Multimodal Transformers

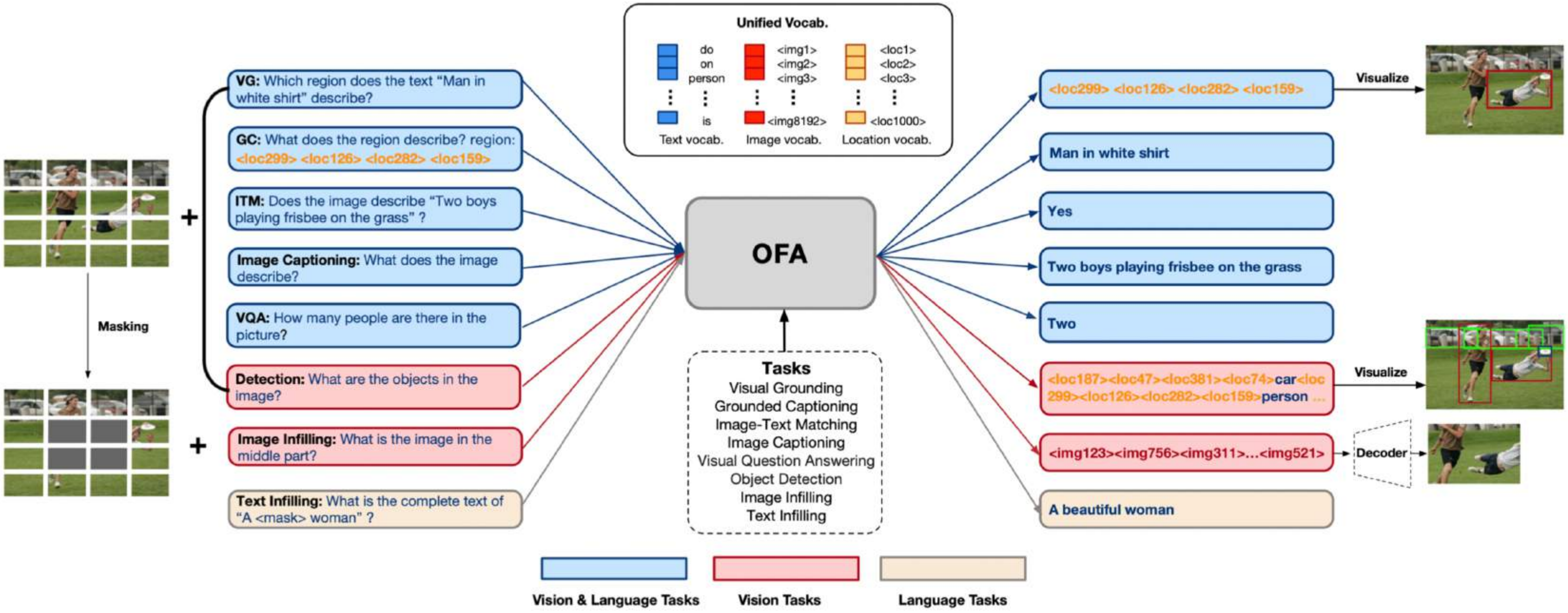
Anton Razzhigaev (@AbstractDL)

13. 04. 2023

Lecture Plan

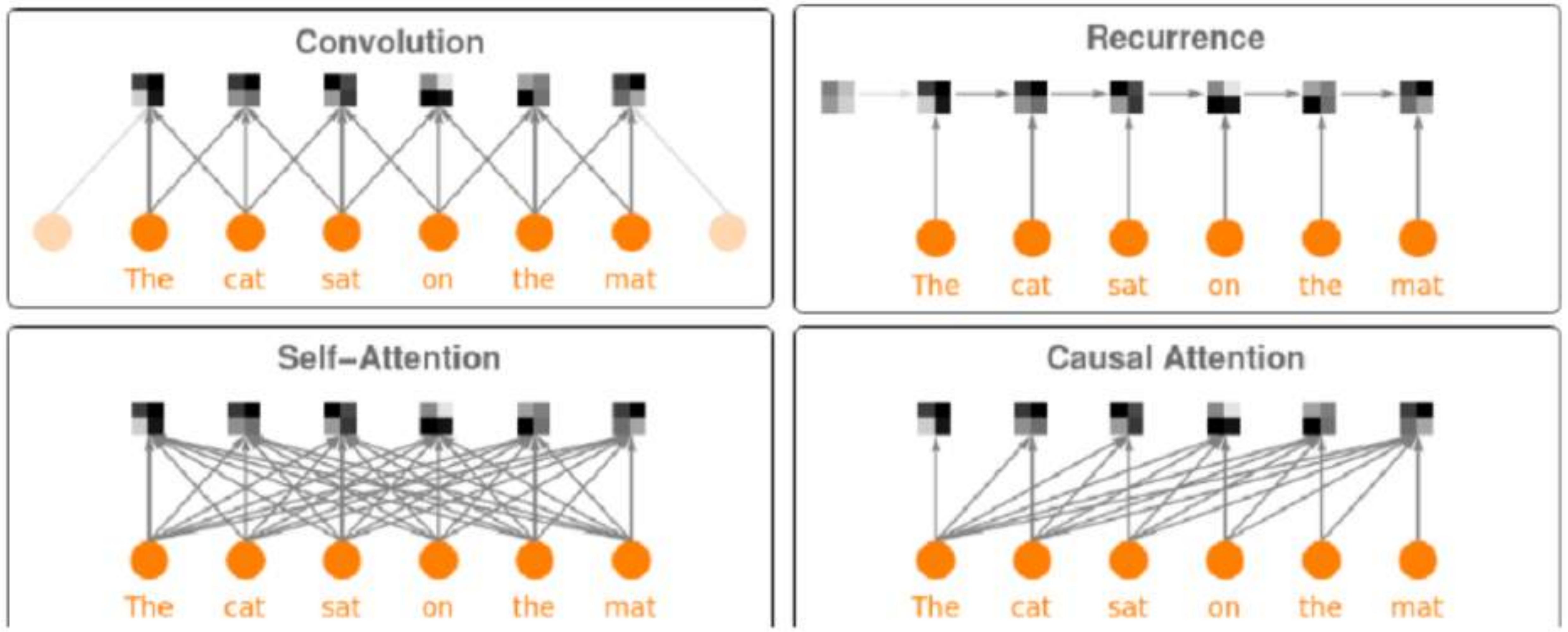
- Multimodality and inductive bias
- ViT, PIXEL, DINO, iGPT
- CLIP, DALL·E, VQ-VAE
- RuDolph
- Diffusion models
  - Dalle 2
  - Kandinsky 2.0, 2.1
- OFA
- Flamingo
- FROMAGe

Multimodality



OFA: UNIFYING ARCHITECTURES, TASKS, AND MODALITIES THROUGH A SIMPLE SEQUENCE-TO-SEQUENCE LEARNING FRAMEWORK

Inductive Bias



Inductive bias — it is a a-priory knowledge about the nature of data, which a human inserts in the ml model.

- CNNs have locality inductive bias.
- RNNs have sequential inductive bias.

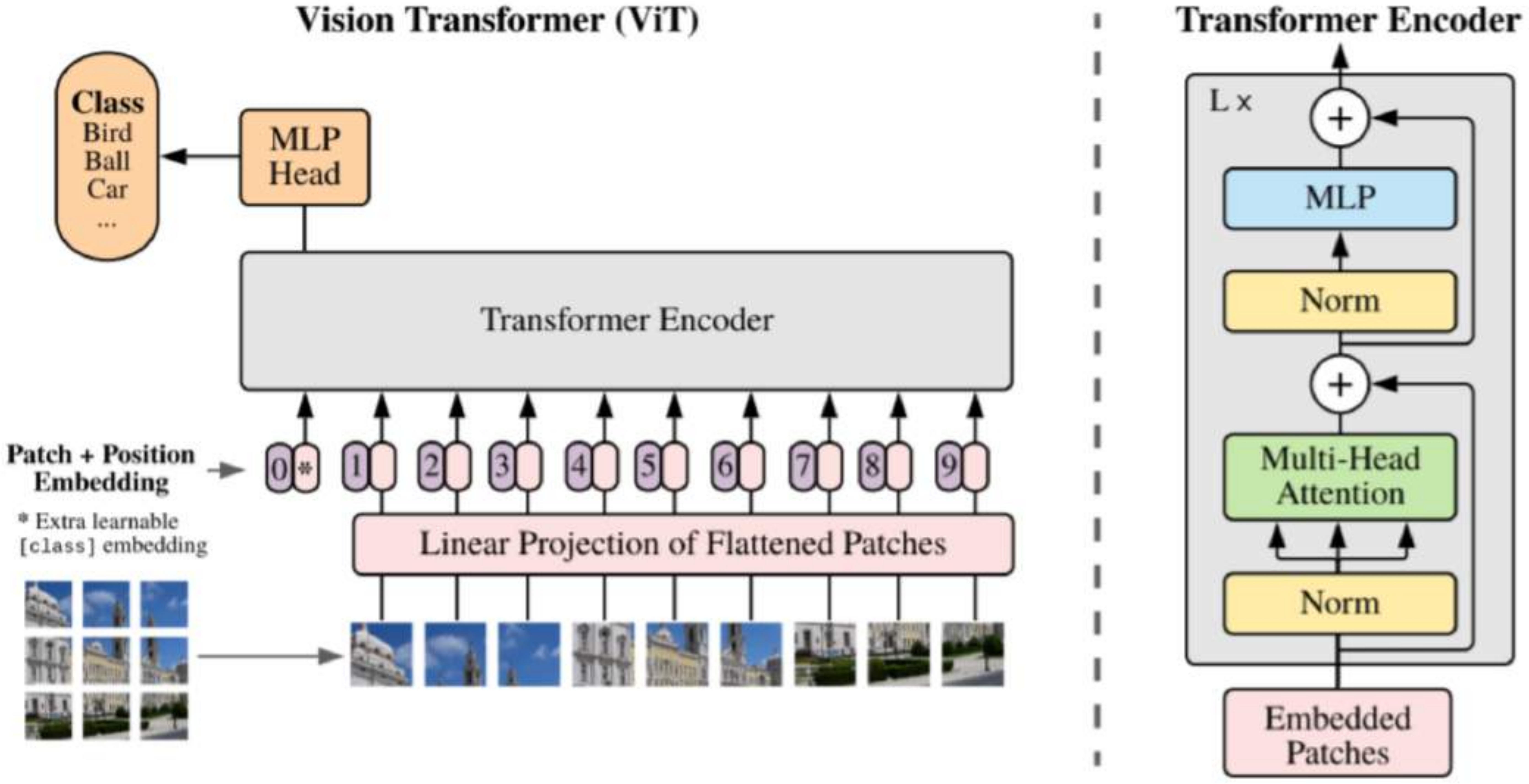
Strong inductive bias makes it easier to train the model. But models with strong inductive bias are less suitable for out of domain data (CNNs work not very well with texts).

In case we have **large enough datasets** or, different modalities **it is better to use weak inductive bias**, like fully-connected architectures, or transformers.

That is why **transformers are more flexible** and demonstrate better performance, but require much more data to be trained.

Visual Tranformers

ViT



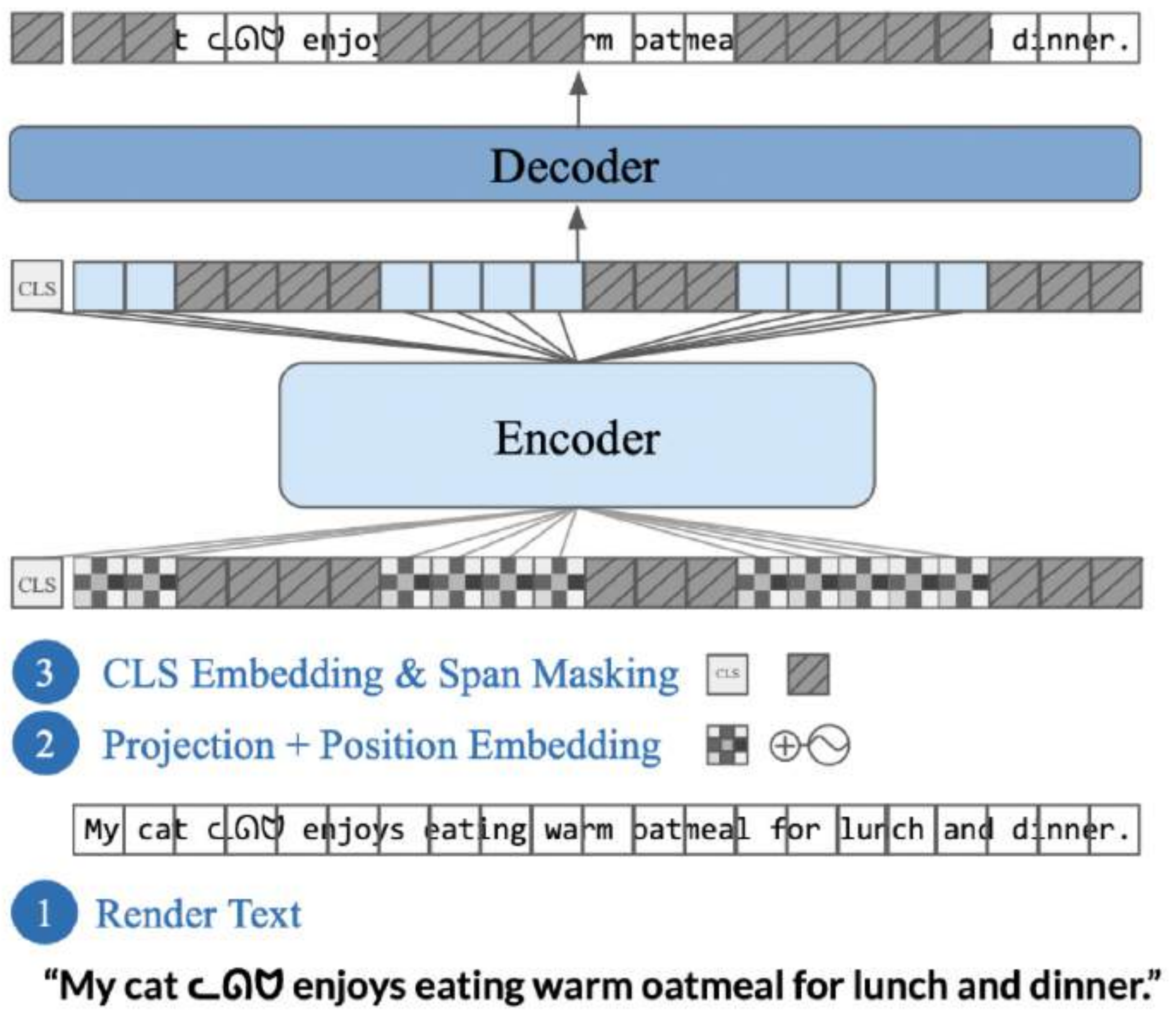
Attention-based alternative to CNN-resnets:

- **Full-image receptive field**
- **Weak inductive bias**
- **Better performance (on large data)**
- **More flexible representations**

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

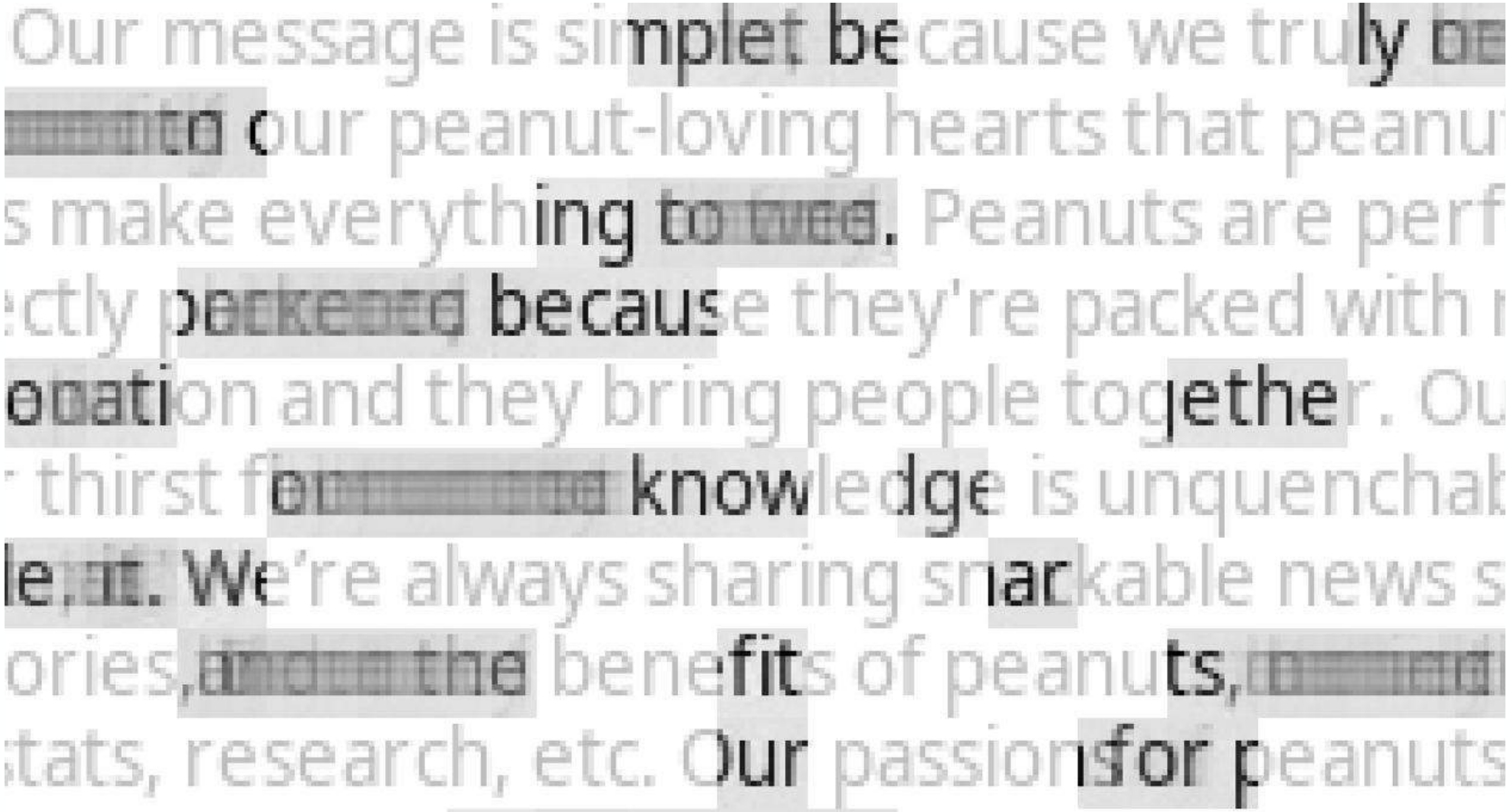
PIXEL

- Character pixels instead of text tokens
- Masked LM over «screenshots»
- BERT-like architecture Understands
- ᗪEE尸 ㄥE闩尺𝒩ｉ𝒩G - can understand this text
- More robust to adversarial attacks



(a) PIXEL pretraining

PIXEL


Our message is simple because we truly
to our peanut-loving hearts that peanu
s make everything to . Peanuts are perf
ectly because they're packed with
ation and they bring people together. Ou
r thirst for knowledge is unquenchab
le. We're always sharing snackable news s
ories, the benefits of peanuts,
stats, research, etc. Our passion for peanuts

DINO

- ViT architecture
- Self-supervised objective
- Attention maps work as unsupervised segmentation

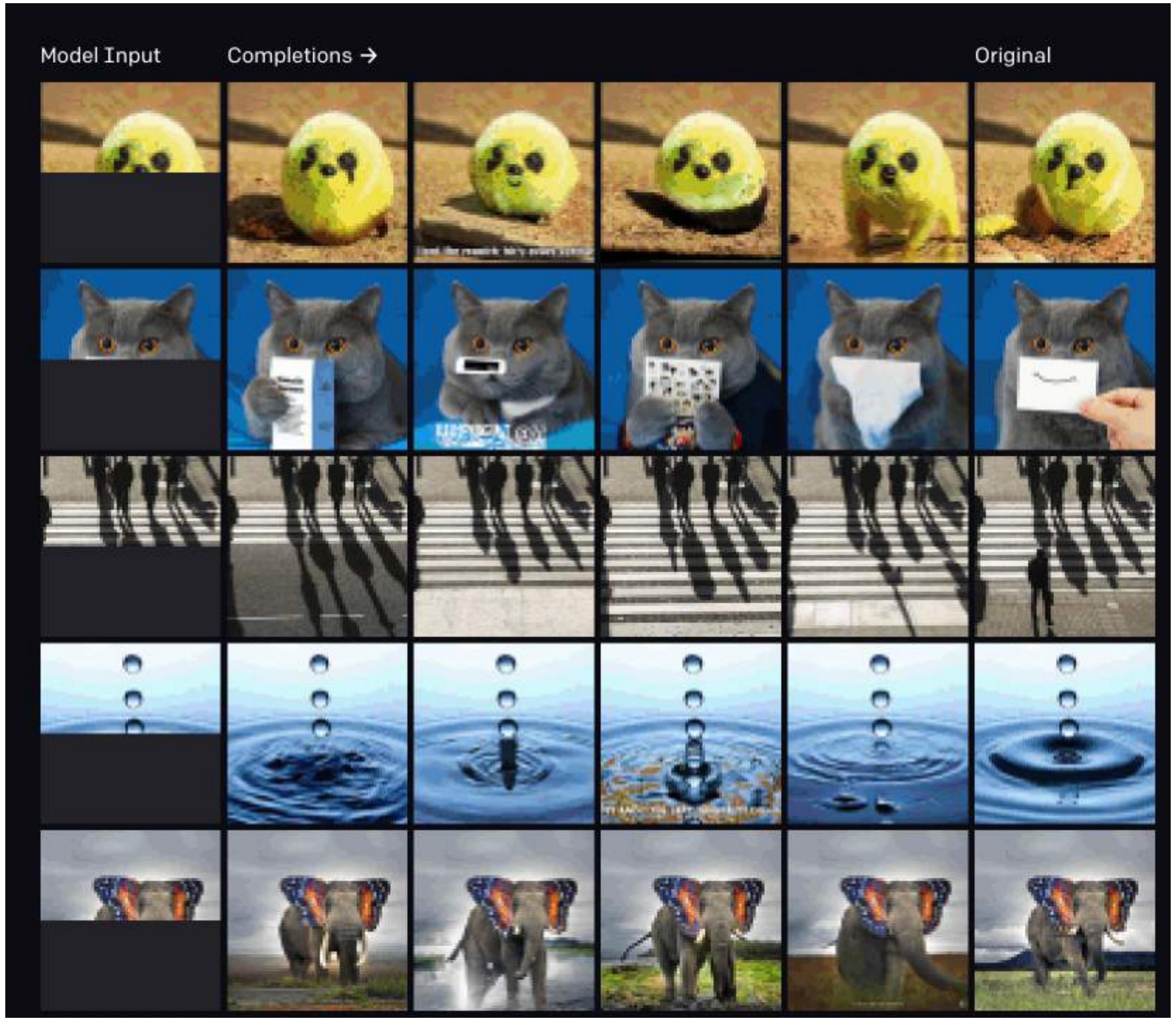
Emerging Properties in Self-Supervised Vision Transformers

DINO

- Two networks: student and teacher
- Random crops of an image go to teacher an student
- CrossEntropy Loss between outputs of student and teacher
- Teacher = exp_avg(Student)

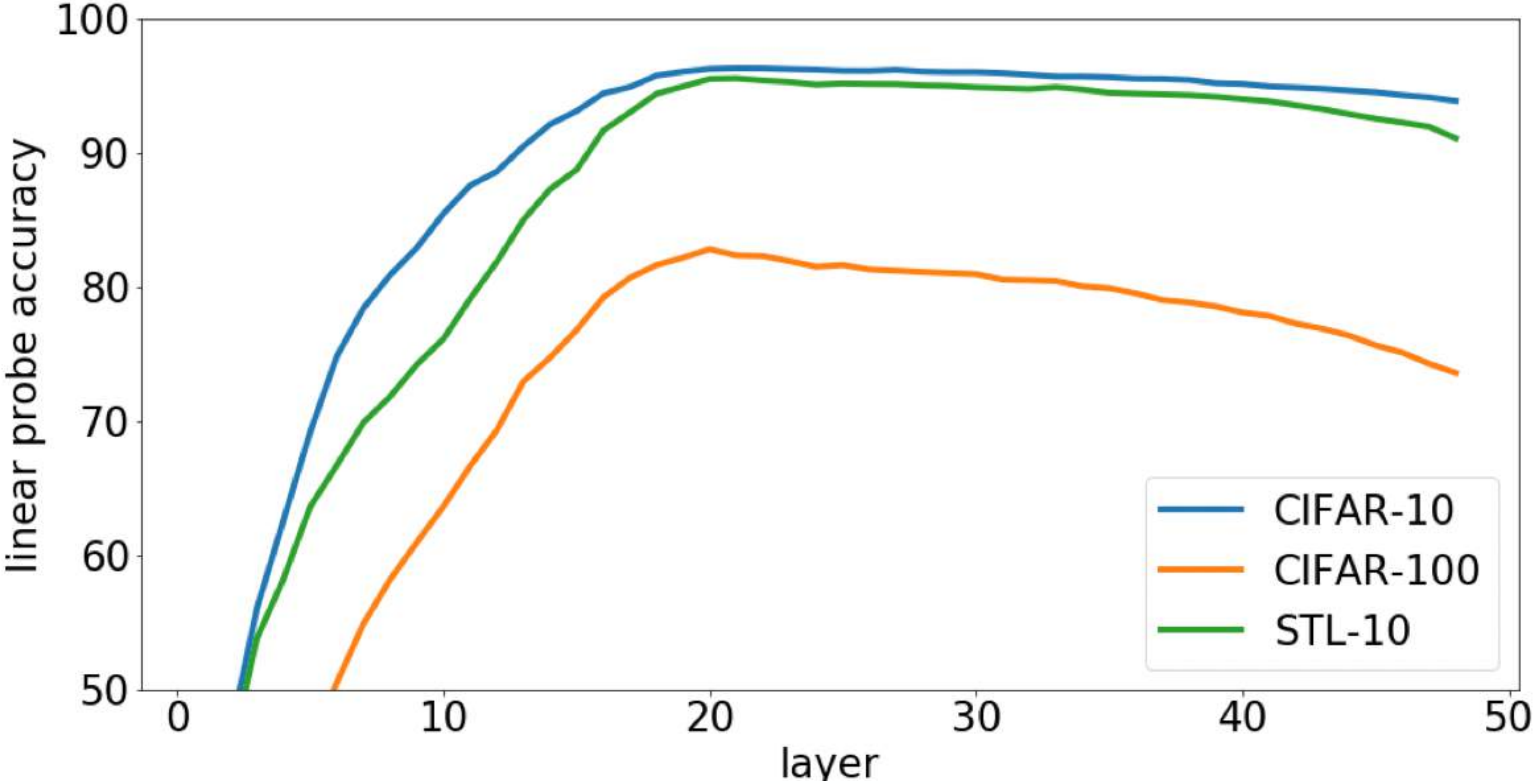
iGPT

- The same architecture as GPT2
- Pretreining over pixel values with cross-entropy
- Can complete images and generate from scratch
- Embeddings can be used for downstream tasks

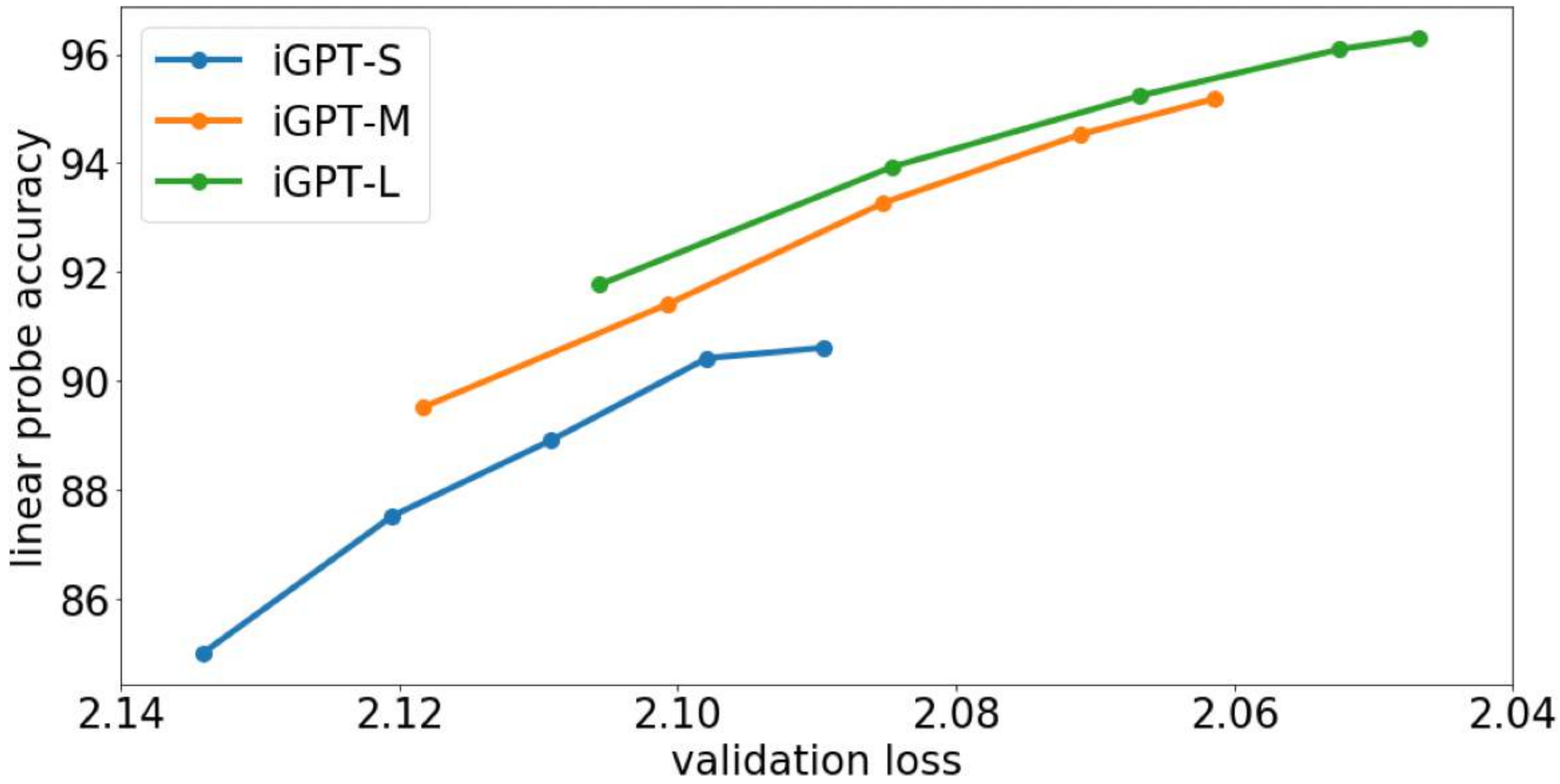


Generative Pretraining from Pixels

iGPT: per layer linear probe



iGPT: the larger the better



Multimodal Tranformers

CLIP and Dall·E

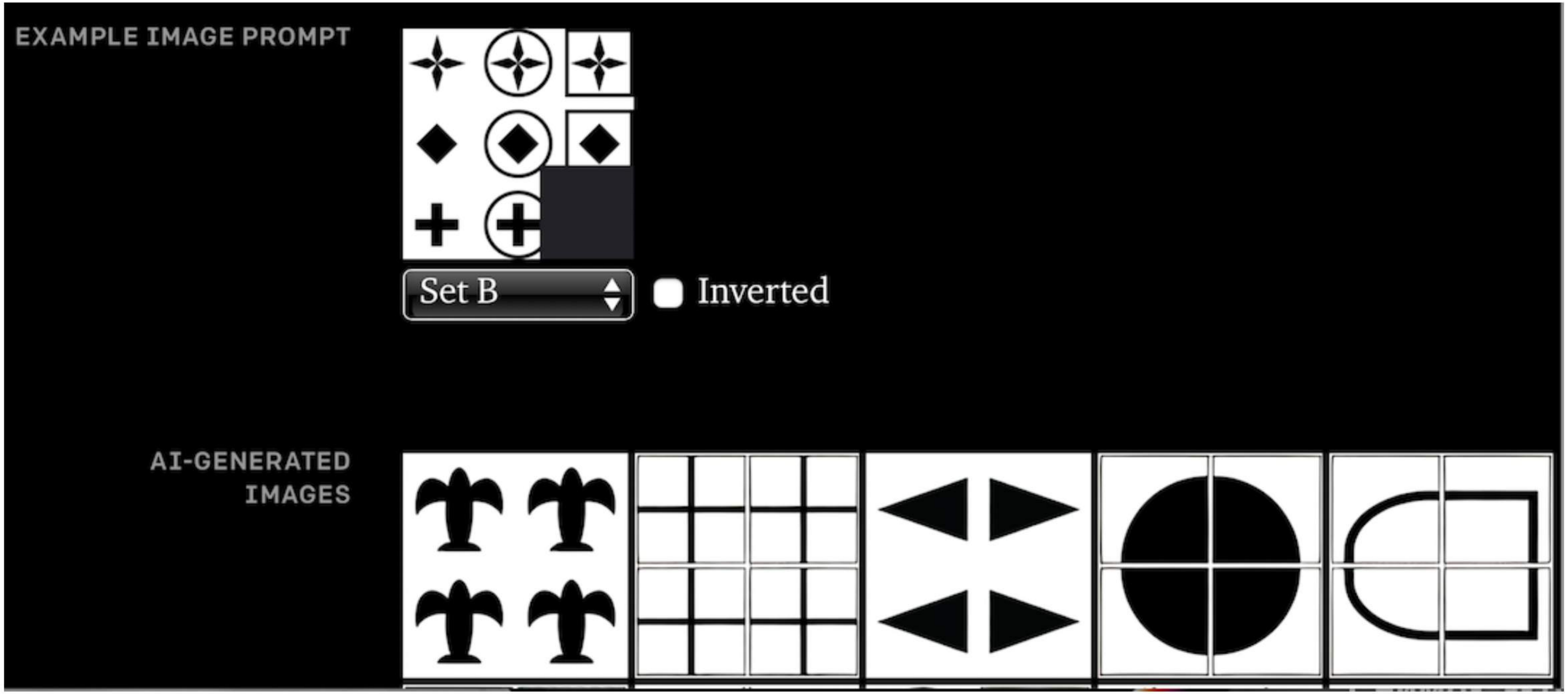
a pentagonal green clock. a green clock in the shape of a pentagon.

a cube made of porcupine. a cube with the texture of a porcupine.

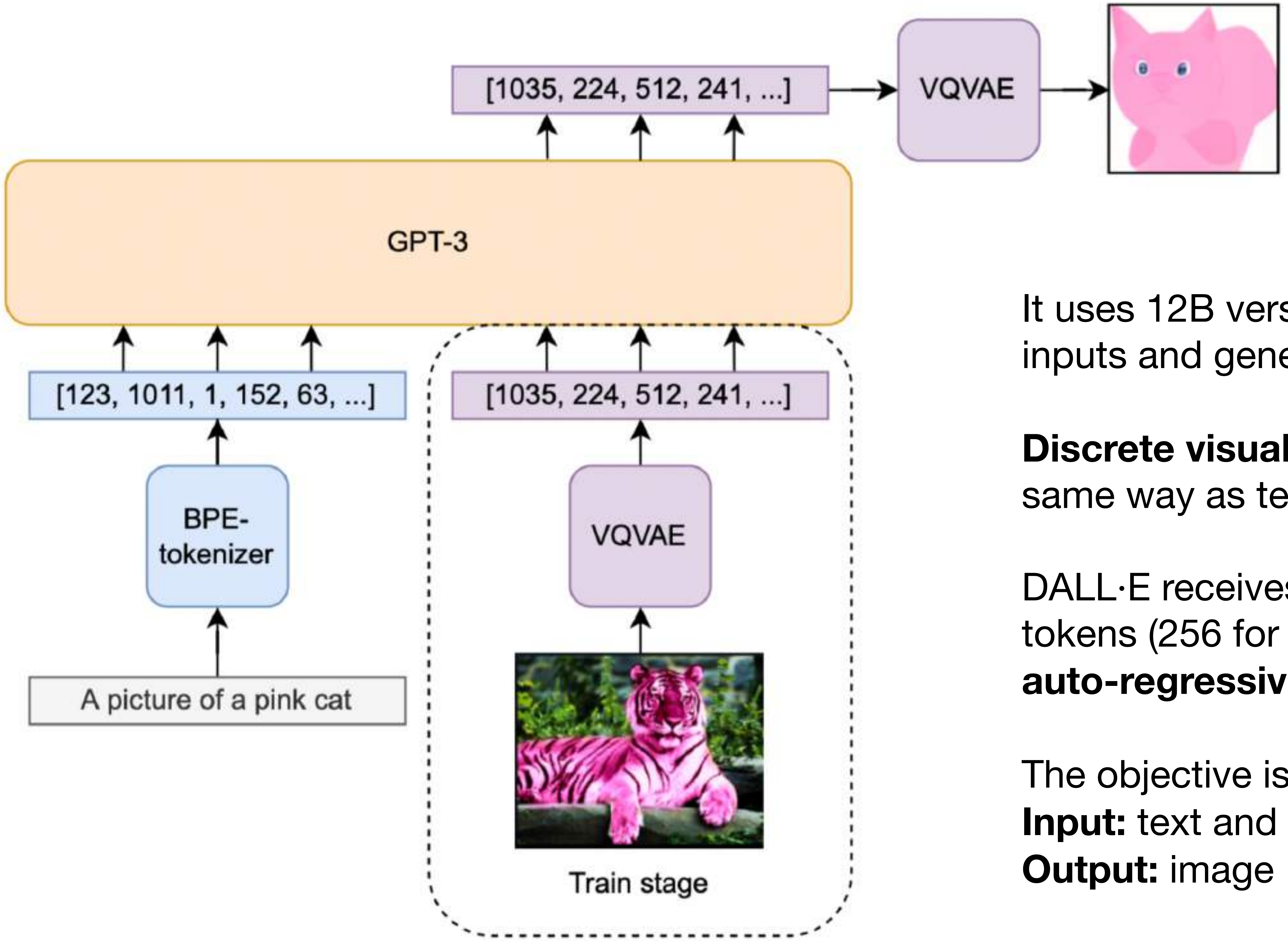
a collection of glasses is sitting on a table

Zero-Shot Text-to-Image Generation

Visual Understanding of Dall·E



DALL·E



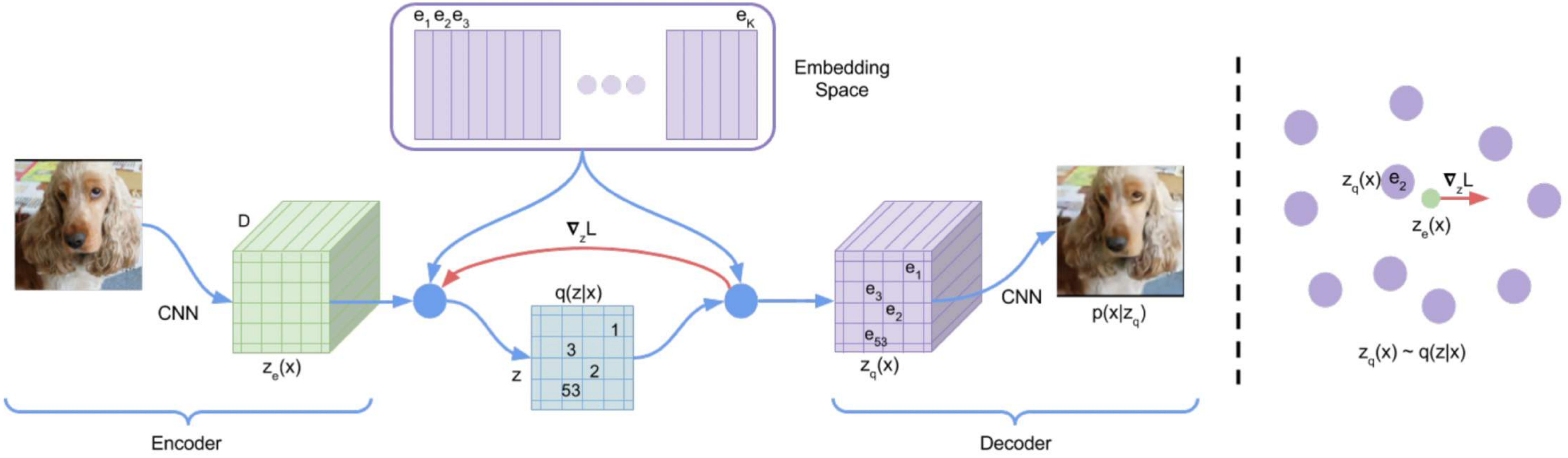
It uses 12B version of the **GPT-3 model to interpret natural language** inputs and generate corresponding images.

Discrete visual features from VQVAE are used as visual tokens in a same way as text tokens, which then can be decoded back to images.

DALL·E receives both the text and the image as a single stream of 1280 tokens (256 for the text and 1024 for the image) and models all of them **auto-regressively**.

The objective is a simple **cross-entropy loss**.
Input: text and (optionally) part of an image
Output: image

VQ-VAE

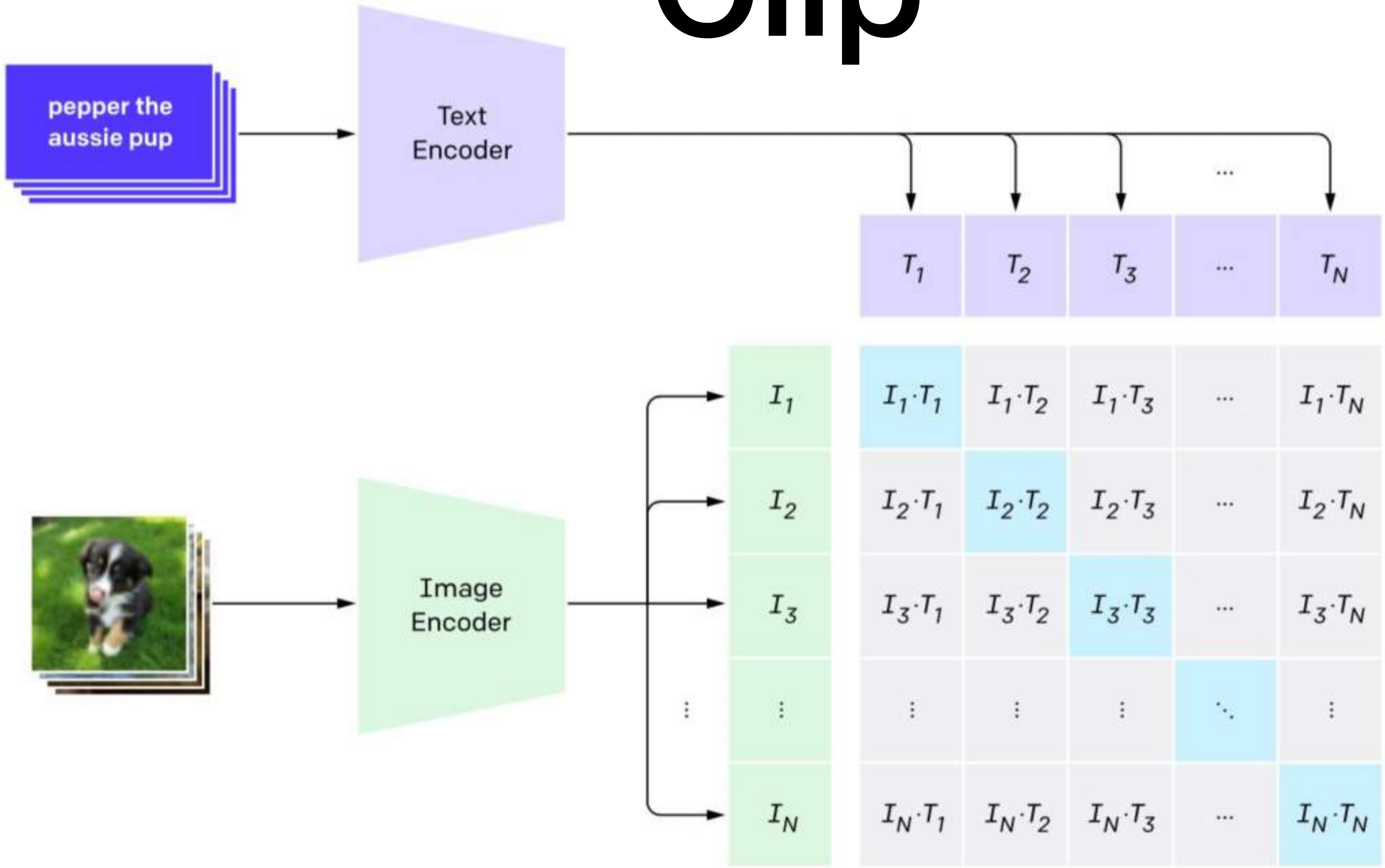


The very important part of DALL·E is **image tokenizer** — the part of the model which transforms an image from pixels to a list of discrete tokens.

It is a usual practice to use **VQVAE — Vector Quantized AutoEncoder** — a special typer of autoencoders, which use discrete latent space (a kind of quantized embeddings).

Generating Diverse High-Resolution Images with VQ-VAE

Clip



The idea is pretty simple: two encoders for text and images which provide similar embeddings for images and their descriptions. It is pretrained on a large dataset of image and captions with contrastive loss.

Input: image or text
Output: embedding

Learning Transferable Visual Models From Natural Language Supervision

Clip Abstractions

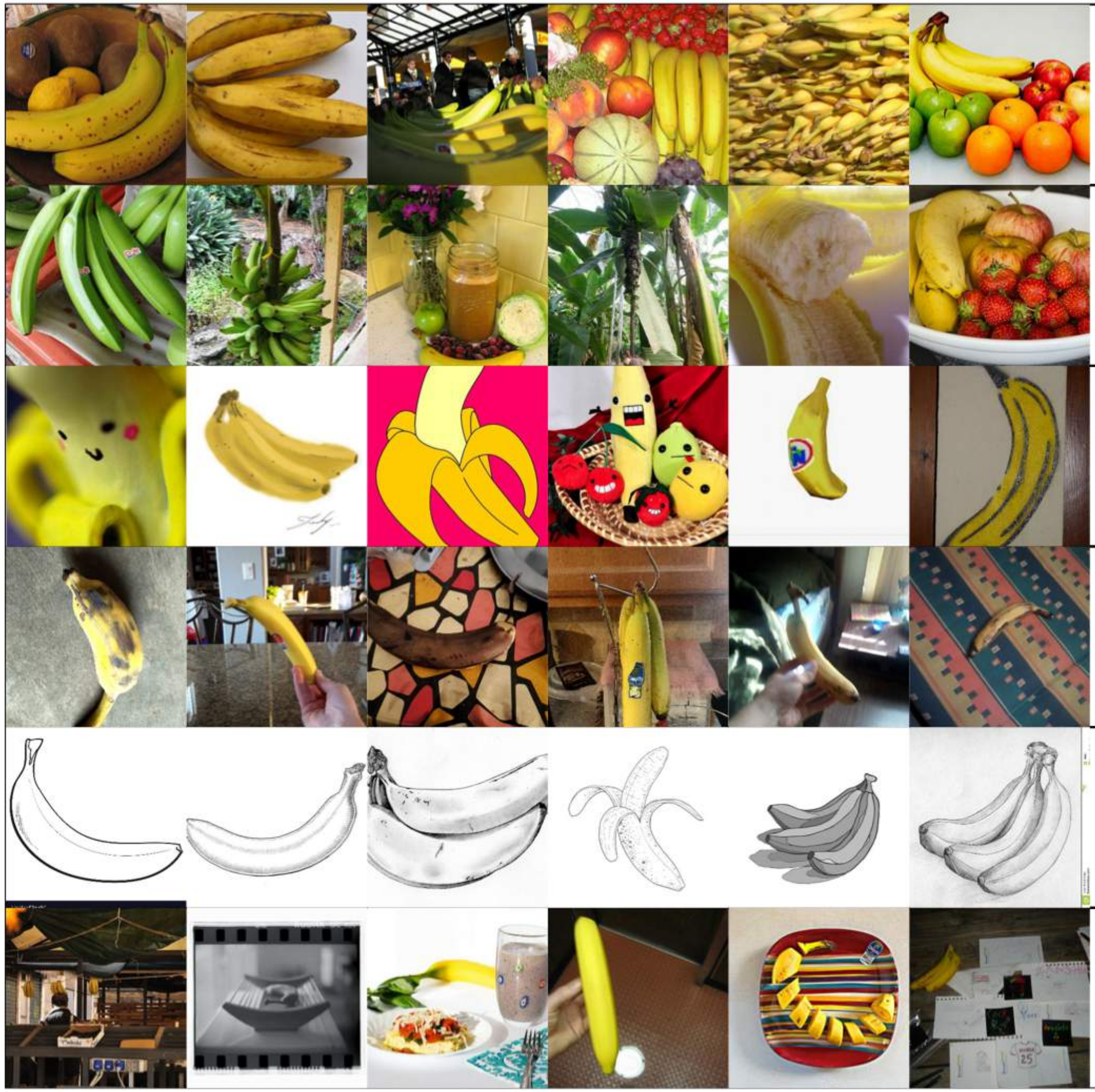
It understands **open-set visual concepts** from natural language and demonstrate unbelievable generalization abilities!

CLIP even understands **high levels of abstractions and implicit relations between them**. Like in the picture with reversed emotions.

Multimodal Neurons in Artificial Neural Networks

iPod

Dataset Examples



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

Zero-shot Applications

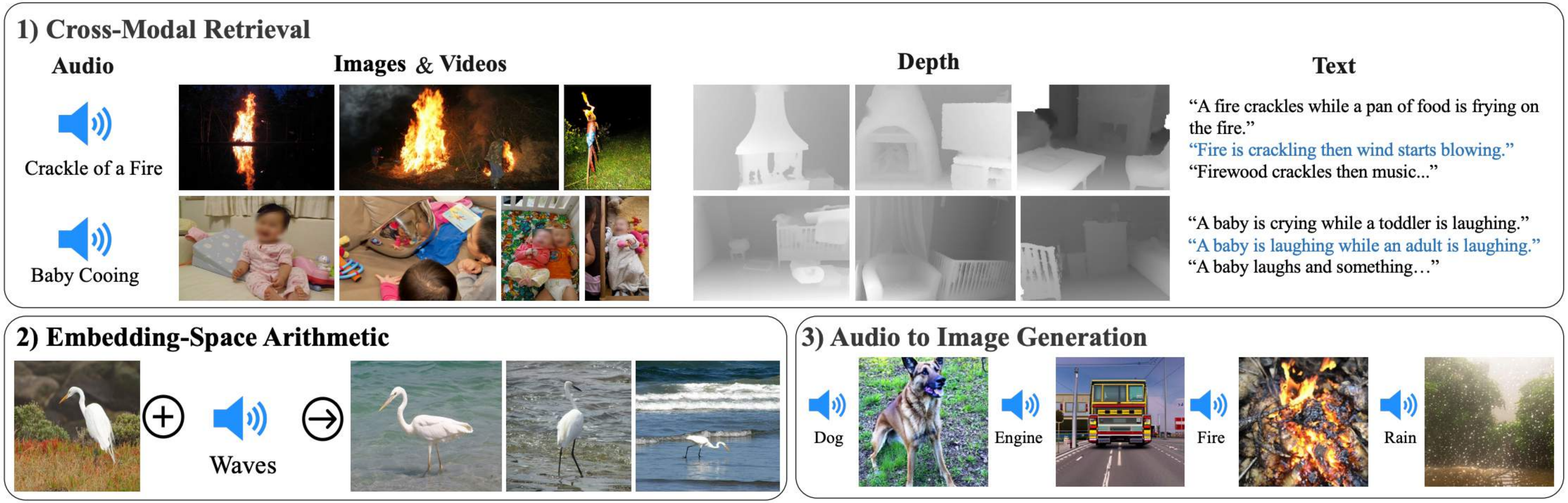
query: "Кто съел всю колбасу?"

https://t.me/abstractDL/92

CLIP can be used for:

- classification
- object detection
- Visual-language salience
- search
- image reranking
- …

ImageBind — CLIP for 7 modalities



IMAGEBIND: One Embedding Space To Bind Them All

RuDolph

It is a **hypermodal neural network** which works **similar to DALL·E**, but more flexible and it **can also generate texts**.

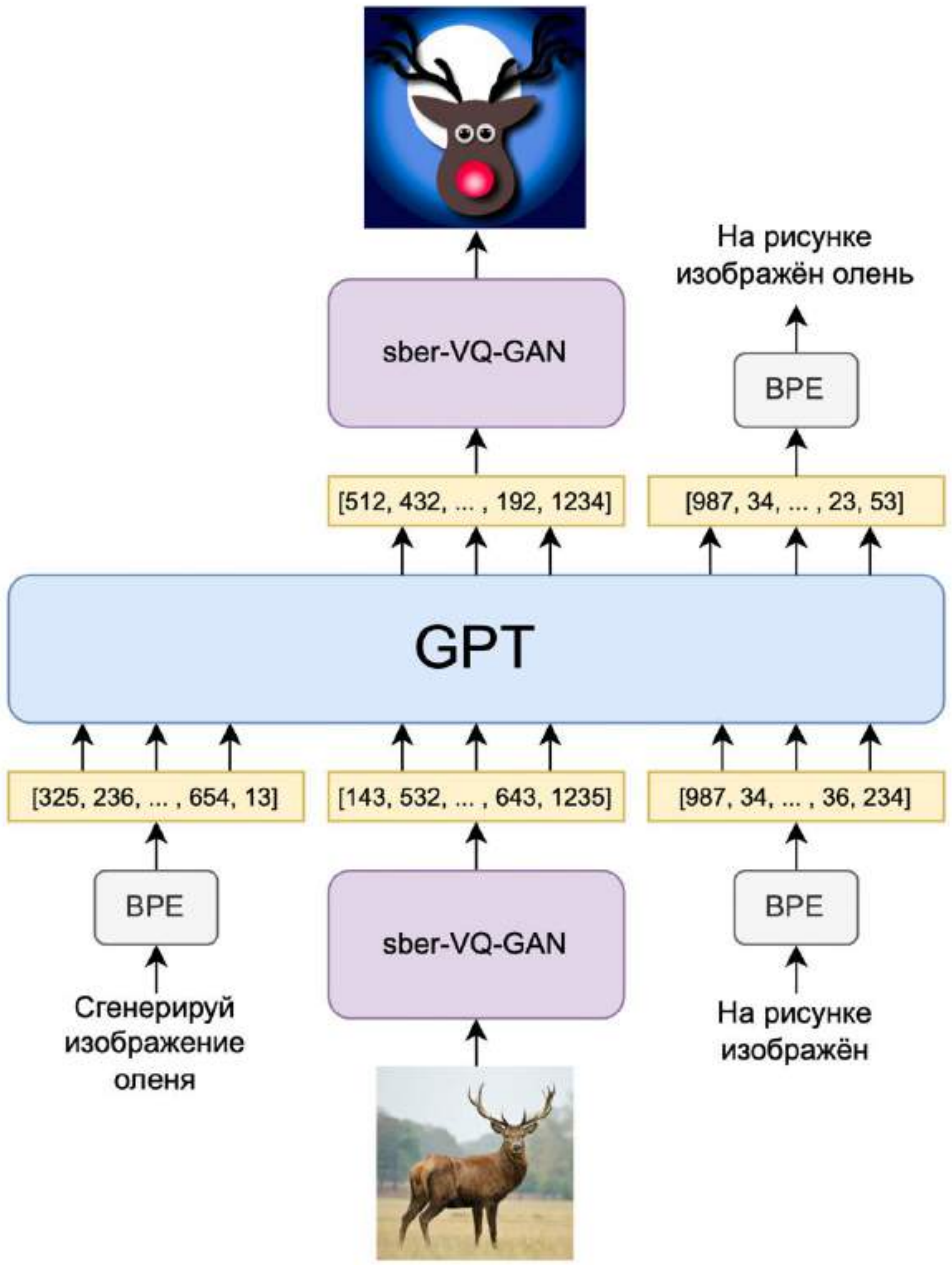
Developed by **SberAI**.

In contrast to DALL·E it has **right and left text contexts:**

- the left one is used for **image generation** (image is in between two contexts)
- the right one is used for **image captioning**.

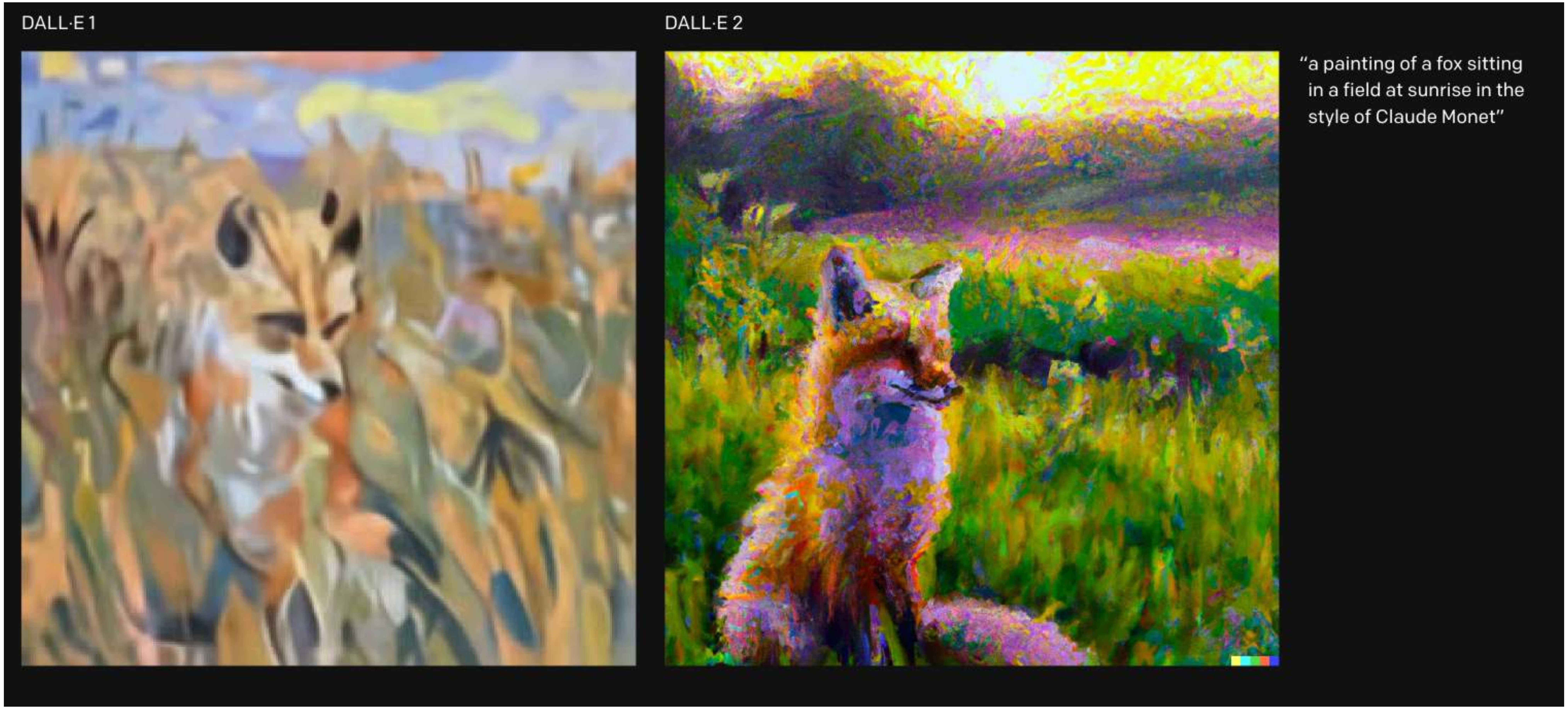
During training this two contexts and tasks alternate each other and the model is trained with **cross-entropy loss.**

Input: text or image
Output: image of text



RuDolph

DallE·2



The architecture is absolutely different: now it is a diffusion model conditioned on CLIP embeddings.
NO AUTOREGRESSION

Hierarchical Text-Conditional Image Generation with CLIP Latents

Diffusion

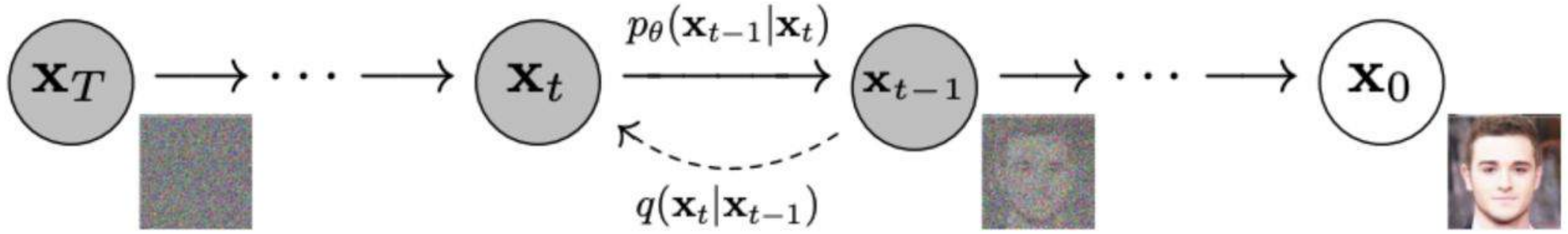
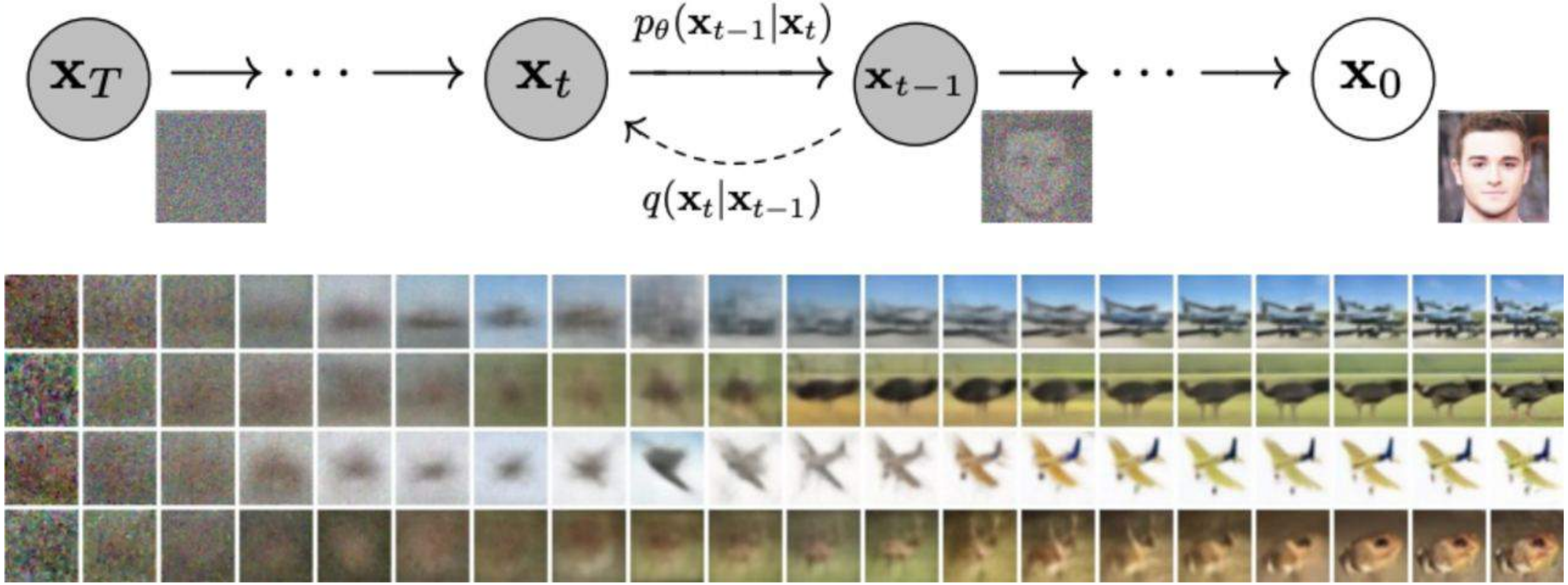


Image Super-Resolution via Iterative Refinement

Diffusion



Diffusion Models Beat GANs on Image Synthesis

It also can do in-painting and even zooming-out (video)

https://t.me/too_motion/455

Kandinsky 2.0

- Based on **Latent Diffusion** — diffusion process in embedding space of KL-VAE
- **Multilingual** — understands more than 100 languages
- Developed by AIRI, SberAI, SberCloud
- Fully open-sourced

Железный человек on the Moon 背景中的烟花

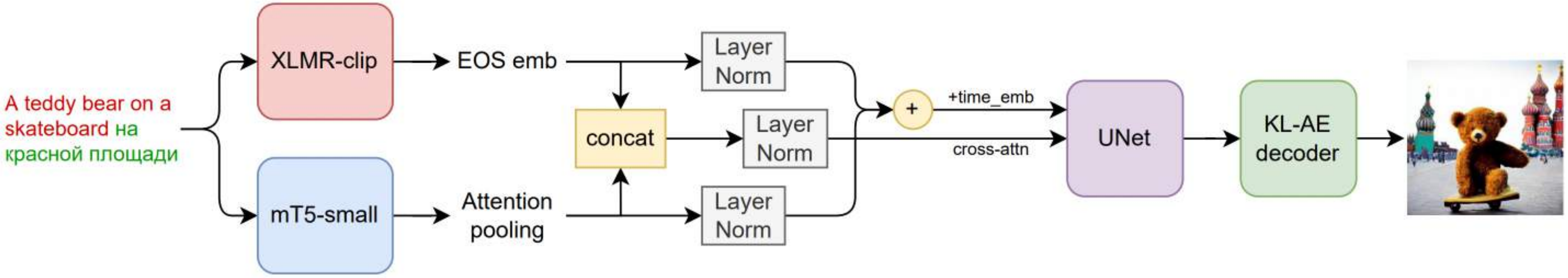
Енот в доспехах

Кресло в форме тыквы

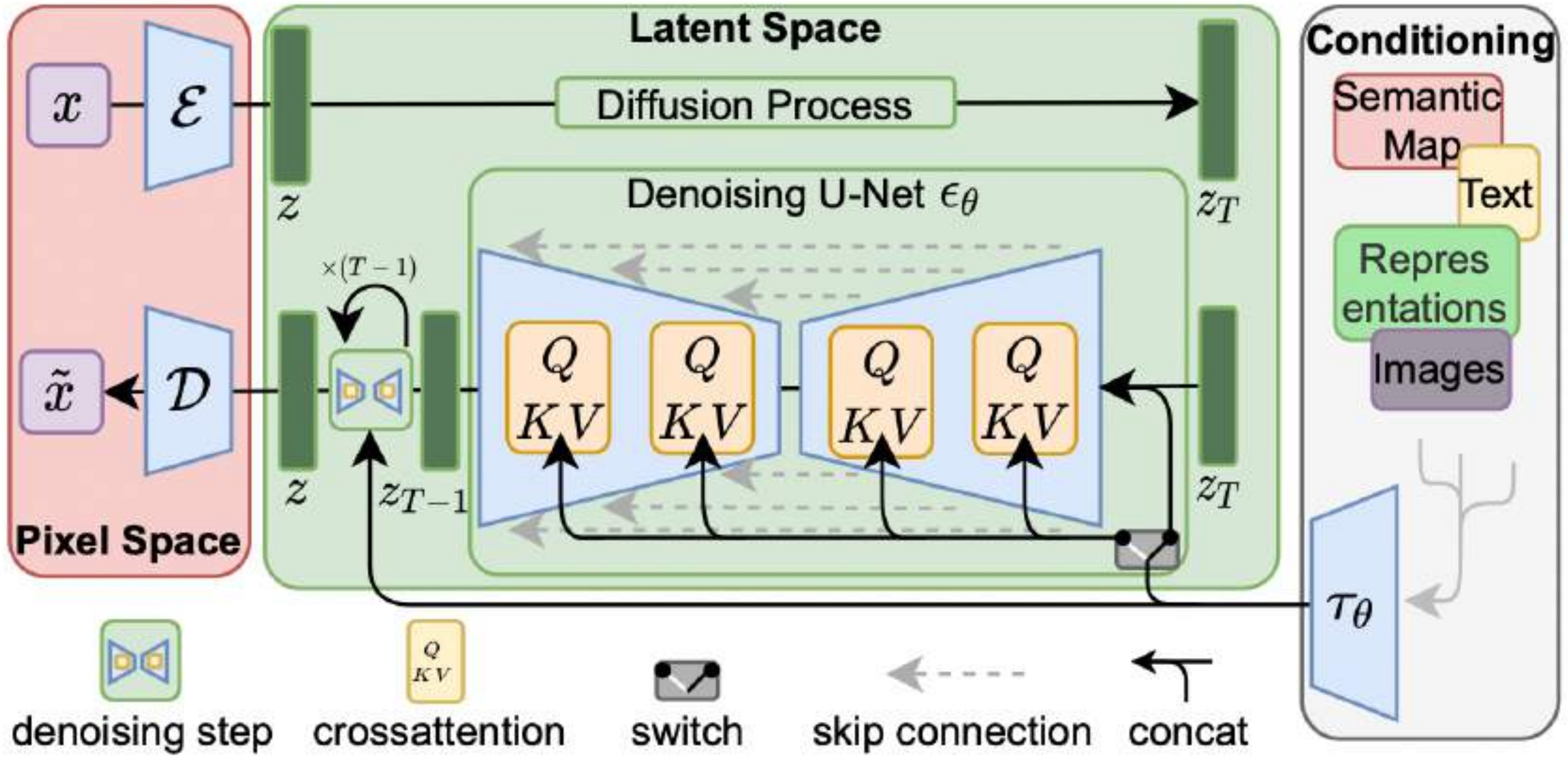
A portrait headshot of sci fi metallic human, bright eyes, complex geometric figure

github   хабр

Kandinsky 2.0 Architecture



- Two multilingual encoders: **XLMR-clip** and **mT5-small**
- 1.2B parameters in UNET
- Dynamic thresholding



Kandinsky 2.0 multilingual generation

Фото человека с
высшим образованием

Photo d'une personne
diplômée de l'enseignement supérieur

受过高等教育的人的照片
(китайский)

Kandinsky 2.0 multilingual generation

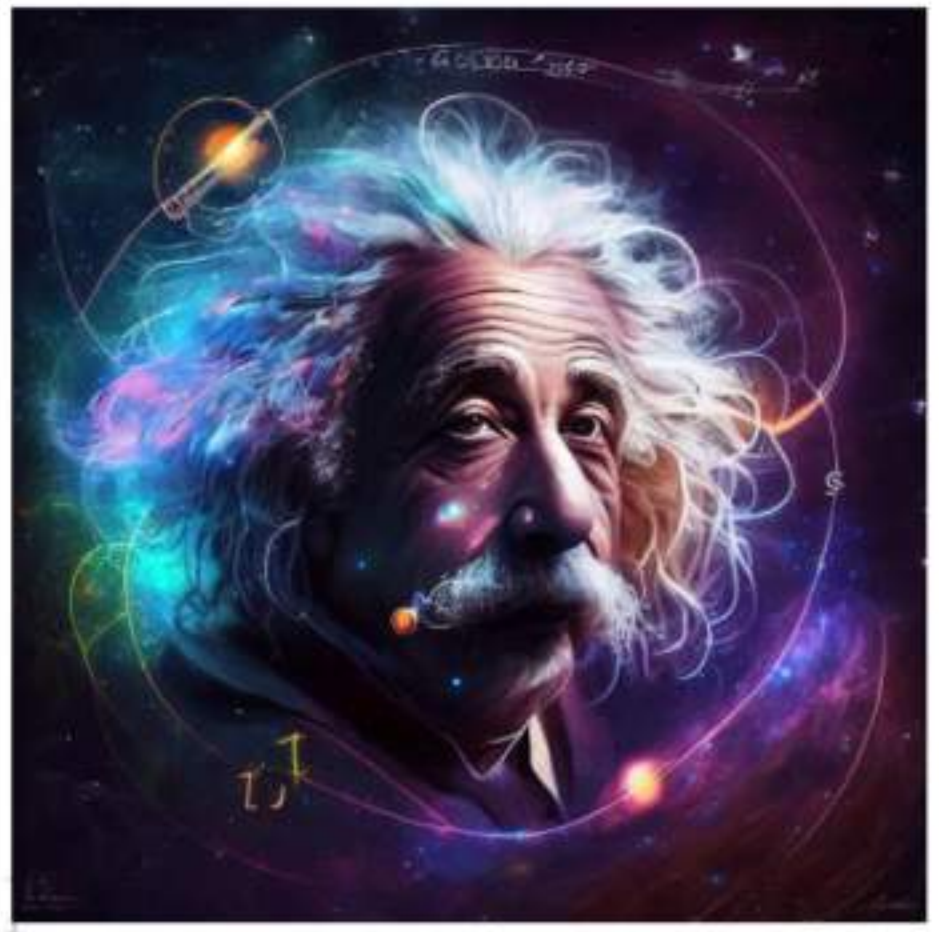
Фото грабителя

A photo of a burglar

एक चोर की तस्वीर (хинди)

Kandinsky 2.1

- Shares the same architecture as Kandinsky 2.0 + diffusion mapping of CLIP embeddings + new decoder (MoVQ)
- Developed by AIRI, SberAI, SberCloud
- Fully open-sourced



Einstein in space around the logarithm scheme

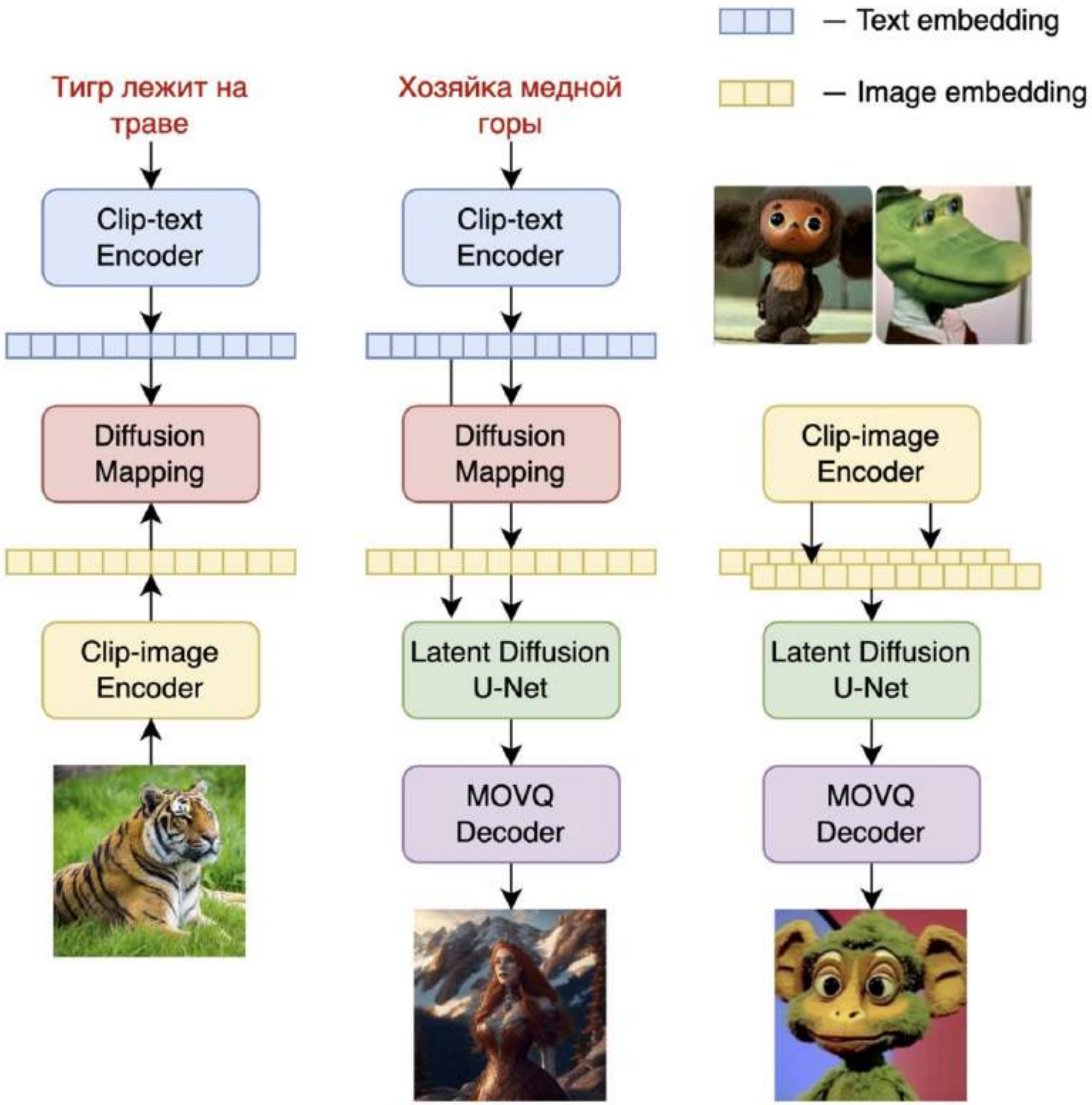
sad clown face 4k

mutant cat in the style of puppet animation in the style of horror film 4k

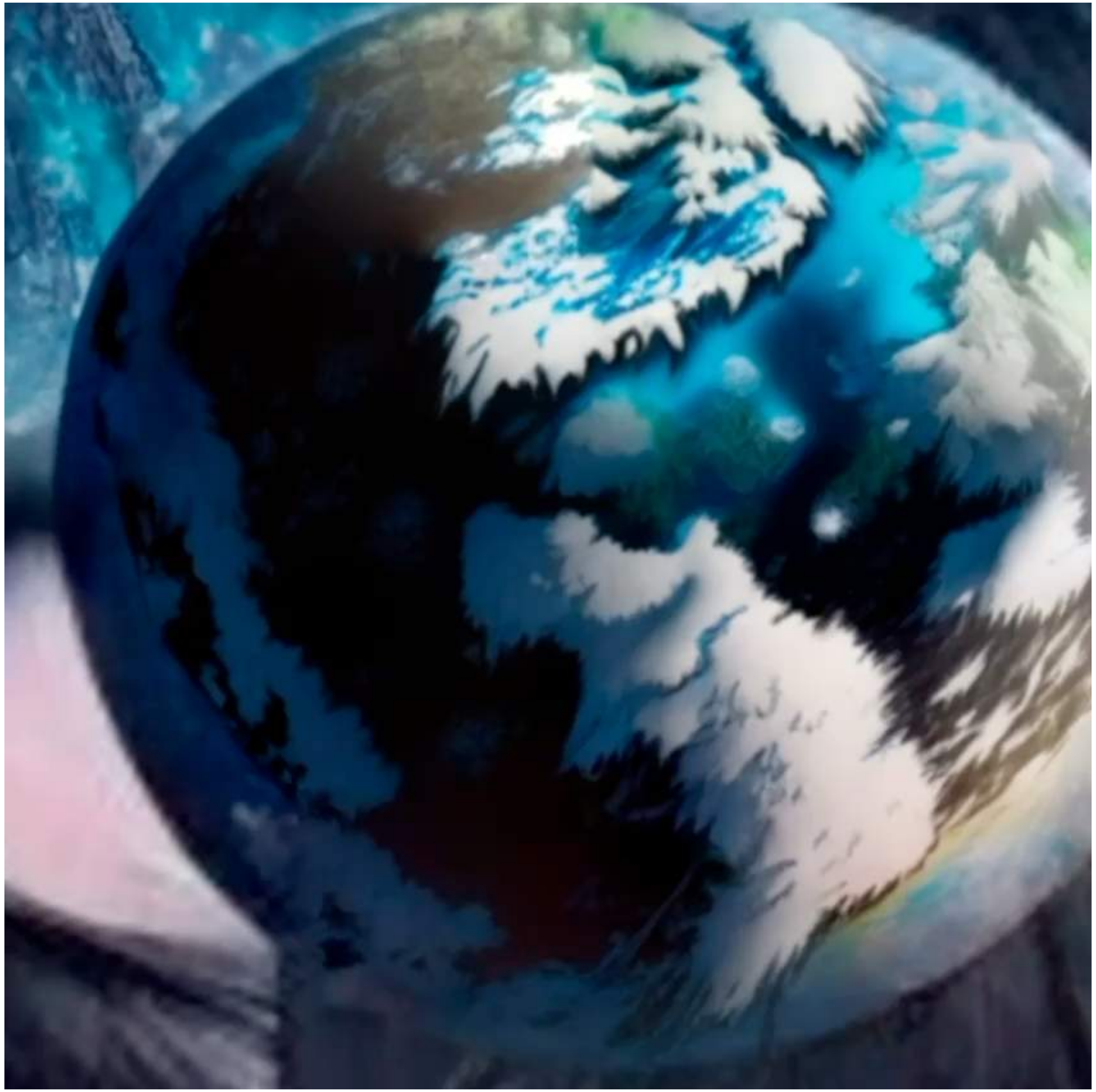
girl in the garden film grain, Kodak portra 800, f1.8, golden hour

github хабр

Kandinsky 2.1

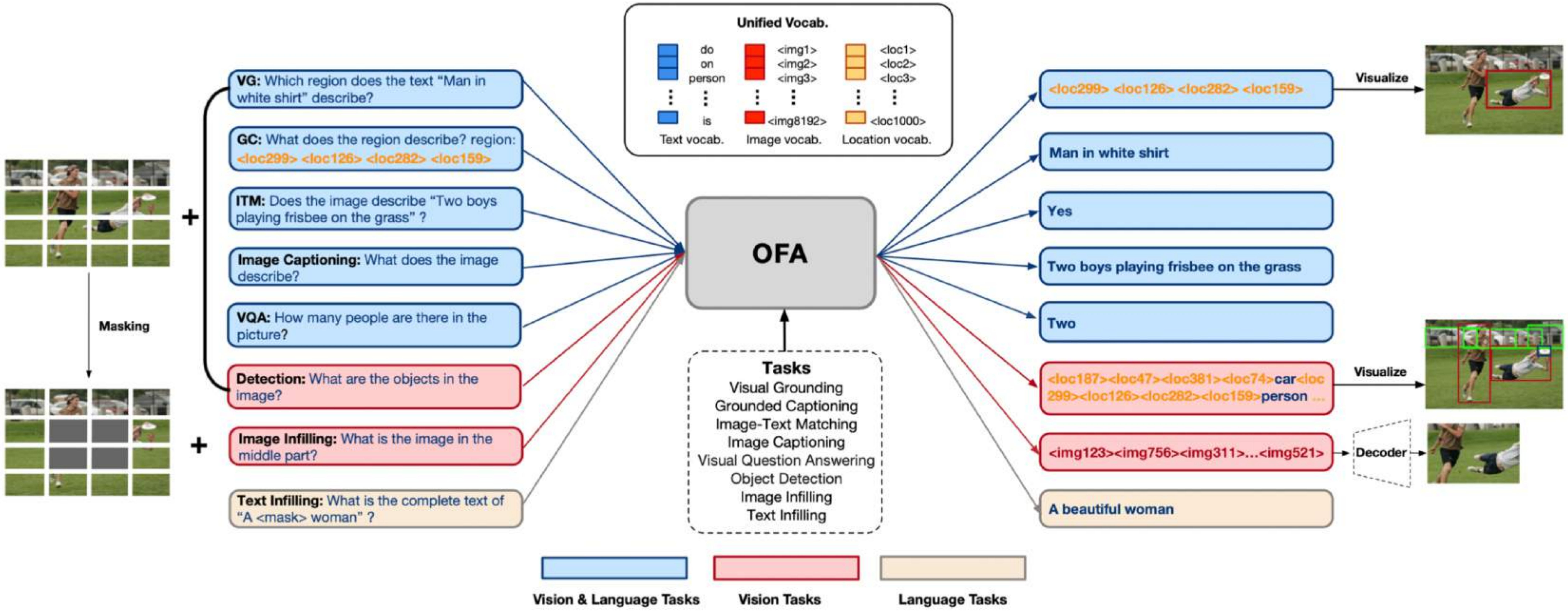


	FID-30K
eDiff-I (2022)	6,95
Imagen (2022)	7,27
Kandinsky 2.1 (2023)	💥**8,21**
Stable Diffusion 2.1 (2022)	8,59
GigaGAN, 512x512 (2023)	9,09
DALL-E 2 (2022)	10,39
GLIDE (2022)	12,24
Kandinsky 1.0 (2022)	15,40
DALL-E (2021)	17,89
Kandinsky 2.0 (2022)	20,00
GLIGEN (2022)	21,04



https://t.me/abstractDL/207

OFA



OFA: UNIFYING ARCHITECTURES, TASKS, AND MODALITIES THROUGH A SIMPLE SEQUENCE-TO-SEQUENCE LEARNING FRAMEWORK

OFA

Q: what color is the car in the region? region: <loc512> <loc483> <loc675> <loc576>
A: gray

One For All — a multimodal network from Alibaba which can solve almost every possible task:

- text2image generating
- image captioning
- image inpainting
- VQA
- object detection
- NLU

The text prompt is used to switch between tasks.
So you should just "ask" the model to do something.

Architecture — **encoder-decoder**, almost the same as BART.
For text tokens, visual tokens and spatial (location) tokens the same representation weights are shared (embeddings).

It is trained with a simple **cross-entropy loss** on multiple tasks.

Input: text (optionally), image (optionally), location (optionally)
Output: text or/and image or/and location

Interestingly, it can solve even tasks that it did not see during training!

Flamingo

Flamingo — is multimodal network. It is noticeable as authors did not train vision and language models from scratch, these models are **pretrained and frozen**.

Only **cross-attention and small adapters** are trained — a kind of connections between modalities.

Training set — **interleaved texts and images**. As it is in web pages.

Parameters: **60B**

Input: interleaved text and images
Output: text

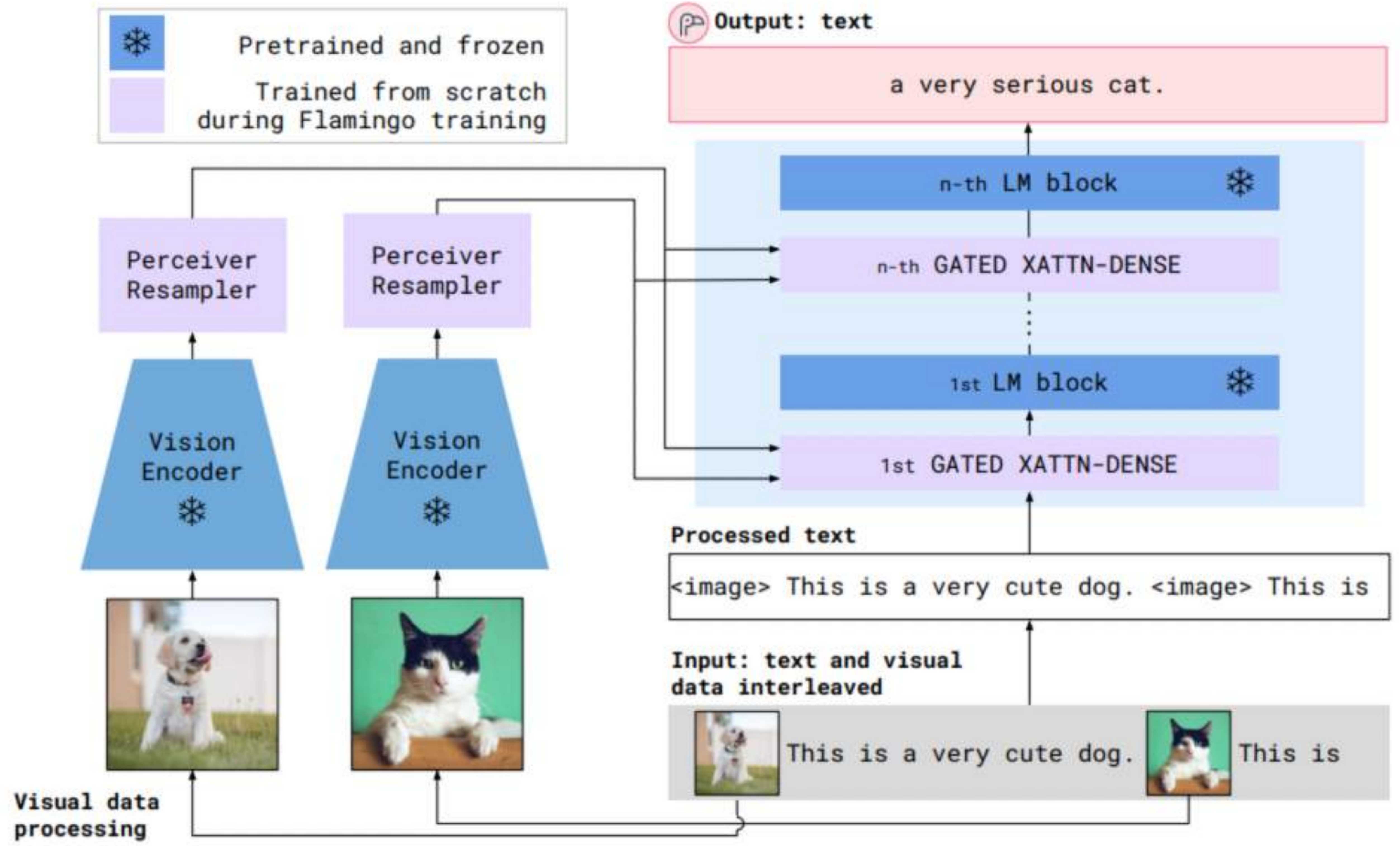


Figure 3 | **Overview of the Flamingo model.** The Flamingo models are a family of visual language model (VLM) that can take as input visual data interleaved with text and can produce free-form text as output. Key to its performance are novel architectural components and pretraining strategies described in Section 3.

🦩 Flamingo: a Visual Language Model for Few-Shot Learning

Flamingo

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

What are they doing?

They are having a conversation.

What object are they using?

It looks like a computer.

Is this surprising?

Yes, it is surprising.

Why is this picture surprising to you?

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

What is the common thing about these three images?

They are all flamingos.

What is the difference between these three images?

The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.

This is an apple with a sticker on it.

What does the sticker say?

The sticker says "iPod".

Where is the photo taken?

It looks like it's taken in a backyard.

Do you think it is printed or handwritten?

It looks like it's handwritten.

What color is the sticker?

It's white.

FROMAGe

FROMAGe — the easiest approach to train a multimodal dialog model. Also it is capable of retrieving images from a given set.

Only **small adapter** is trained — a mapping of visual embeddings to text embeddings and inserted into GPT.

Training set — **image with captions (CC4M)**

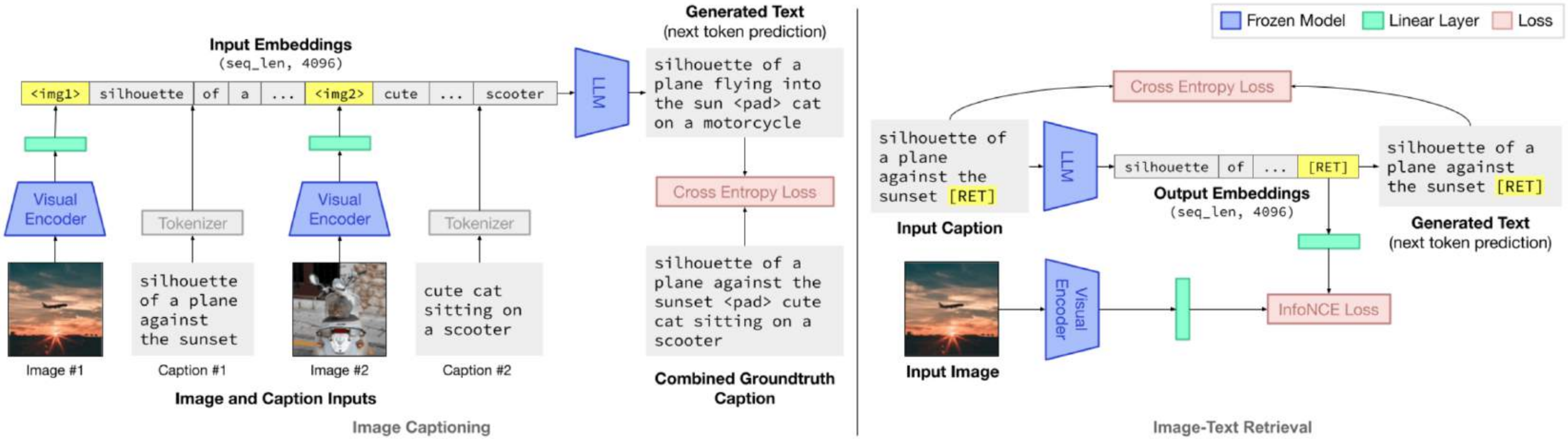
Parameters: **30B** (but only 5M are trainable)

Input: interleaved text and images
Output: text, retrieved images

Grounding Language Models to Images for Multimodal Generation (jykoh.com/fromage)

Grounding Language Models to Images for Multimodal Generation

FROMAGe



Input Embeddings
(seq_len, 4096)
<img1> silhouette of a ... <img2> cute ... scooter
LLM
Generated Text
(next token prediction)
silhouette of a plane flying into the sun <pad> cat on a motorcycle
Cross Entropy Loss
silhouette of a plane against the sunset <pad> cute cat sitting on a scooter
Combined Groundtruth Caption
Visual Encoder
Tokenizer
silhouette of a plane against the sunset
cute cat sitting on a scooter
Image #1
Caption #1
Image #2
Caption #2
Image and Caption Inputs
Image Captioning
Frozen Model
Linear Layer
Loss
Cross Entropy Loss
silhouette of a plane against the sunset [RET]
Input Caption
silhouette of ... [RET]
Output Embeddings
(seq_len, 4096)
silhouette of a plane against the sunset [RET]
Generated Text
(next token prediction)
Input Image
InfoNCE Loss
Image-Text Retrieval