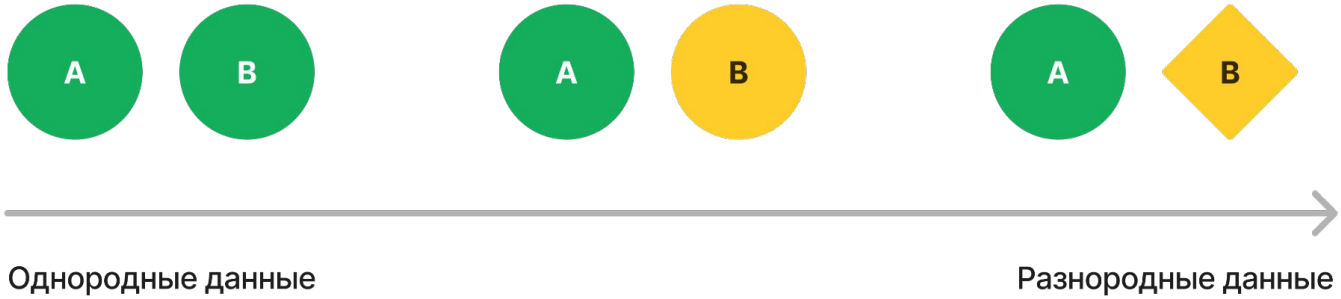
ИТМО

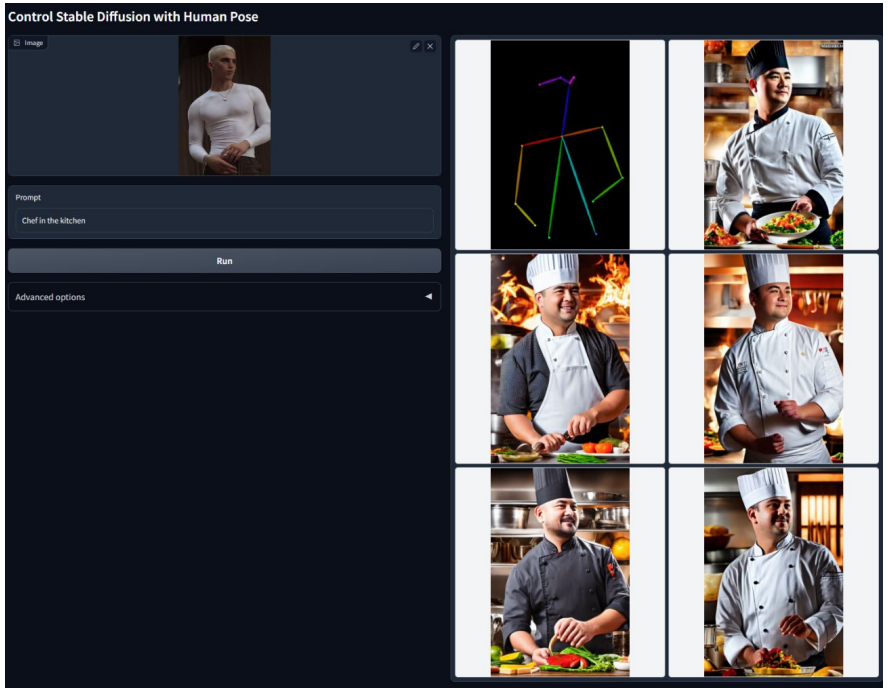
Современные архитектуры нейронных сетей

Мультимодальность

Что такое модальность

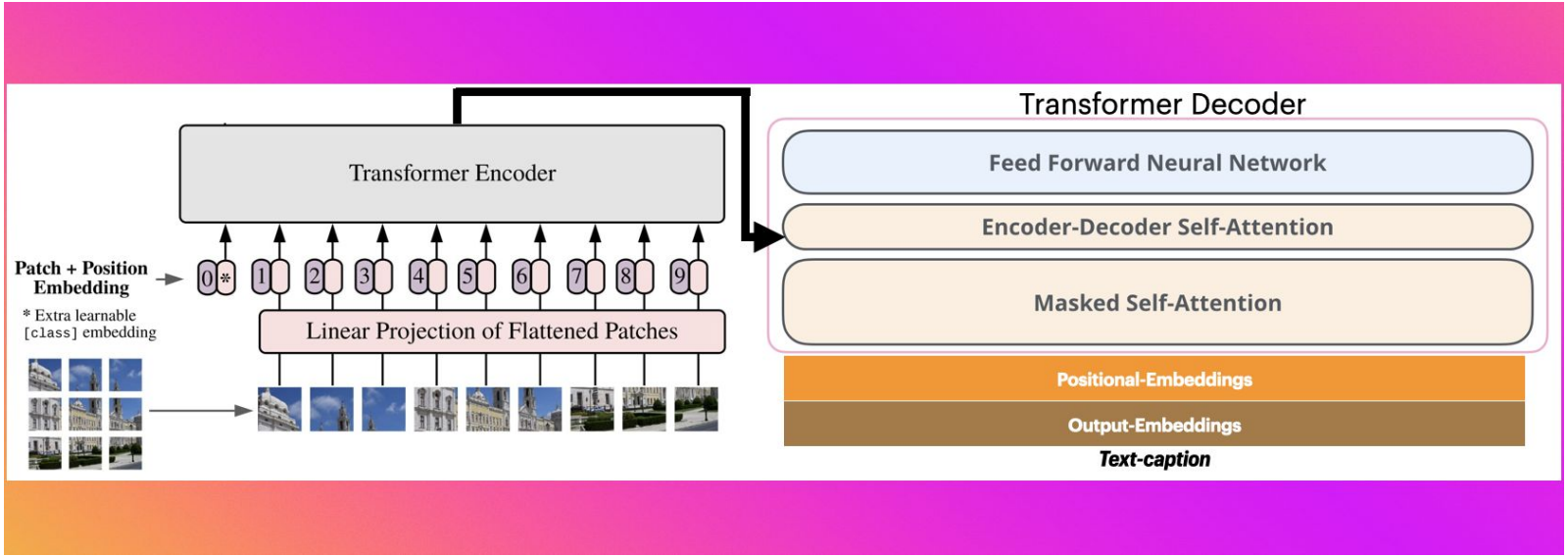


Что такое модальность



Чем больше модальностей - тем лучше

Image Captioning



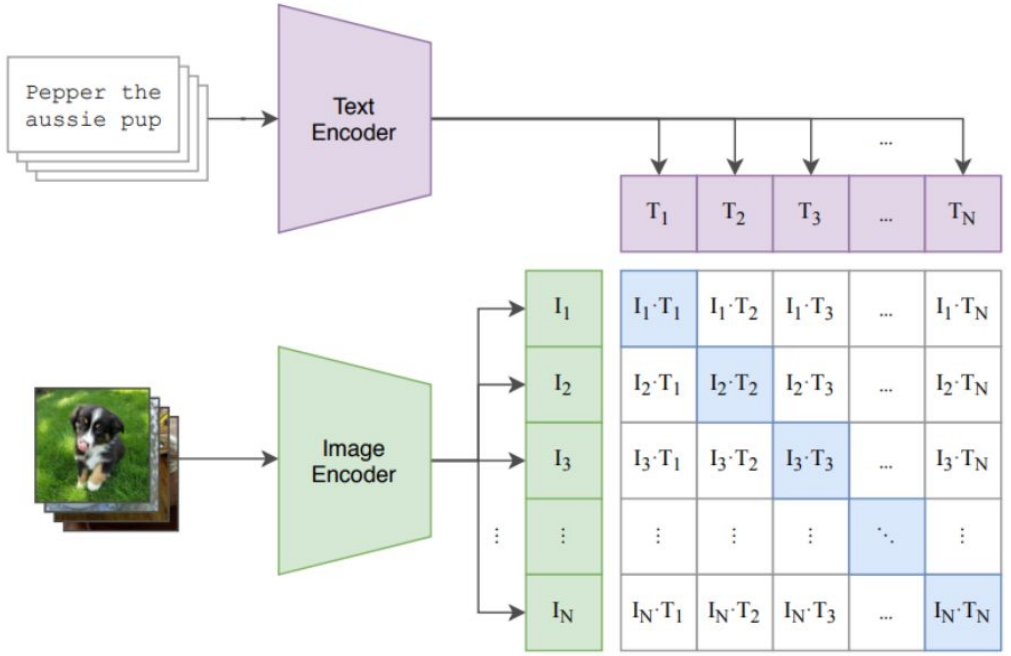
CLIP

ИТМО

$\{I_k\}_{k=1}^{batch\_size}$ - векторы изображений

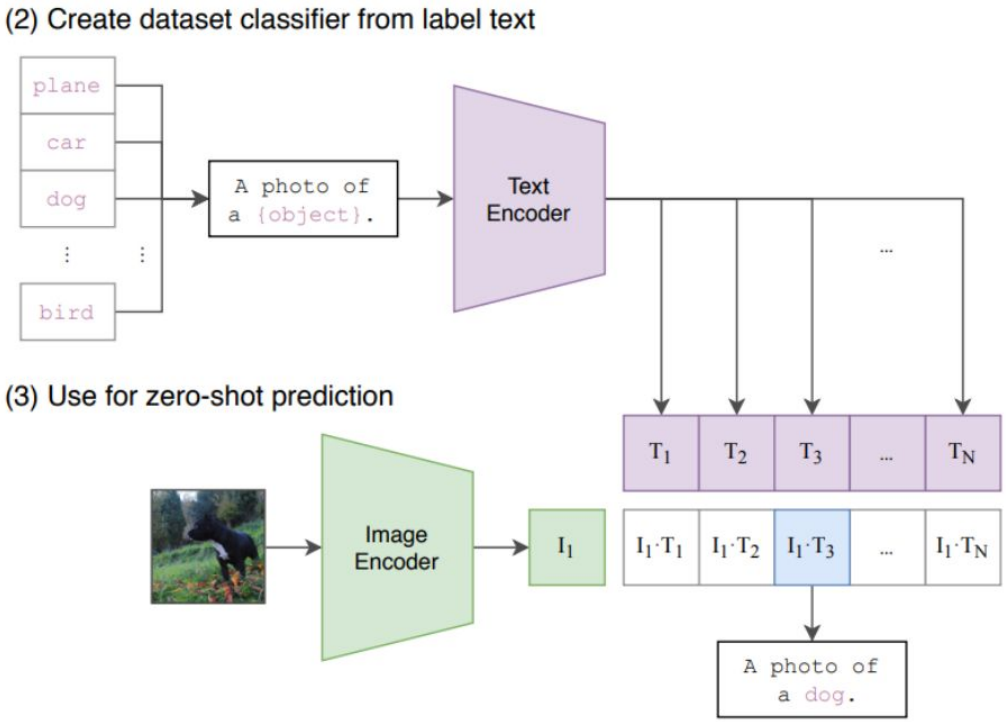
$\{\tau_k\}_{k=1}^{batch\_size}$ - векторы текста

$$cosine\_similarity(I_k, \tau_j) = \frac{(I_k, \tau_j)}{||I_k|| \cdot ||\tau_j||}$$

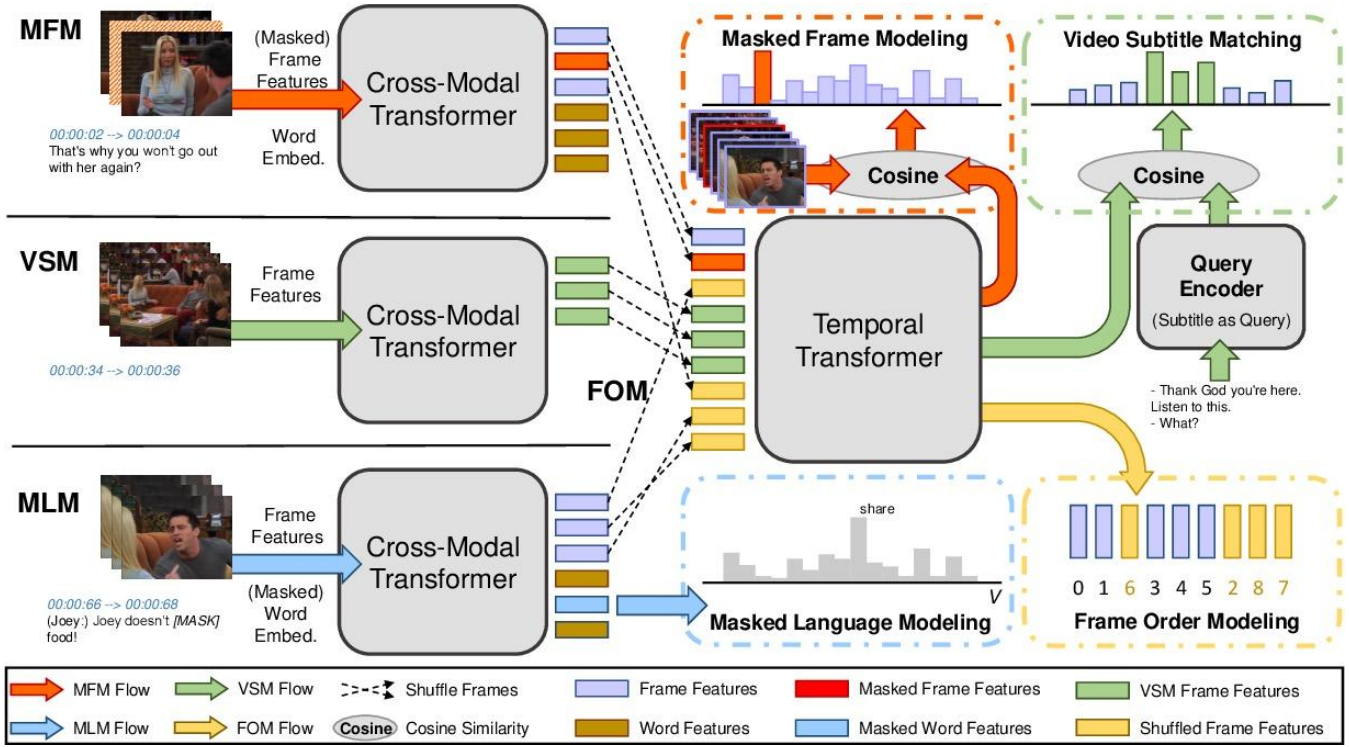


CLIP

ИТМО

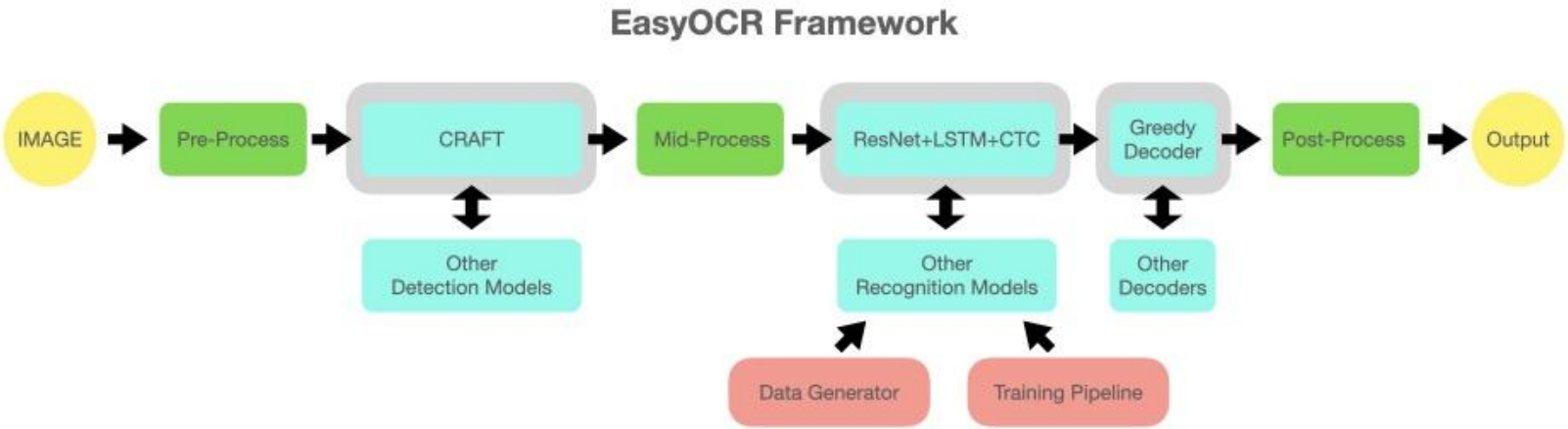

(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Кроссмодальность трансформеров

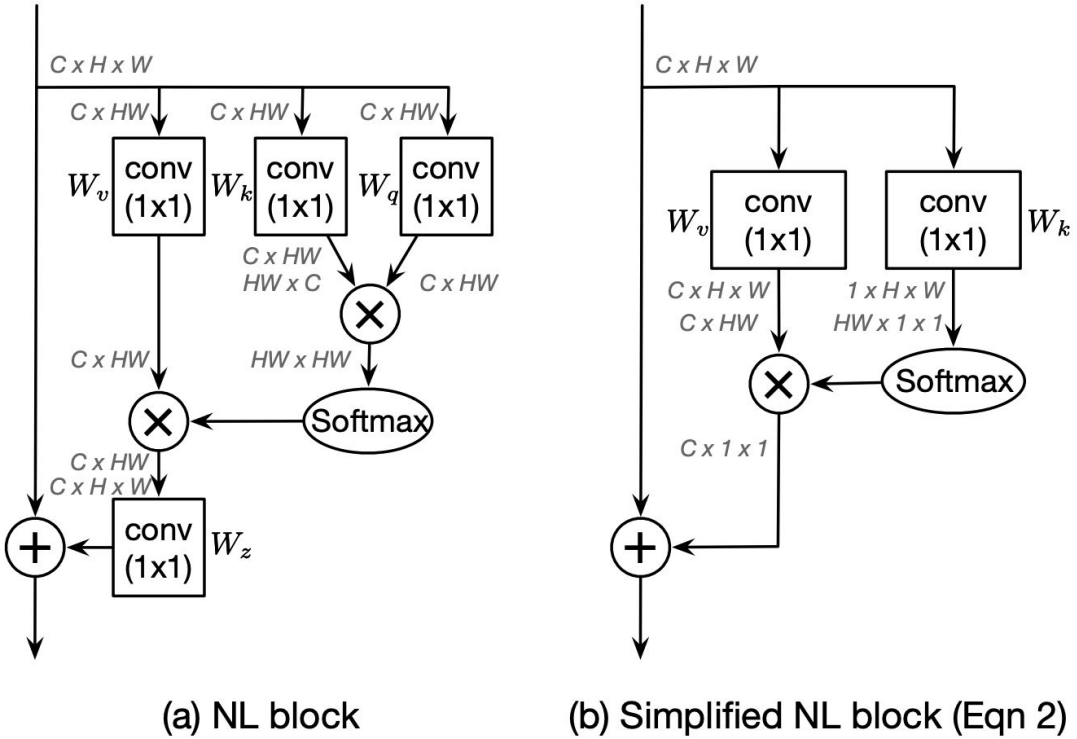


Картинка для устрашения

Свертки vs трансформеры



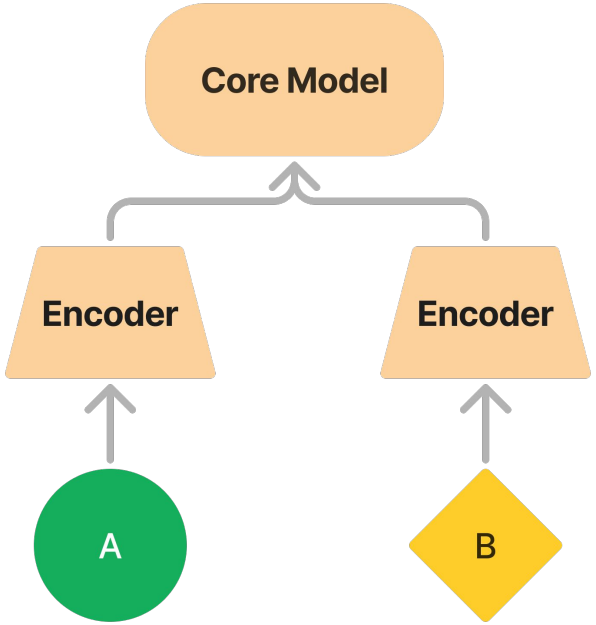
Свертки vs трансформеры



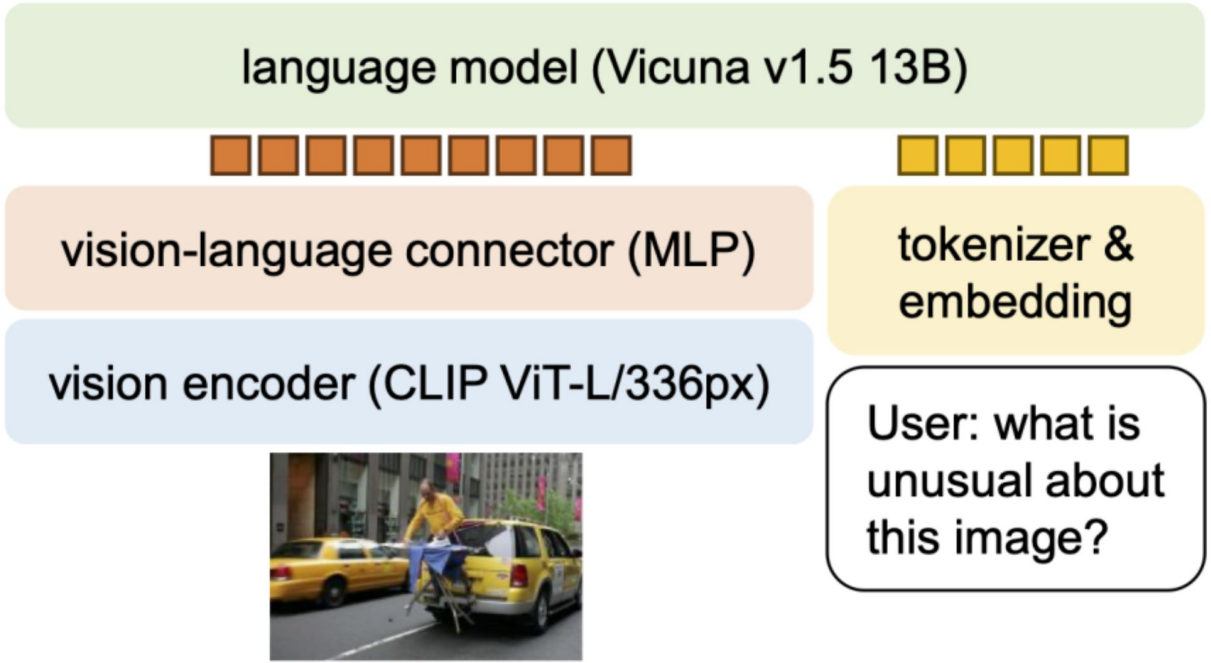
(a) NL block

(b) Simplified NL block (Eqn 2)

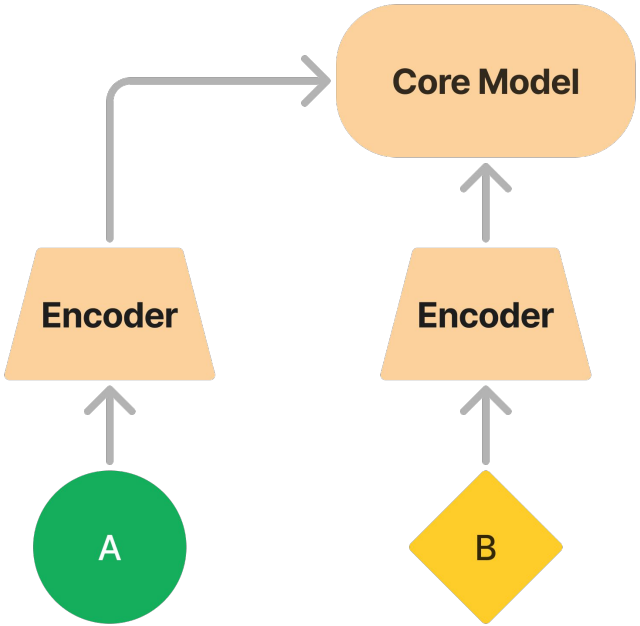
Смешивание модальностей. Early Fusion



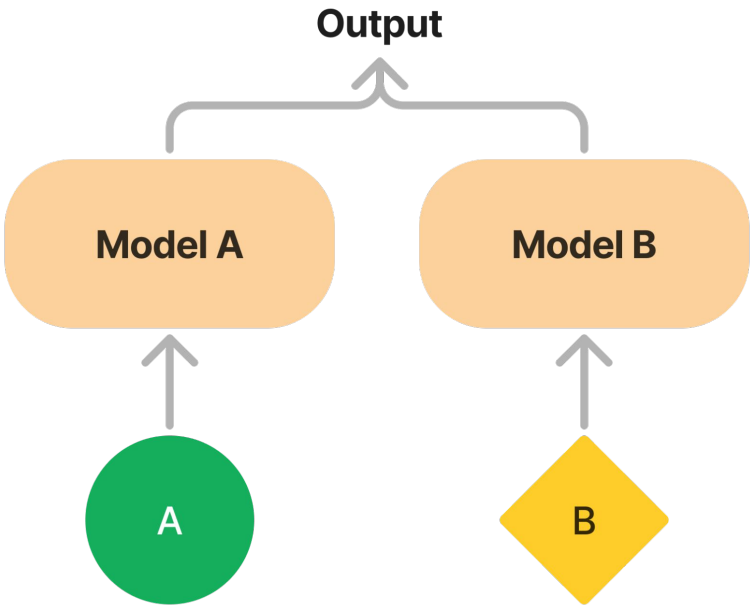
Смешивание модальностей. Early Fusion



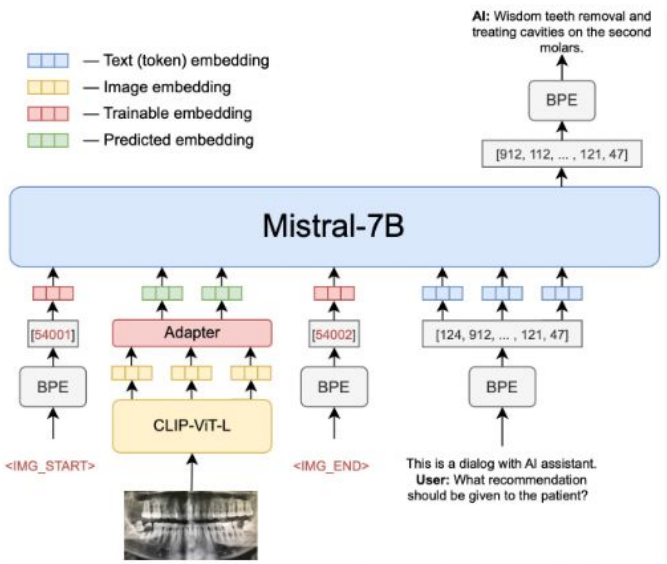
Смешивание модальностей. Deep Fusion

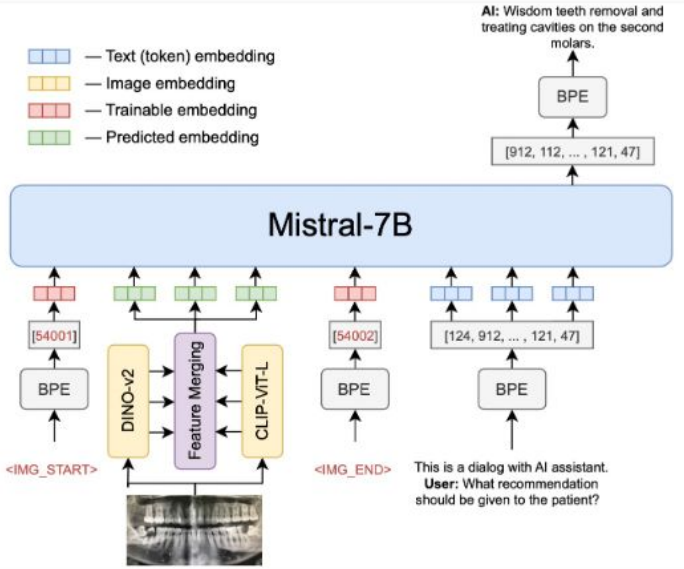


Смешивание модальностей. Late Fusion



Смешивание модальностей. Merge of Encoders





Уменьшение числа токенов

ИТМО

- **Pixel-shuffle**
- **Аддитивное смешивание**

$$z = \sum_{i=1}^{k} w_i x_i,$$

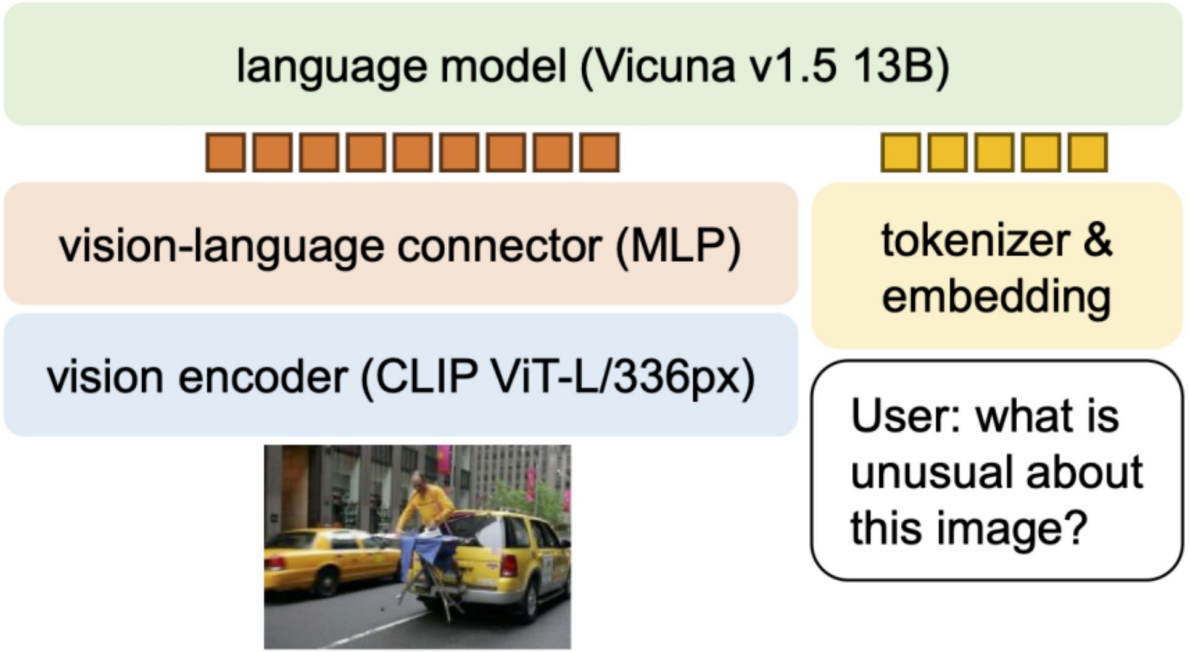
- **Мультипликативное смешивание**

$$z = w(x_1 \times x_2)$$

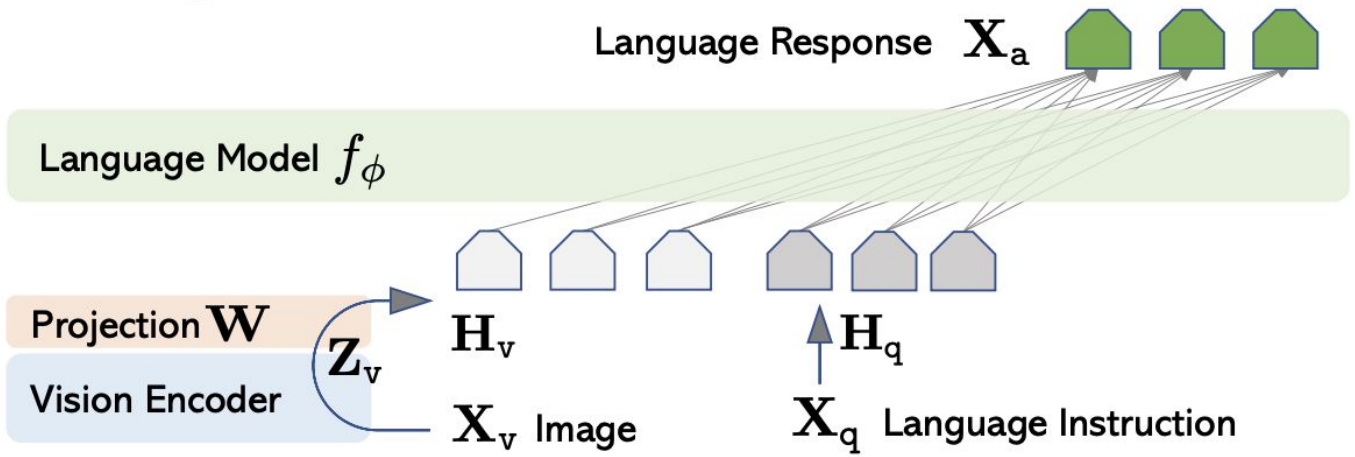
- **Gated Fusion**

$$z = \sum_{i=1}^{k} g_i(x_1, x_2, ..., x_k) x_i$$

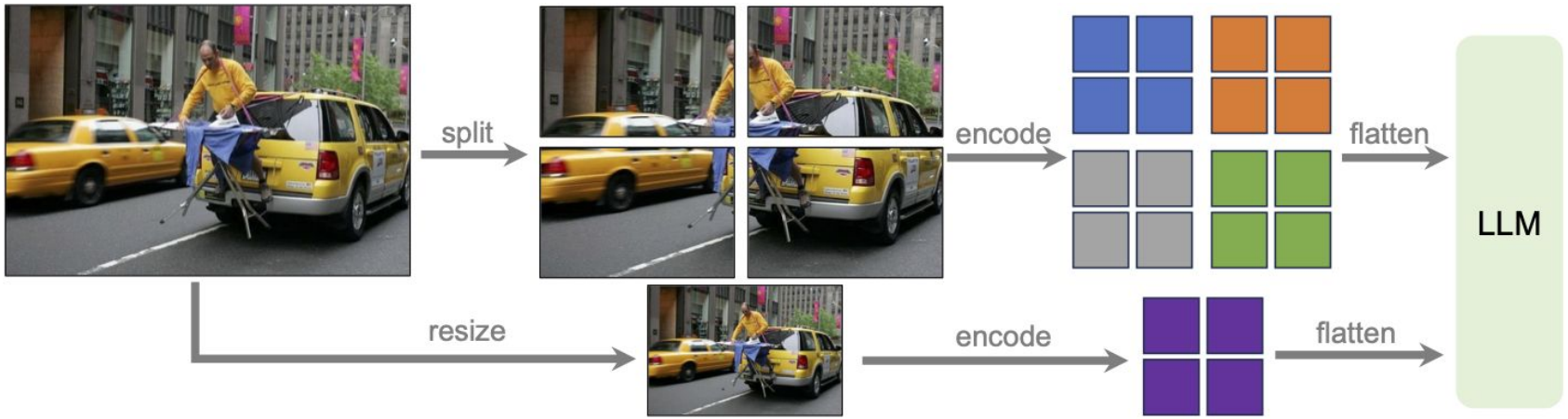
Captioning


language model (Vicuna v1.5 13B)
vision-language connector (MLP)
tokenizer & embedding
vision encoder (CLIP ViT-L/336px)
User: what is unusual about this image?

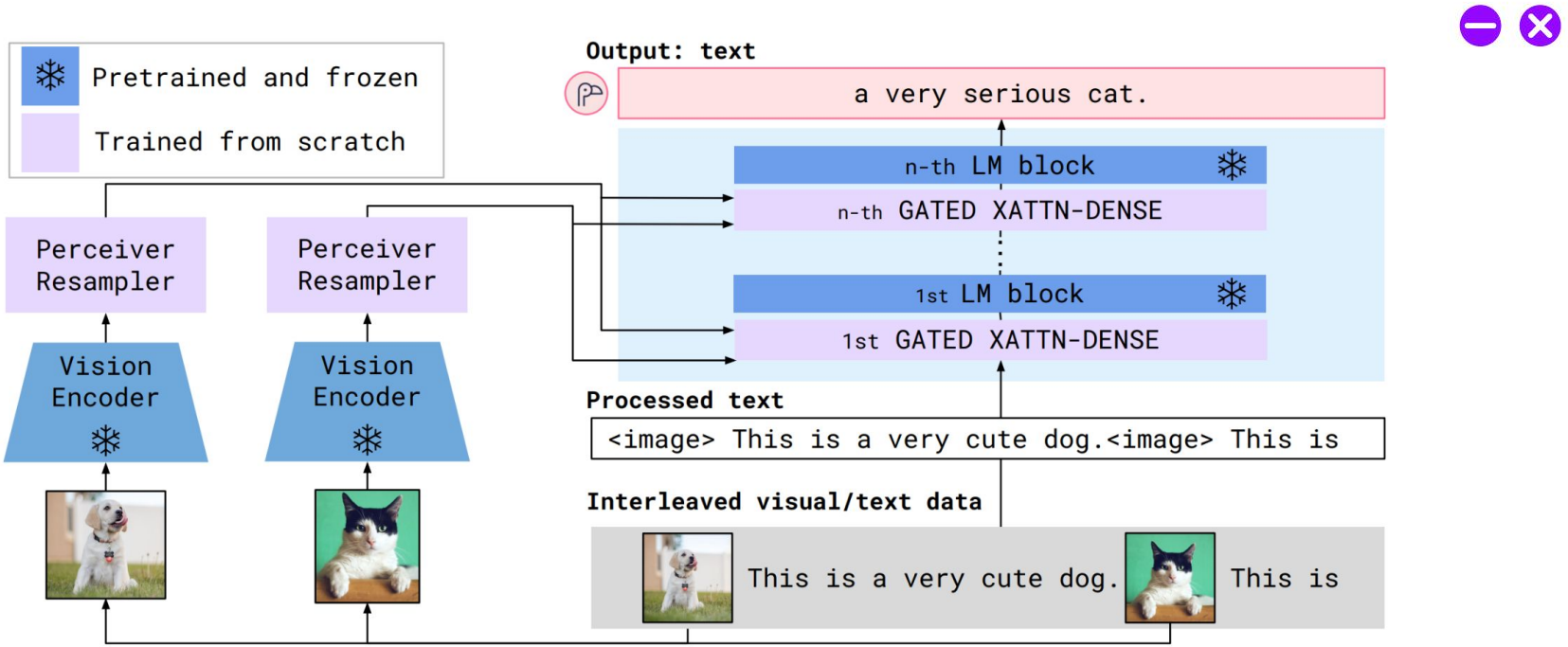
Early Fusion VLM (LLaVA)



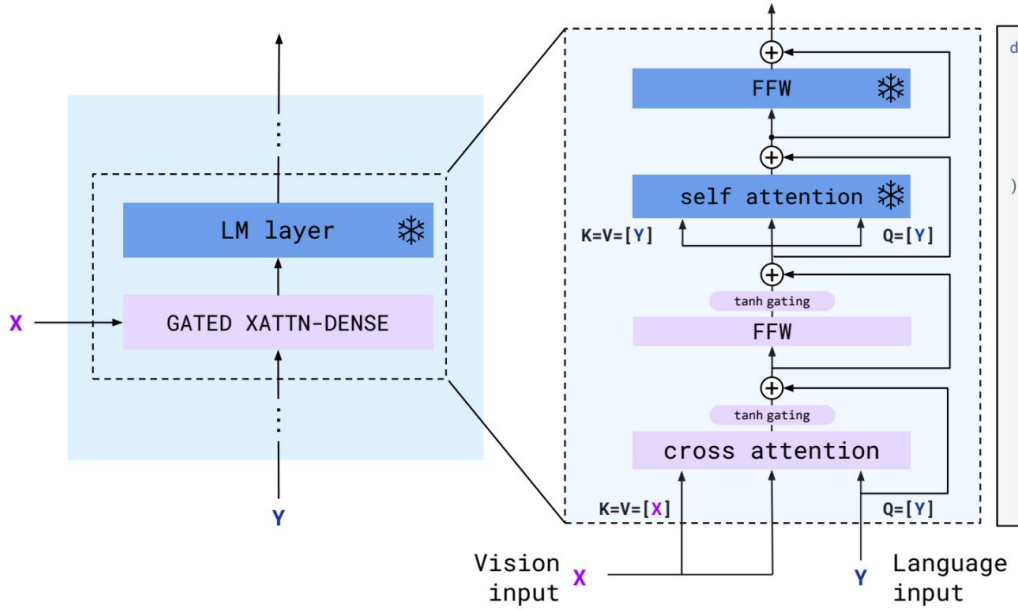
LLaVA Next



Deep Fusion VLM (Flamingo)

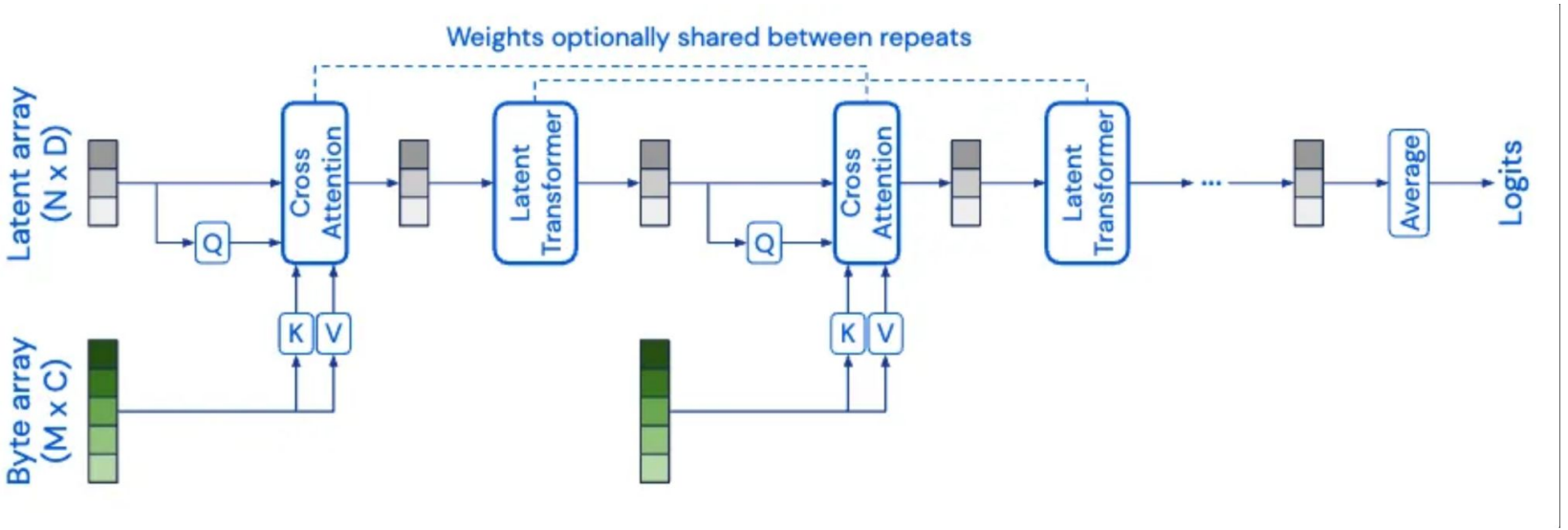


Deep Fusion VLM (Flamingo XAttn)

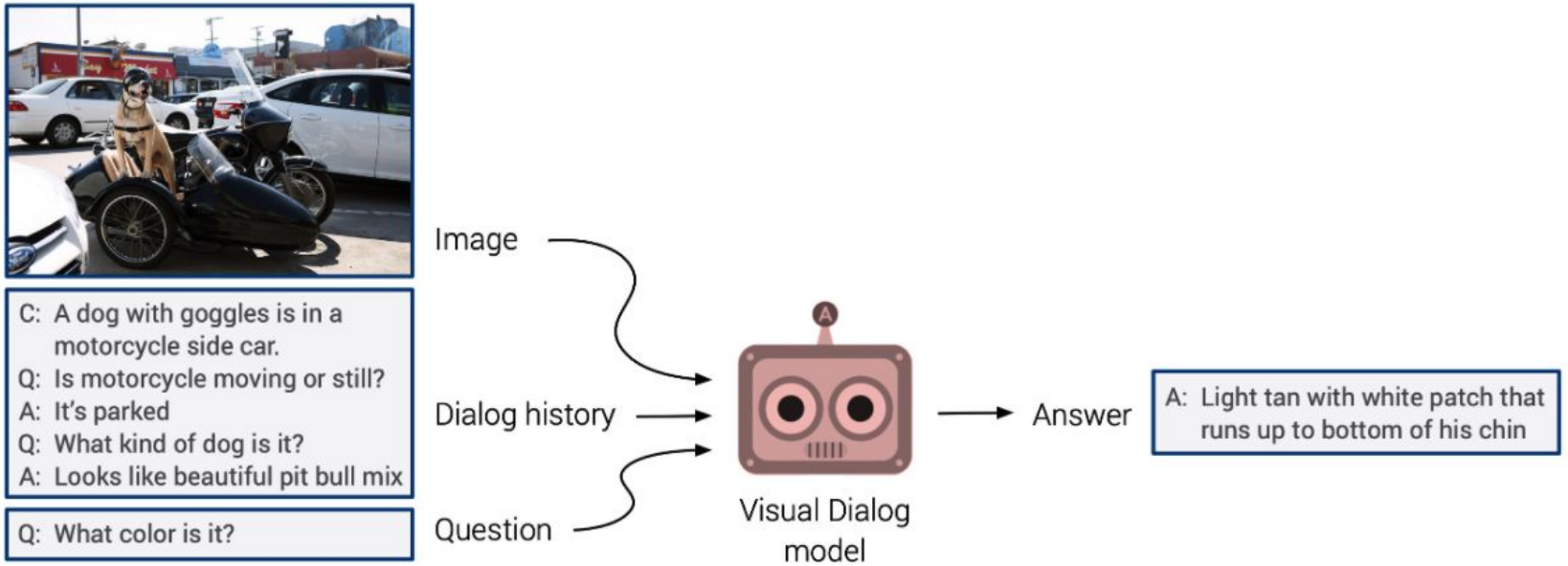



```
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 1. Gated Cross Attention
  y = y + tanh(alpha_xattn) * attention(q=y, kv=x)
  # 2. Gated Feed Forward (dense) Layer
  y = y + tanh(alpha_dense) * ffw(y)

  # Regular self-attention + FFW on language
  y = y + frozen_attention(q=y, kv=y)
  y = y + frozen_ffw(y)
  return y  # output visually informed language features
```

Perceiver Resampler



Visual Question Answering



DALL-E

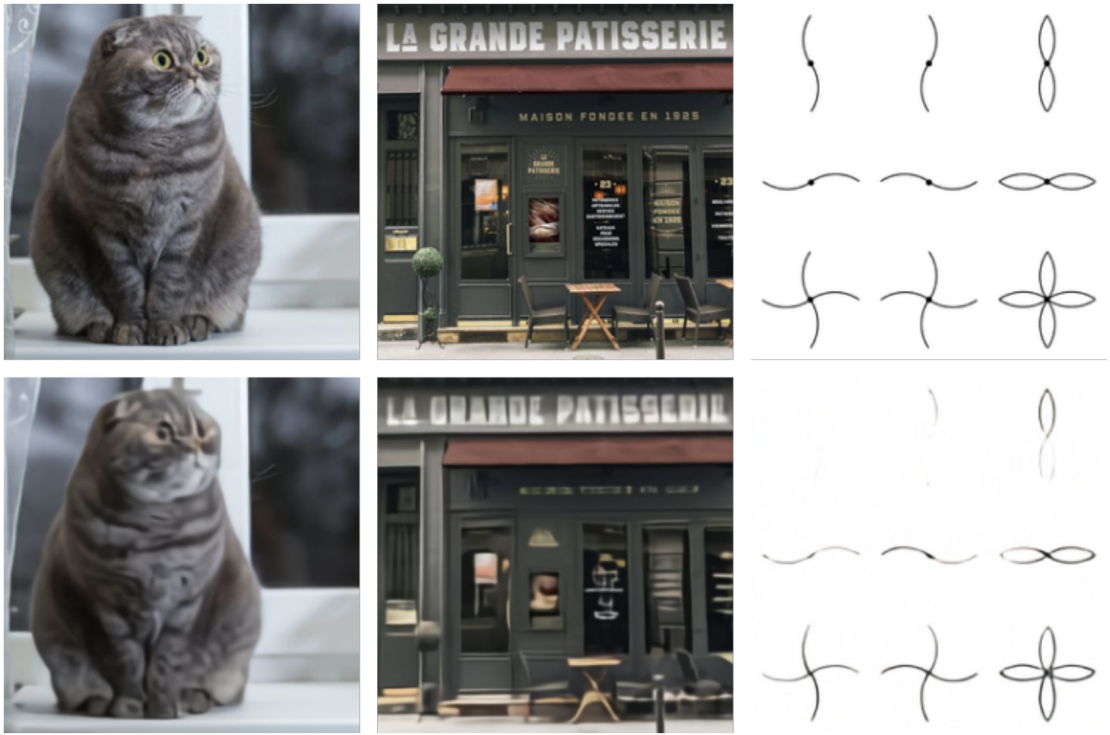
ИТМО

DALL·E 2 can take an image and create different variations of it inspired by the original.

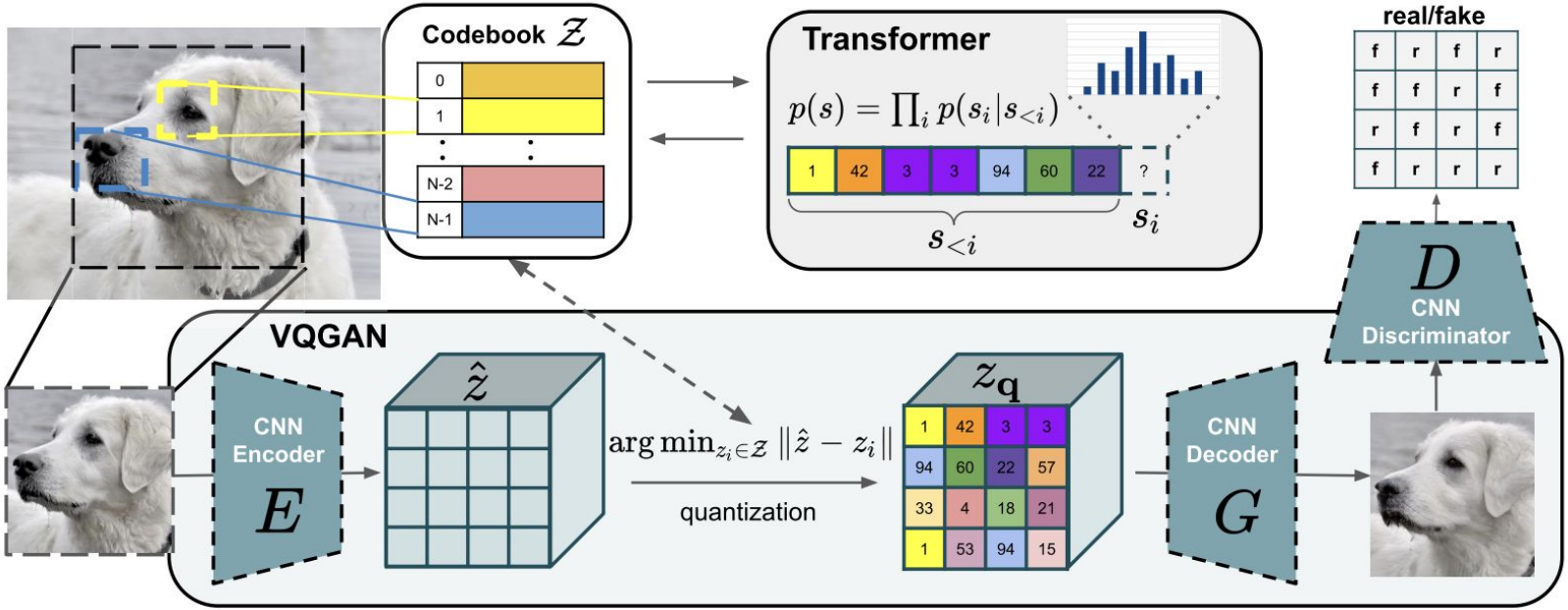
ORIGINAL IMAGE

DALL·E 2 VARIATIONS

Кодирование VAE



VQ-GAN


Codebook $\mathcal{Z}$
0
1
N-2
N-1
Transformer
$p(s) = \prod_i p(s_i|s_{<i})$
1 42 3 3 94 60 22 ?
$s_{<i}$
s_i
real/fake
f r f r
f f r f
r f r f
f r r r
D
CNN Discriminator
VQGAN
CNN Encoder
E
$\hat{z}$
$\arg\min_{z_i \in \mathcal{Z}} \|\hat{z} - z_i\|$
quantization
$z_{\mathbf{q}}$
1 42 3 3
94 60 22 57
33 4 18 21
1 53 94 15
CNN Decoder
G