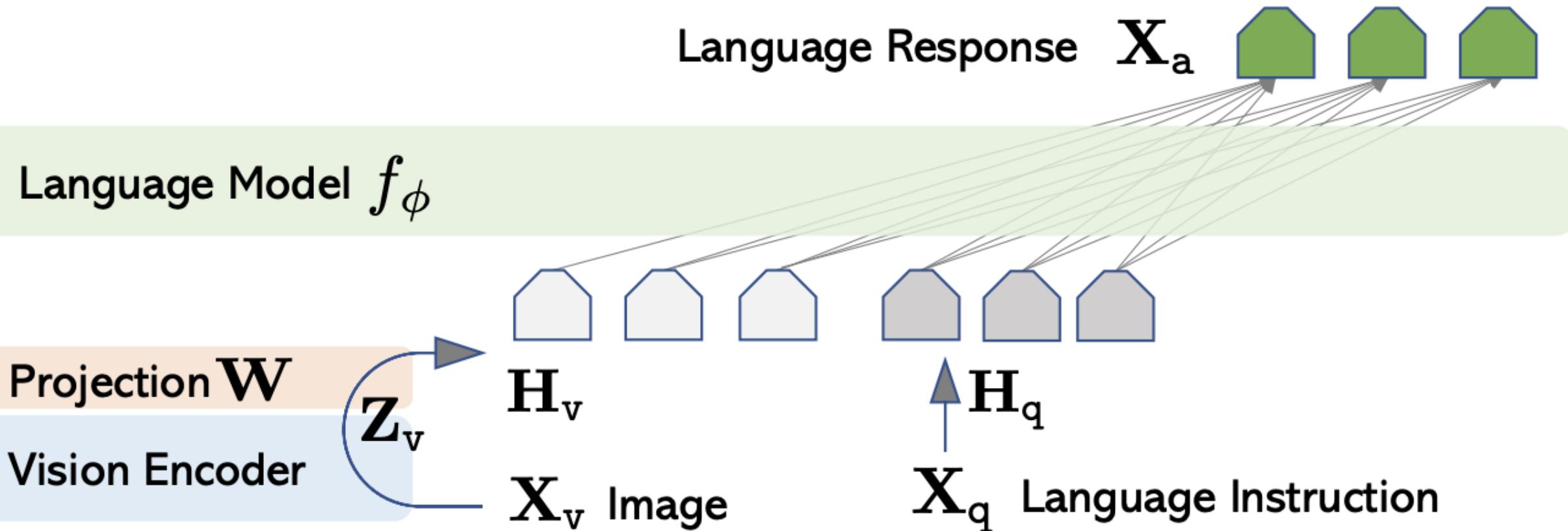
Write down the facts that you know about this flower.

8

It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.

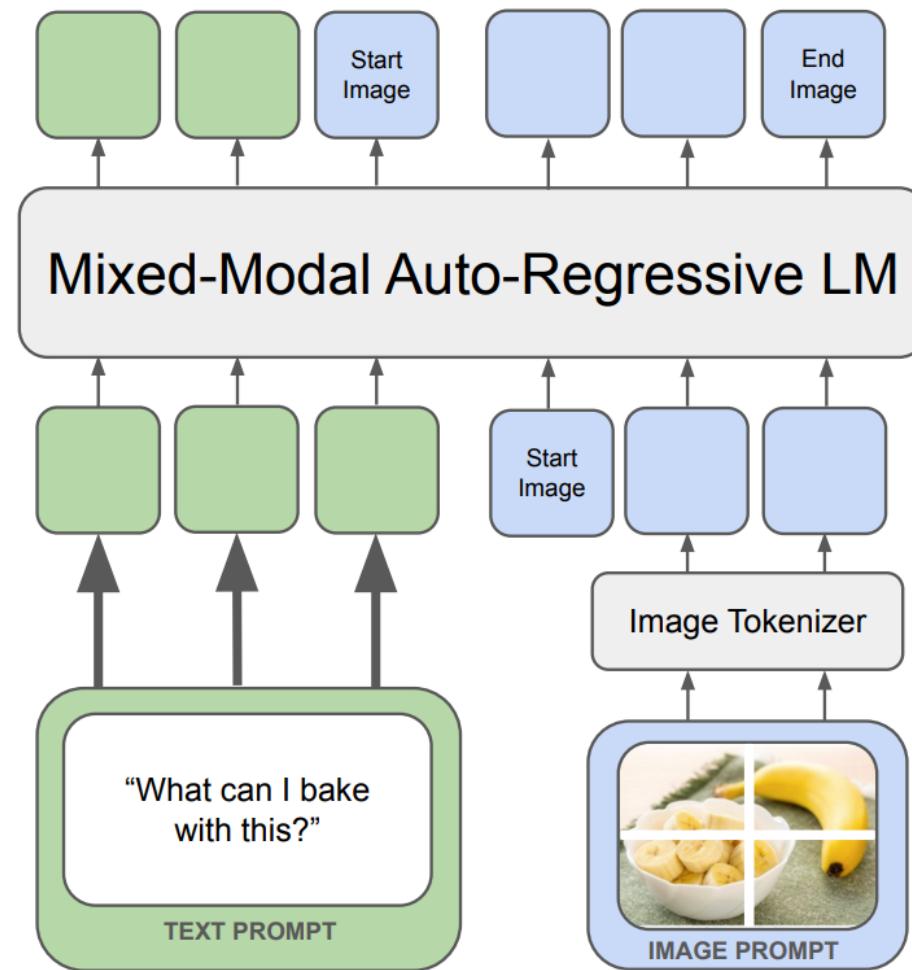


LLaVA

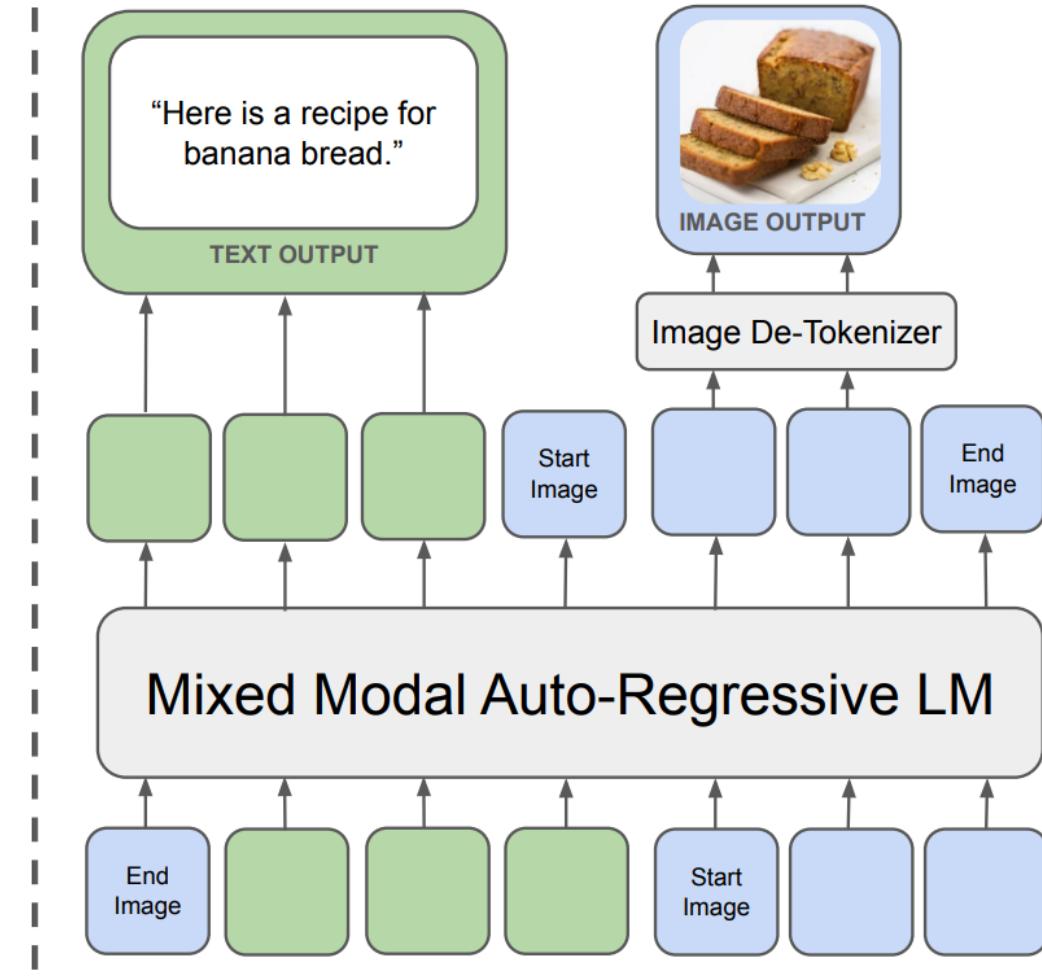


Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2024).

Chameleon



(a) Mixed-Modal Pre-Training



(b) Mixed-Modal Generation

Team, Chameleon. "Chameleon: Mixed-modal early-fusion foundation models." arXiv preprint arXiv:2405.09818 (2024).

Maybe Soon On WhatsApp



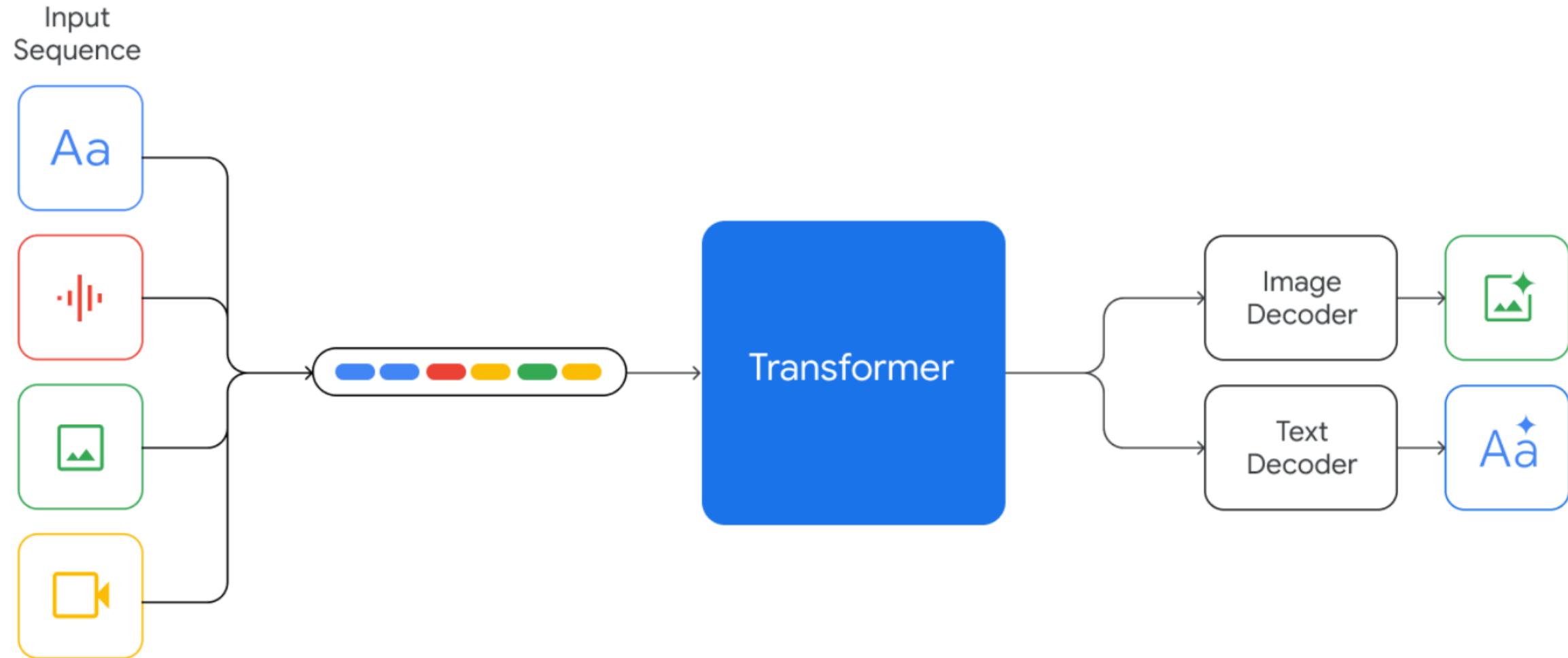
Prompt: I'm bored. Could you show me some cool, quirky-looking birds? Also tell me about them in 2–3 sentences.



Response: Absolutely! Here are three birds with some unique features that you might find interesting:

Keel-Billed Toucan The Keel-Billed Toucan is a brightly colored bird native to South American rainforests and woodlands. Its most distinctive feature is its colorful beak, which can grow up to seven inches long and has a distinctive, curved shape. The beak is used for plucking fruit from trees and for social displays. Despite its large size, the Keel-Billed Toucan is a skilled flyer and can often be seen flitting through the treetops. Here is an image of this vibrant bird:

Gemini



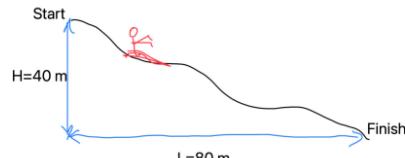
Team, Gemini, et al. "Gemini: a family of highly capable multimodal models." arXiv preprint arXiv:2312.11805 (2023).

Interesting!

Prompt

Here's a solution to a Physics problem by a student.

23. A skier slides down the frictionless slope as shown. What is the skier's speed at the bottom?



Solution:

The conservation of energy means that the energy of the skier is same at the beginning and at the end.

$$\text{Start: } E = mgL$$

$$\text{End: } E = \frac{1}{2}mv^2$$

$$\Rightarrow mgL = \frac{1}{2}mv^2$$

$$\Rightarrow v = \sqrt{2gL} = \sqrt{2 \times 9.81 \times 80} = 39.6 \text{ m.s}^{-1}$$

Try to reason about the question step by step.

Did the student get the correct answer? If the solution is wrong, please explain what is wrong and solve the problem.

Make sure to use LaTeX for math and round off the final answer to two decimal places.

Response

The student did not get the correct answer. The student made a mistake in the calculation of the potential energy at the start of the slope. The potential energy at the start of the slope is given by mgh , where m is the mass of the skier, g is the acceleration due to gravity, and h is the height of the slope. The student used the length of the slope, L , instead of the height, H , in the calculation of the potential energy.

The correct solution is:

The conservation of energy means that the total energy of the skier is the same at the beginning and at the end of the slope.

At the start of the slope, the skier has only potential energy, which is given by mgh .

At the end of the slope, the skier has only kinetic energy, which is given by $\frac{1}{2}mv^2$

Therefore, we can write:

$$mgh = \frac{1}{2}mv^2$$

Cancelling out the mass m on both sides, we get:

$$gh = \frac{1}{2}v^2$$

Solving for v , we get:

$$v = \sqrt{2gh}$$

Substituting the given values, we get:

$$v = \sqrt{2 \times 9.81 \times 40} = 28.01 \text{ m/s}$$

Therefore, the skier's speed at the bottom of the slope is 28.01 m/s.

GPT-4o

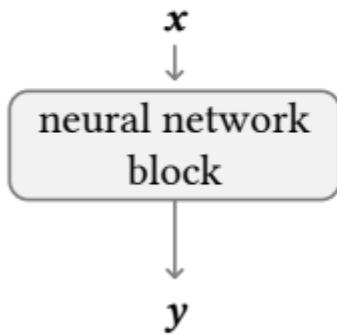


Wearable Tech FTW

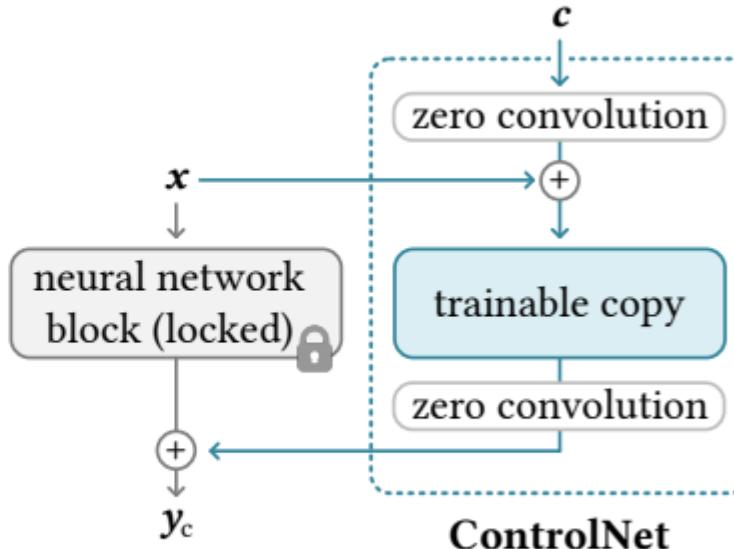


How Do You Control Generation?

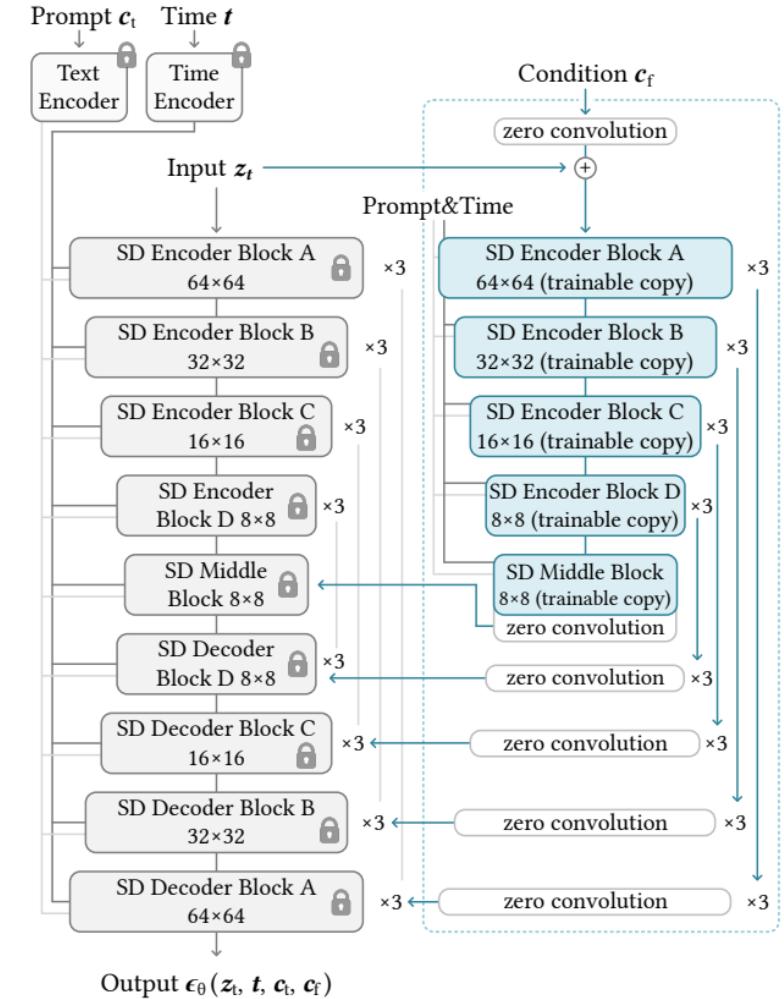
For Novel Inputs Retraining Is Too Slow 😞



(a) Before



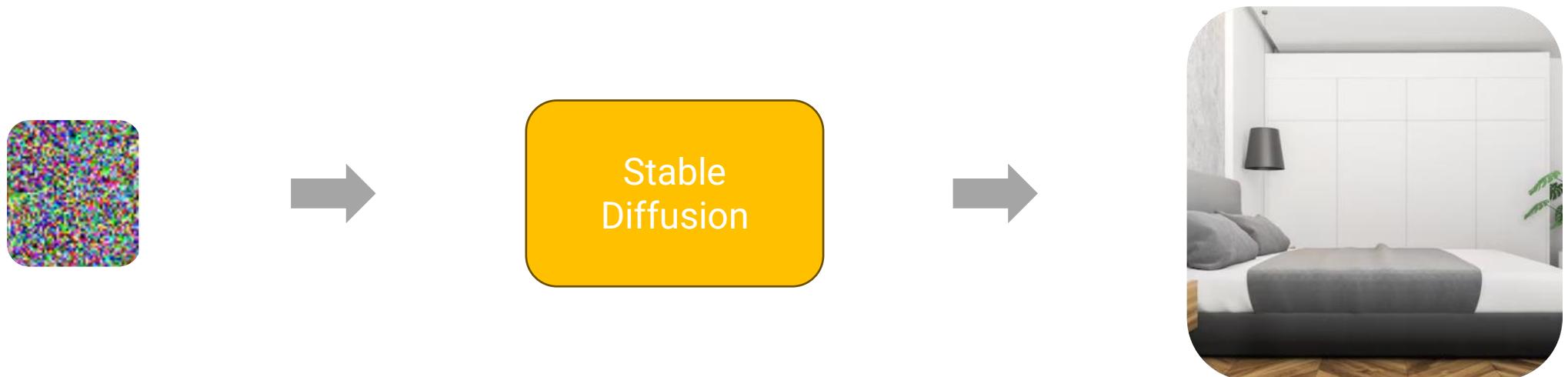
(b) After



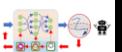
(a) Stable Diffusion

(b) ControlNet

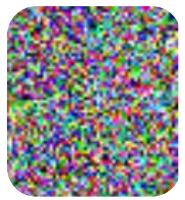
CtrlNet Is Amazing!



🧐 Ok, but what IS ControlNet? ([youtube.com](https://www.youtube.com))



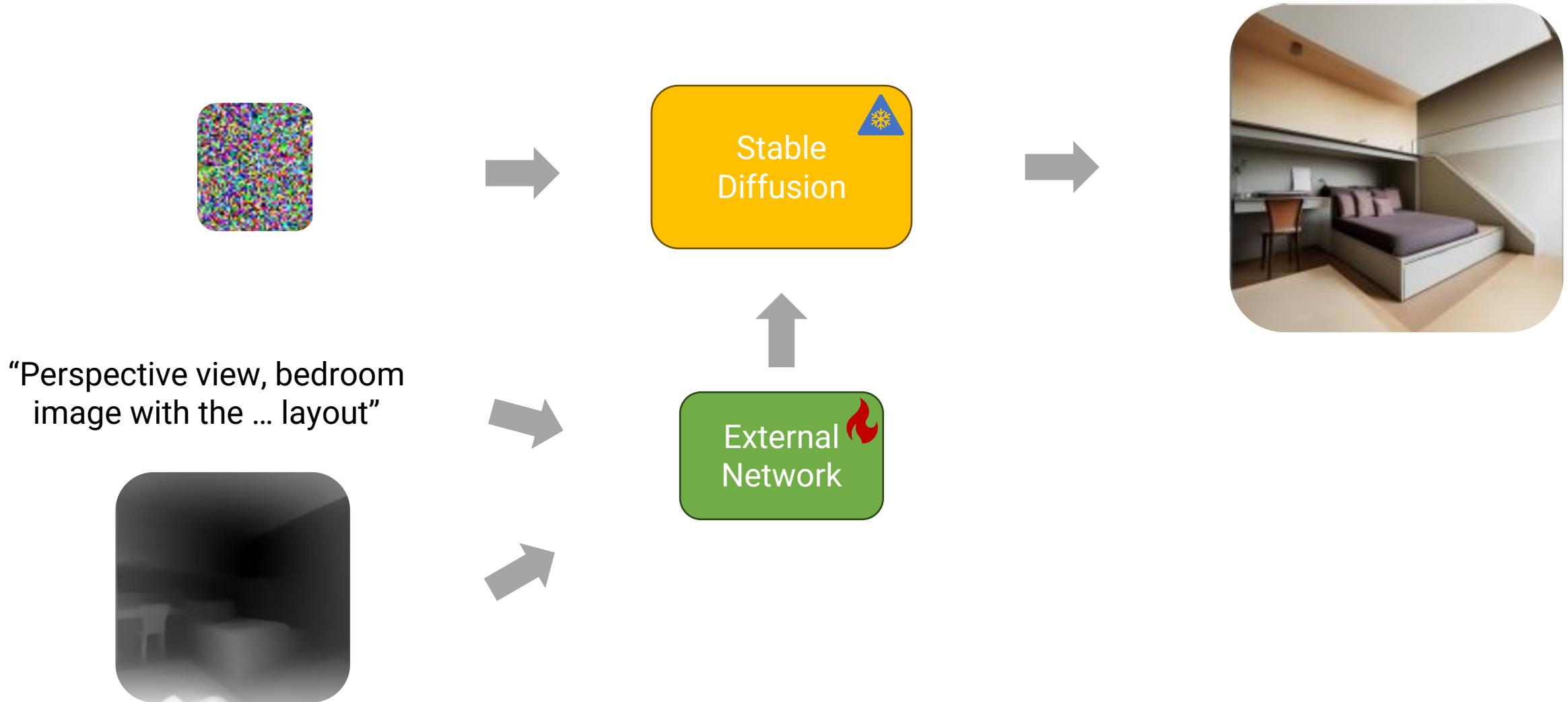
CtrlNet Is Amazing!



“Perspective view, bedroom
image with the ... layout”



CtrlNet Is Amazing!



Can Do Cool Stuff!

Sketch

Normal map

Depth map

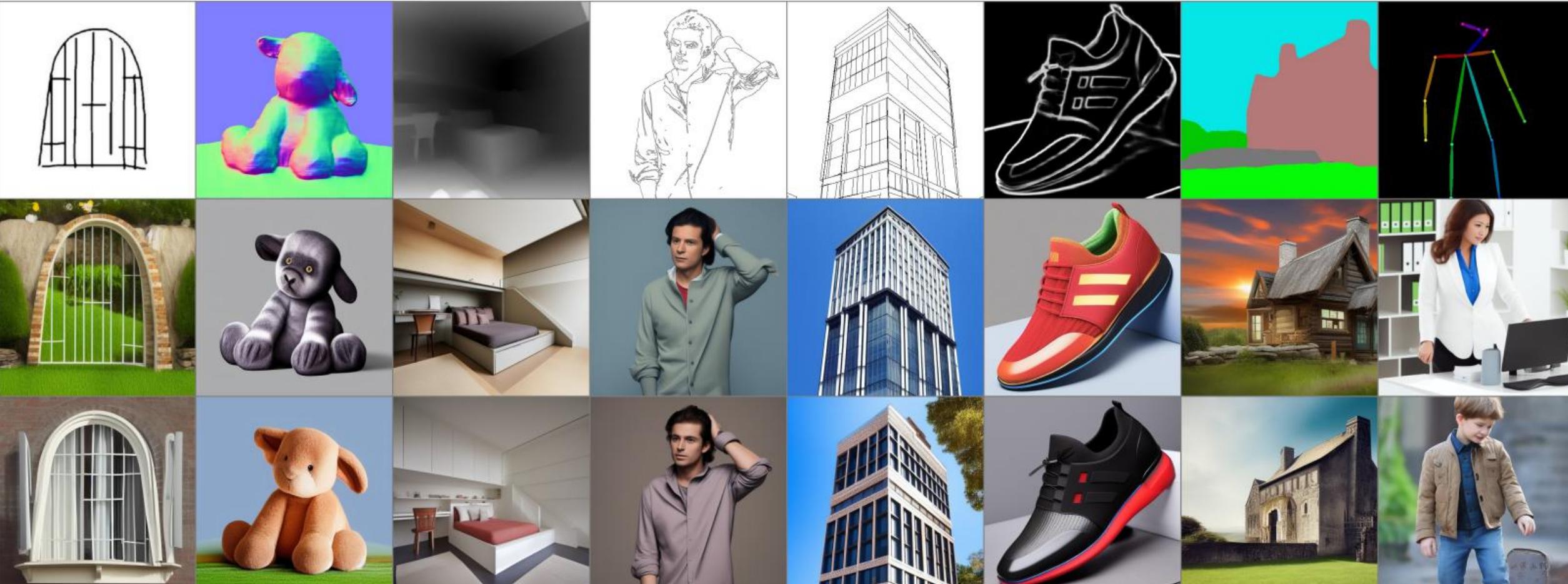
Canny[11] edge

M-LSD[24] line

HED[91] edge

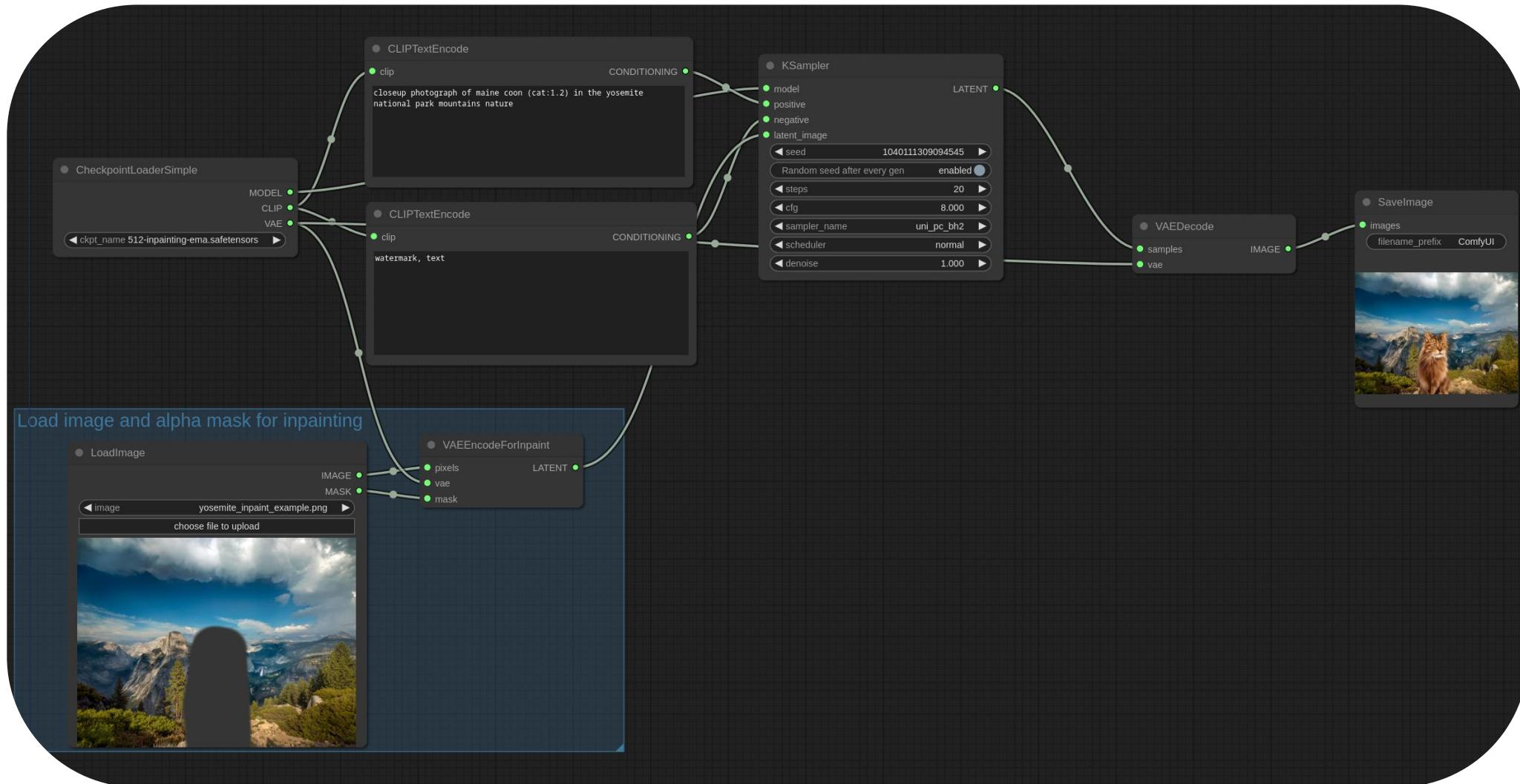
ADE20k[96] seg.

Human pose

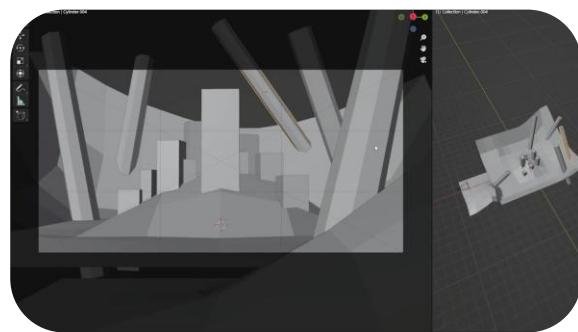


Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

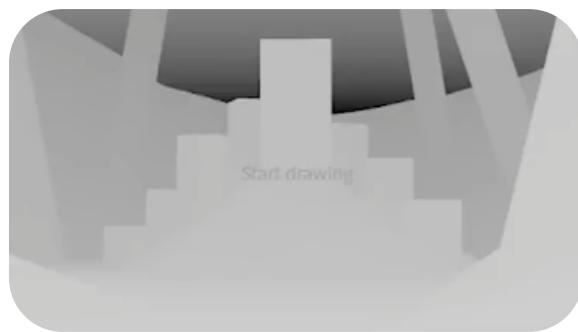
ComfyUI



GenAI DataGen!



Simple Geometry



Depth Map

Can also do Semantic Maps



(16229) Use AI to create amazing 3D Animations! [Blender + Stable Diffusion] - YouTube

A photograph of a person standing on a rocky cliff edge, looking up at the sky with their arms outstretched. The person is wearing a blue jacket, red pants, and sunglasses. The background is a bright, cloudy sky.

What A Time To Be Alive Right?

AKA Everything Is Amazing, Right?

Recall CLIP
400M Data samples
250-600 GPUs for 18 days!

Scalability



How Much Does 1 ChatGPT Search Cost?



ChatGPT
Google

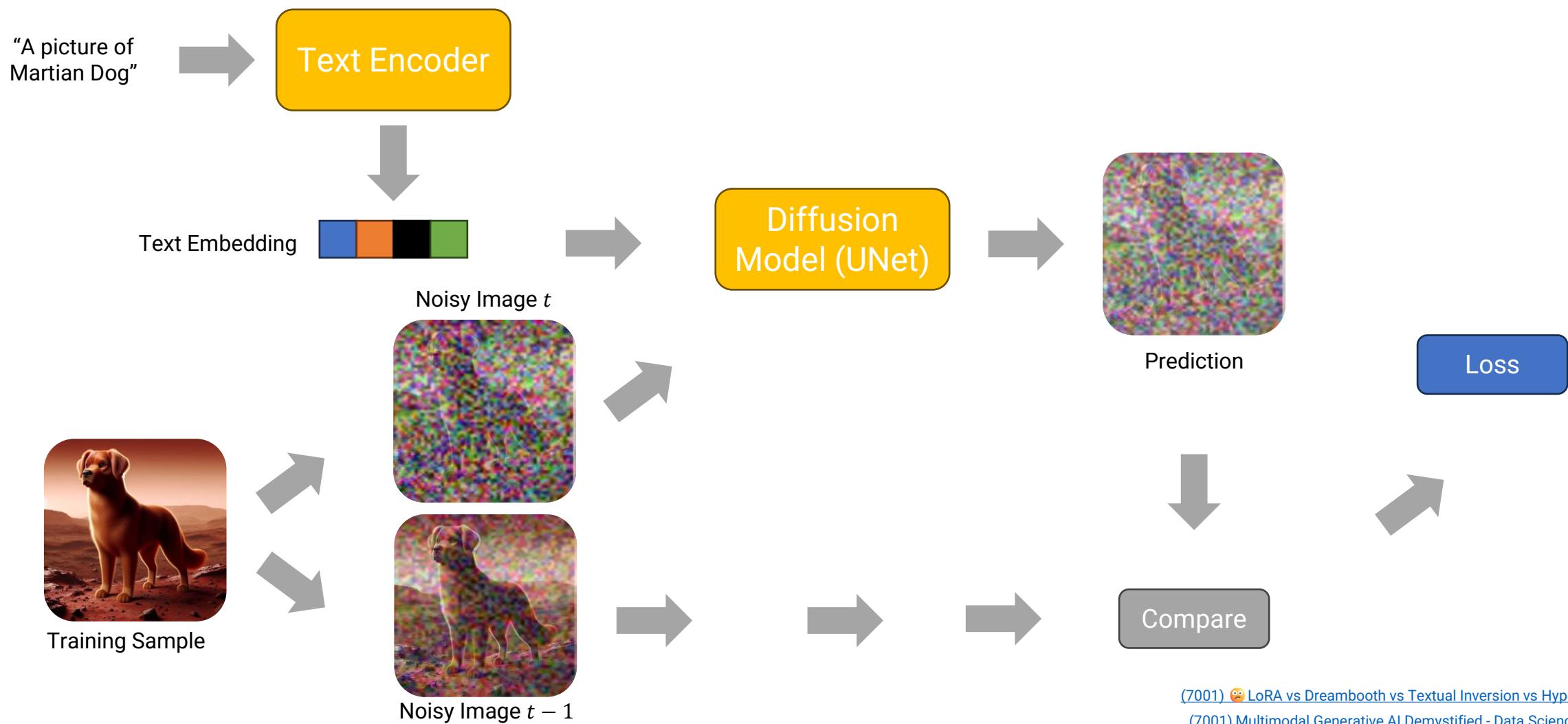


ChatGPT
Google



ChatGPT
Google

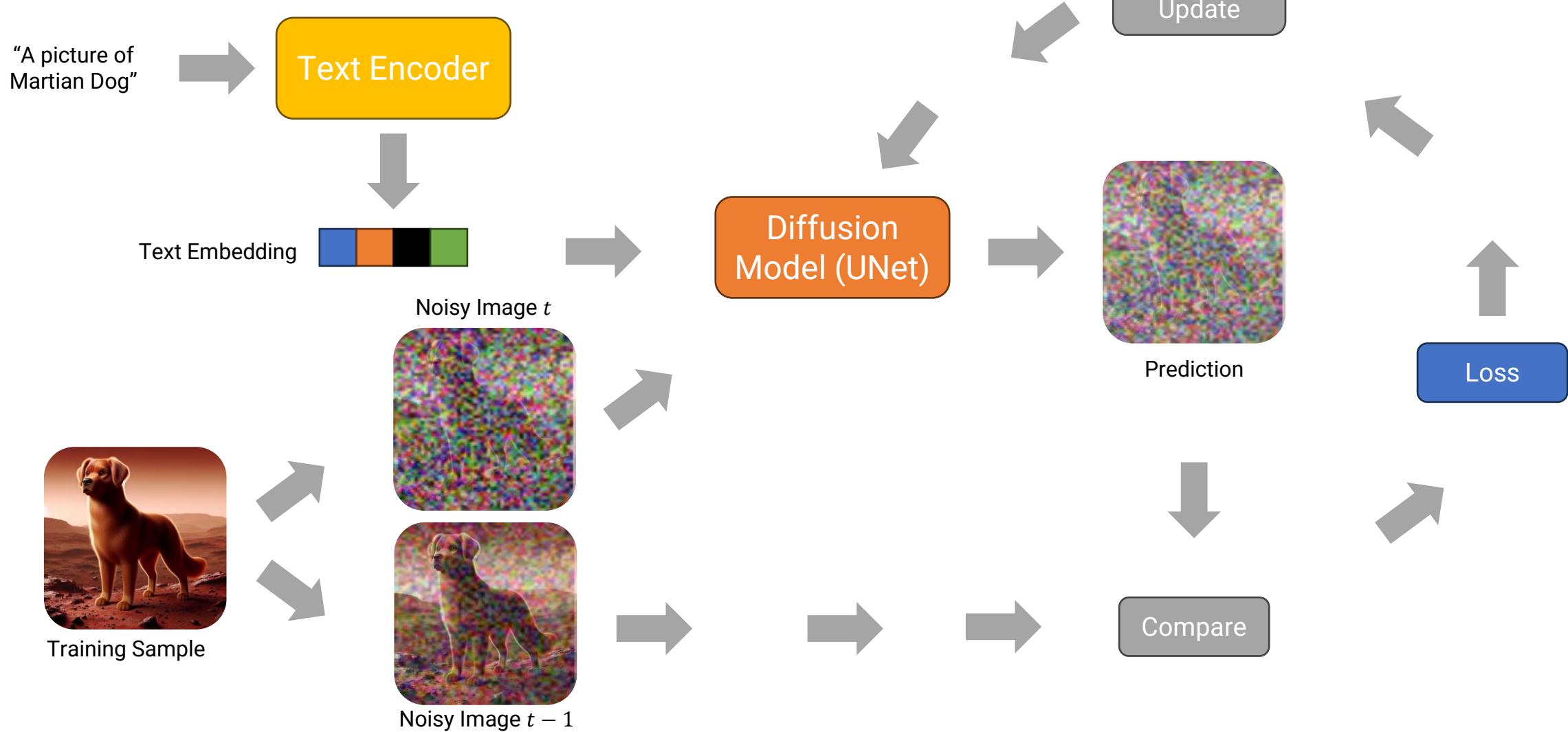
Dreambooth



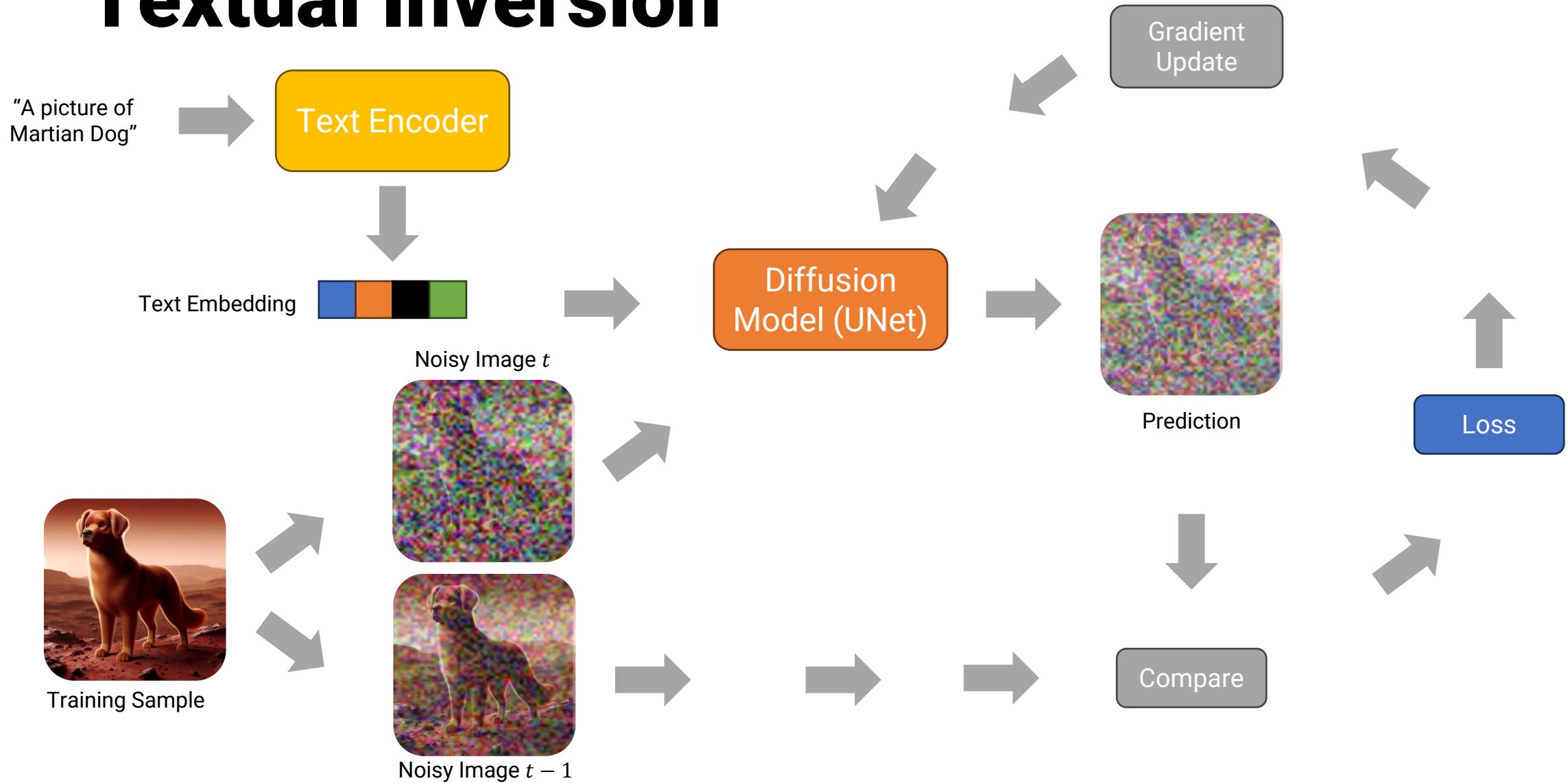
(7001) 🤖 LoRA vs Dreambooth vs Textual Inversion vs Hypernetworks - YouTube
(7001) Multimodal Generative AI Demystified - Data Science Festival - YouTube

- + Very effective
- Storage Inefficient

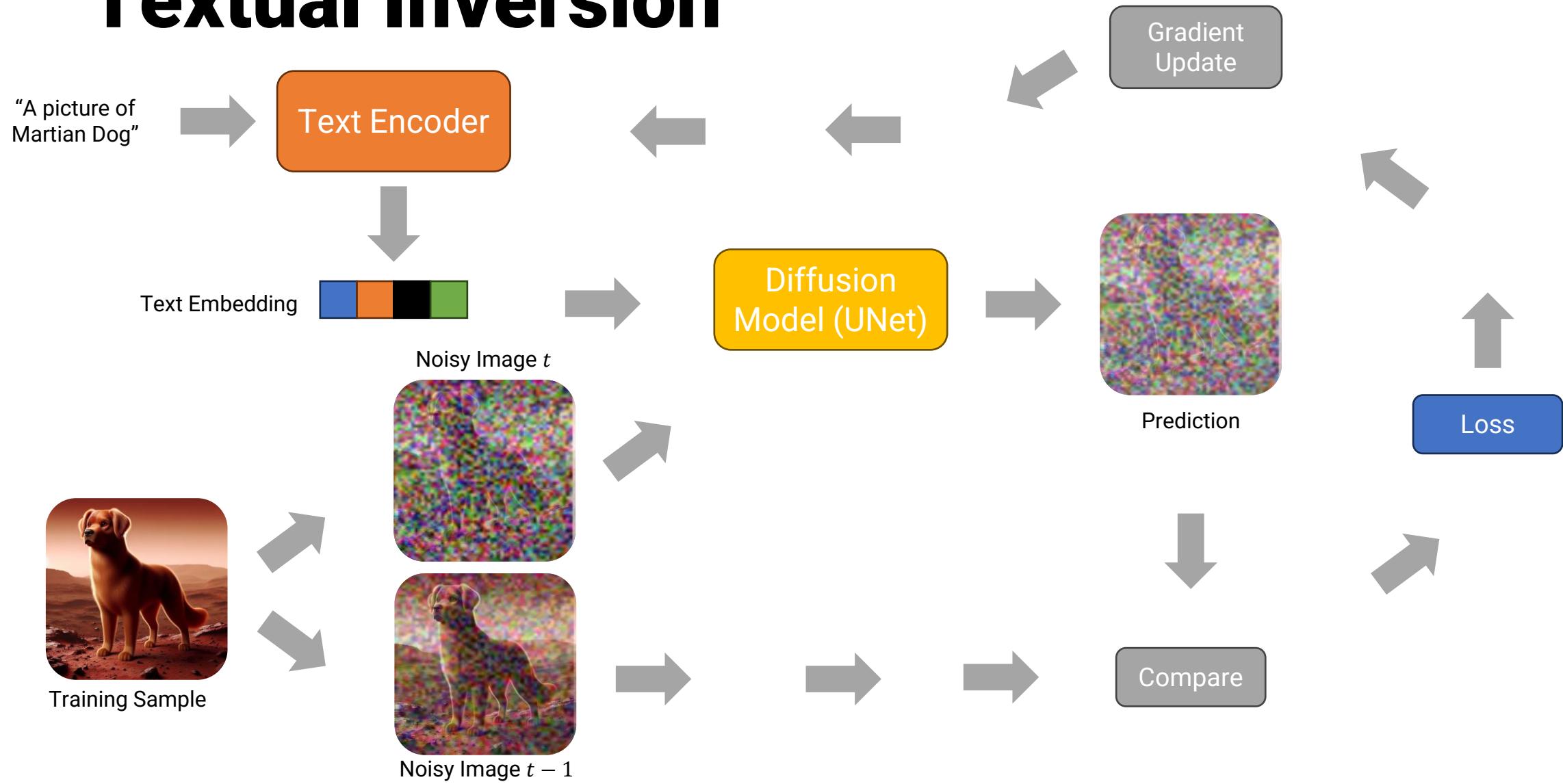
Dreambooth



Textual Inversion



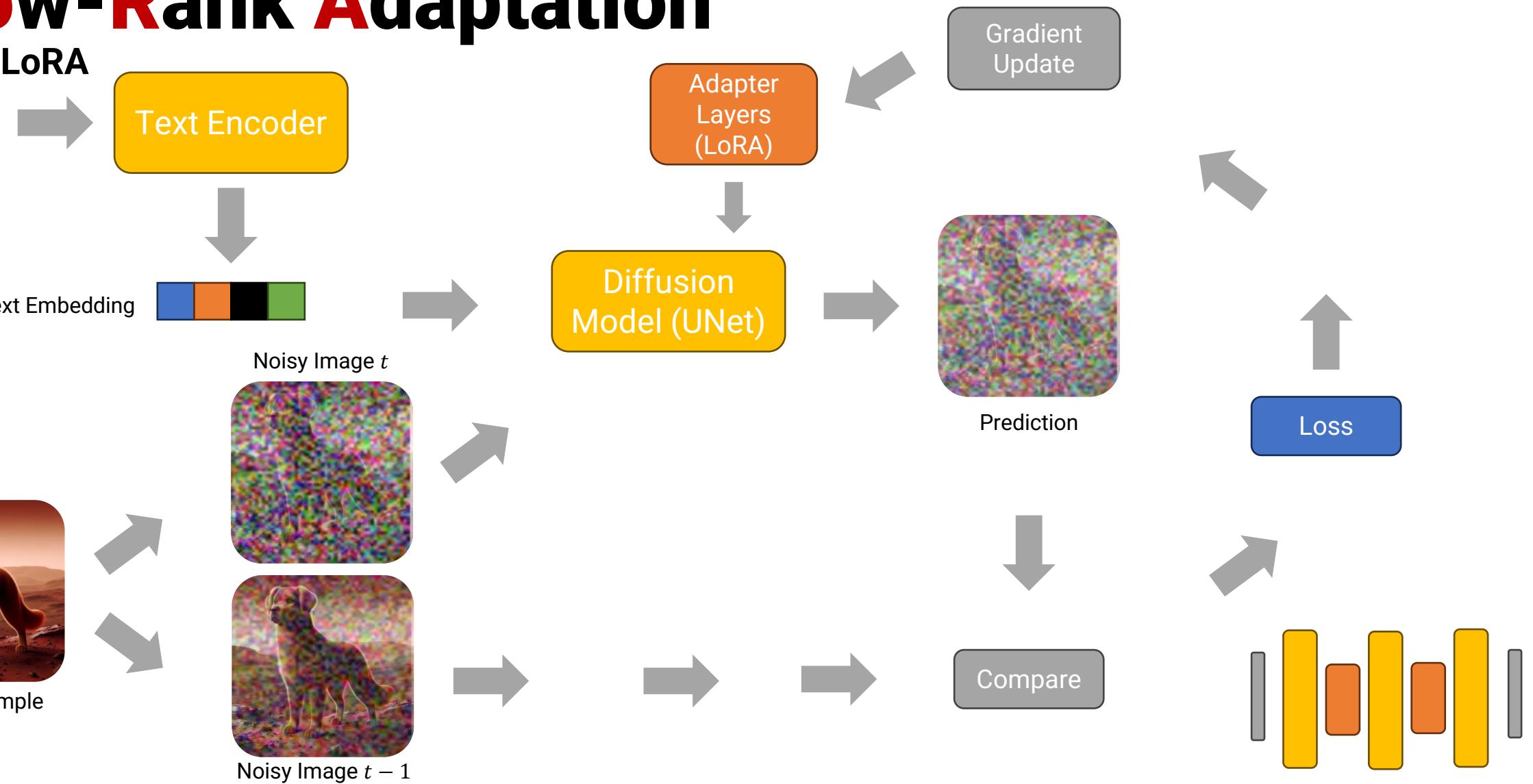
Textual Inversion



Low-Rank Adaptation

AKA LoRA

"A picture of
Martian Dog"



Large Models Are Data Gluttonous!



Where Does Data Come From?



Ekaterina Sirazitdinova



Biases In Data



`prompt = "a picture of young
ohwx woman, digital painting
by Atey Ghailan, cinematic
light, dark hair roots, blond
hair highlights, new york
background, looking in the
camera, hazel eyes"`

`negative_prompt = "low
quality, jpeg artifacts, ugly,
distorted, 3d, rendered,
cross-eyed, lazy eyes, long
hair, wavy hair, blue eyes,
signature, text"`



Biases In Data



New York Background



Chinatown Background



New York Background 0.5,
Chinatown Background 0.5



By Sandra Kuck



By Takeshi Koike

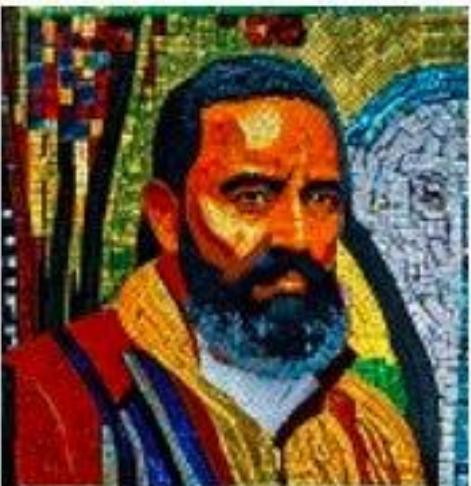


By Sandra Kuck 0.5,
By Takeshi Koike 0.5

(7001) Multimodal Generative AI Demystified - Data Science Festival - YouTube



Who Owns It?



mosaic art



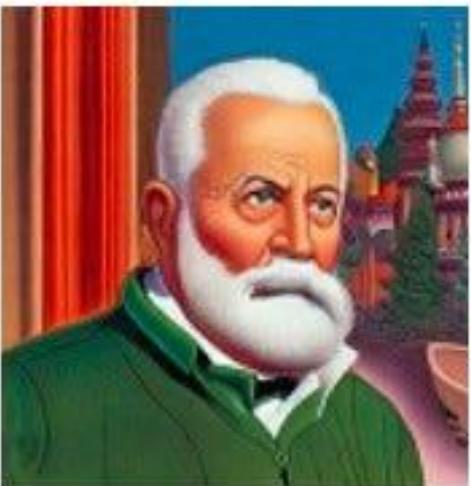
Harry Clarke



video game character



<belen> cartoon



cartoon by Carl Barks



<hanfu-anime-style>



<naoki_saito> anime

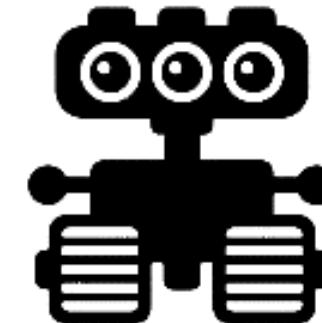
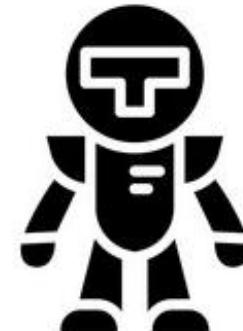
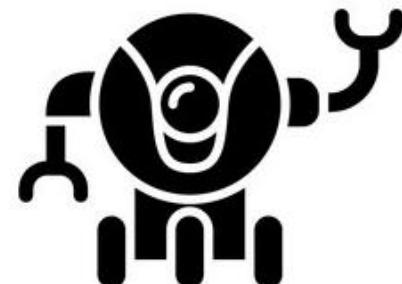
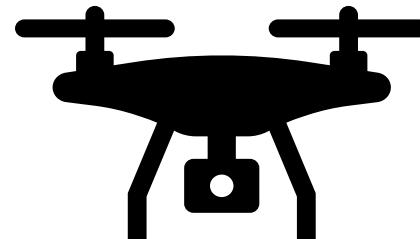
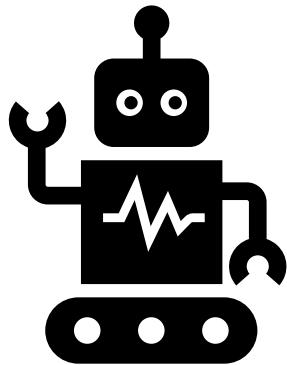


<pixelart-stardew>



Role Of Universities

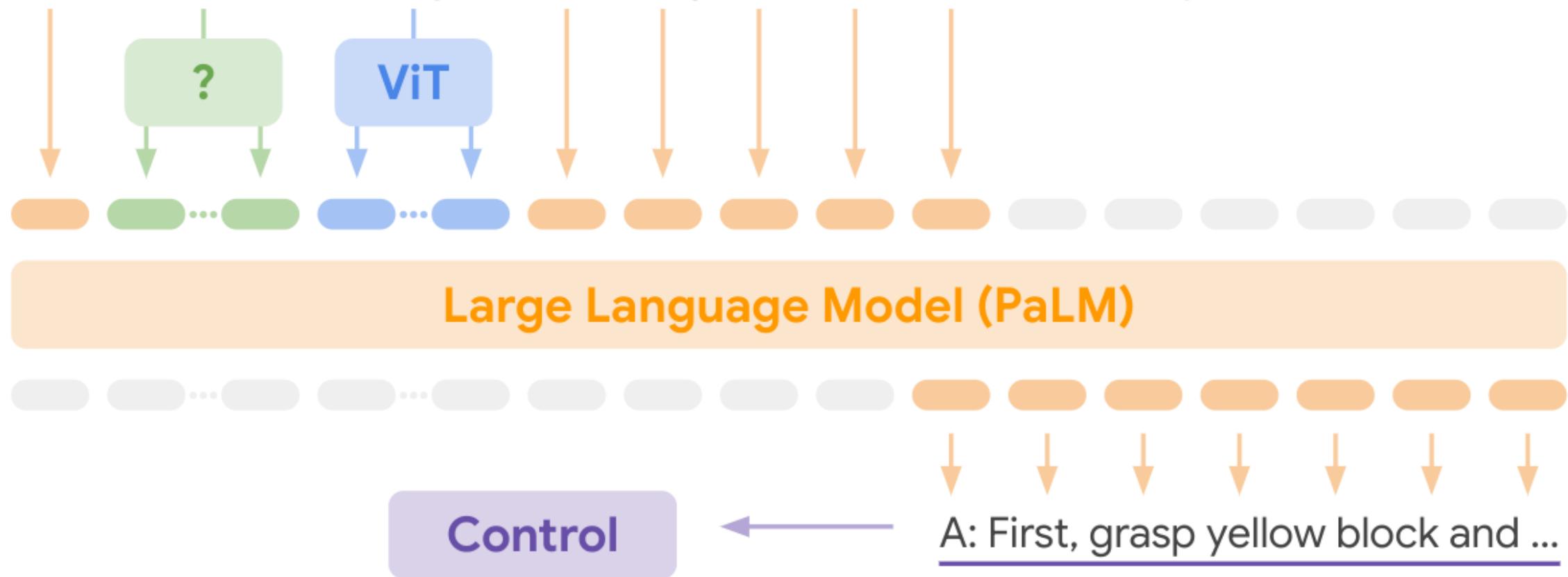
Robots Are Embodied Agents



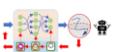
PaLM-E

PaLM-E: An Embodied Multimodal Language Model

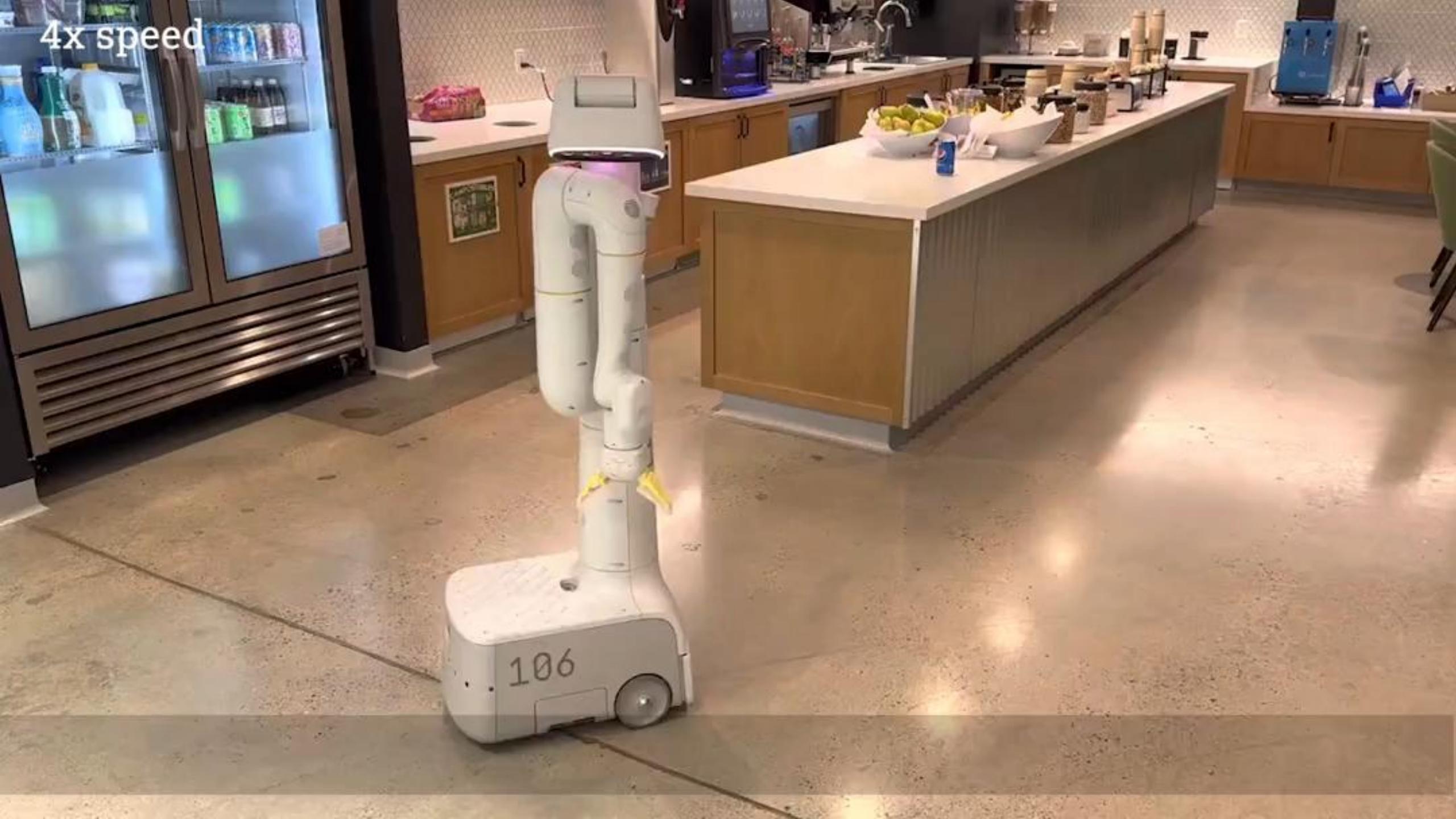
Given <emb> ... Q: How to grasp blue block? A: First, grasp yellow block



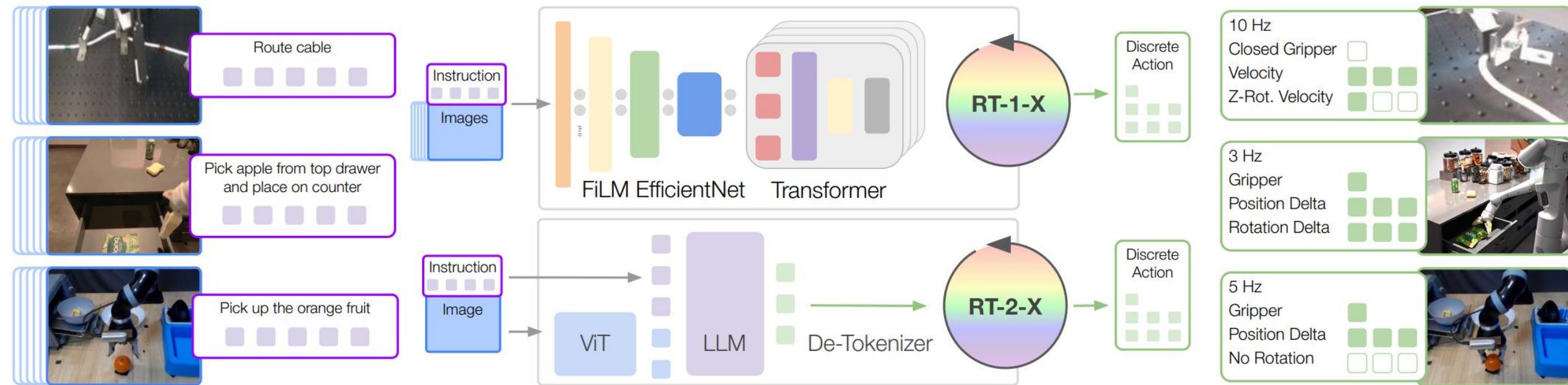
Driess, Danny, et al. "Palm-e: An embodied multimodal language model." arXiv preprint arXiv:2303.03378 (2023).



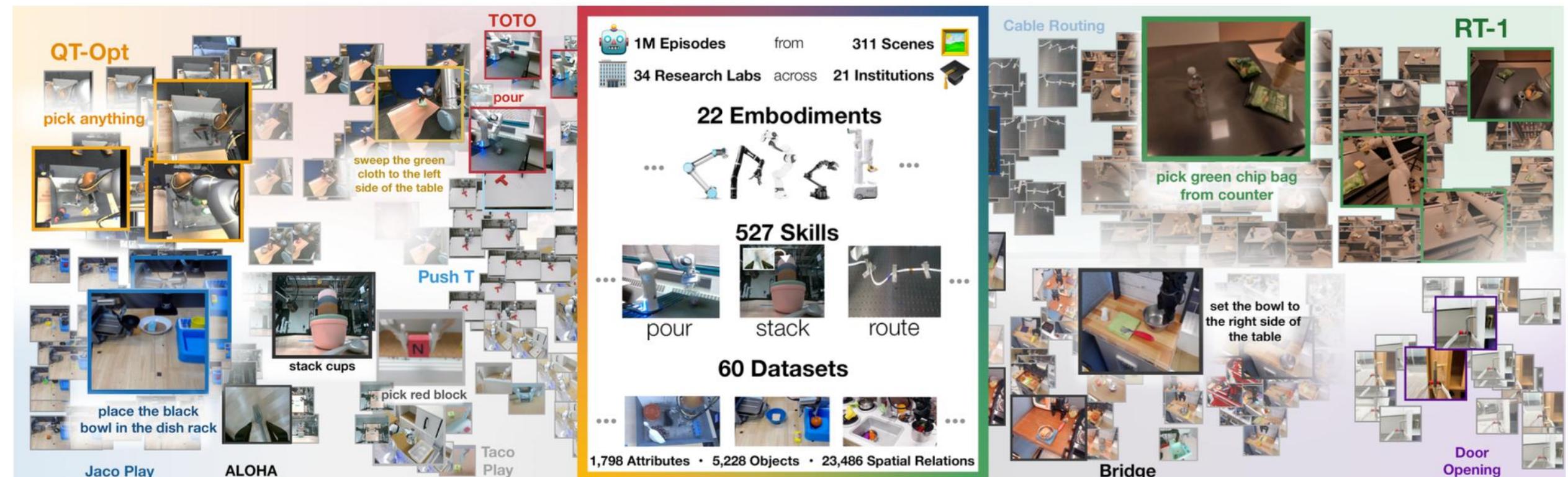
4x speed



Open X-Embodiment



Open X-Embodiment



RT-2-X (55B): one of the biggest models to date
performing unseen tasks in academic labs





Bring Back Your Heros

RIP Akira Toriyama



You can recreate Akira's Magic with GenAI Tools!



Deep Fakes



NEW

DeepFake Video

FaceFusion

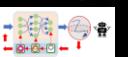


Best Digital Art

Catch was that it was AI generated!



People Have Begun to Sell Their Prompts for AI-Generated Artwork - ExtremeTech

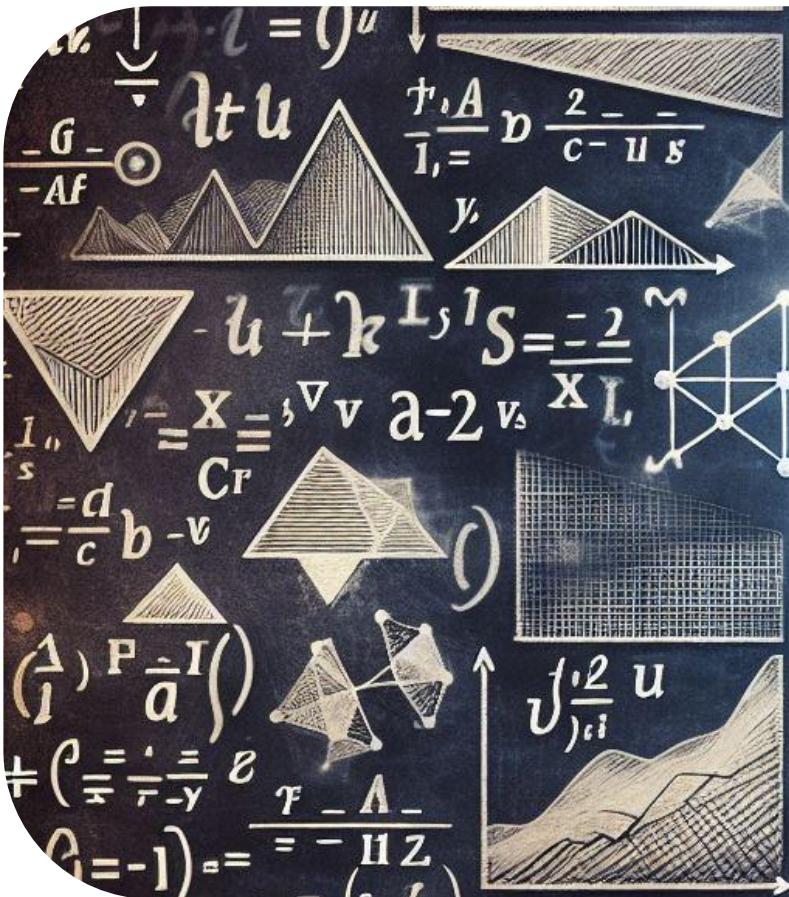


12/5/2025

80

WPI

Next Class!



Course Project Presentations