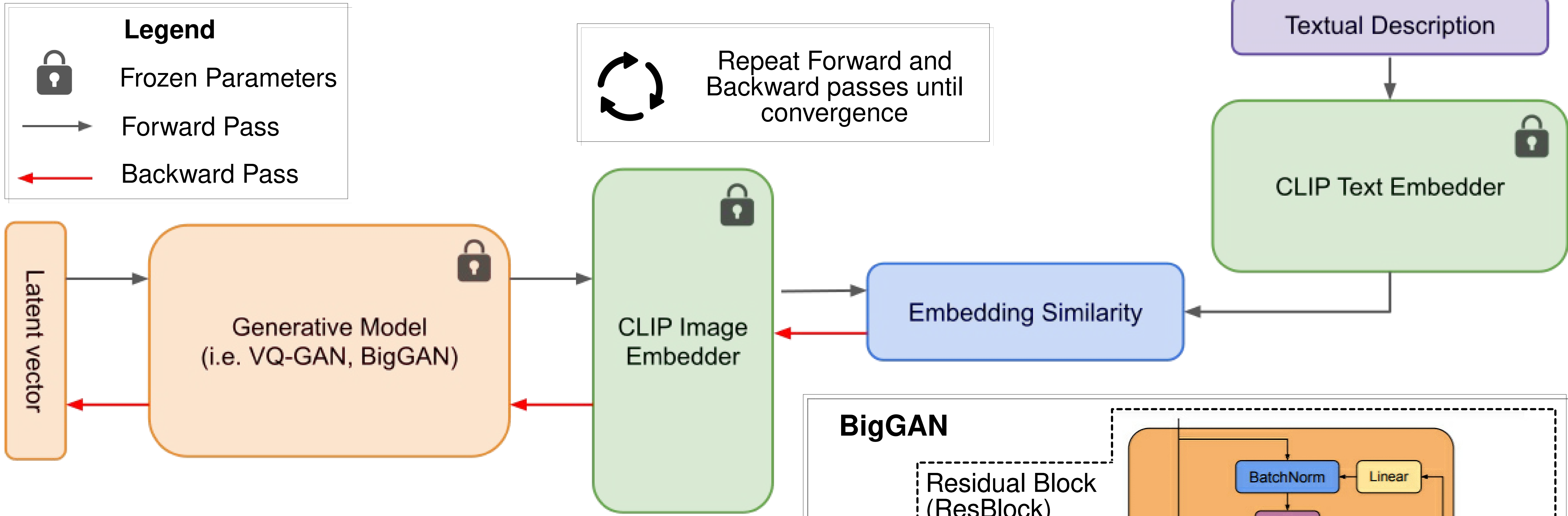# Text-to-Image Generation
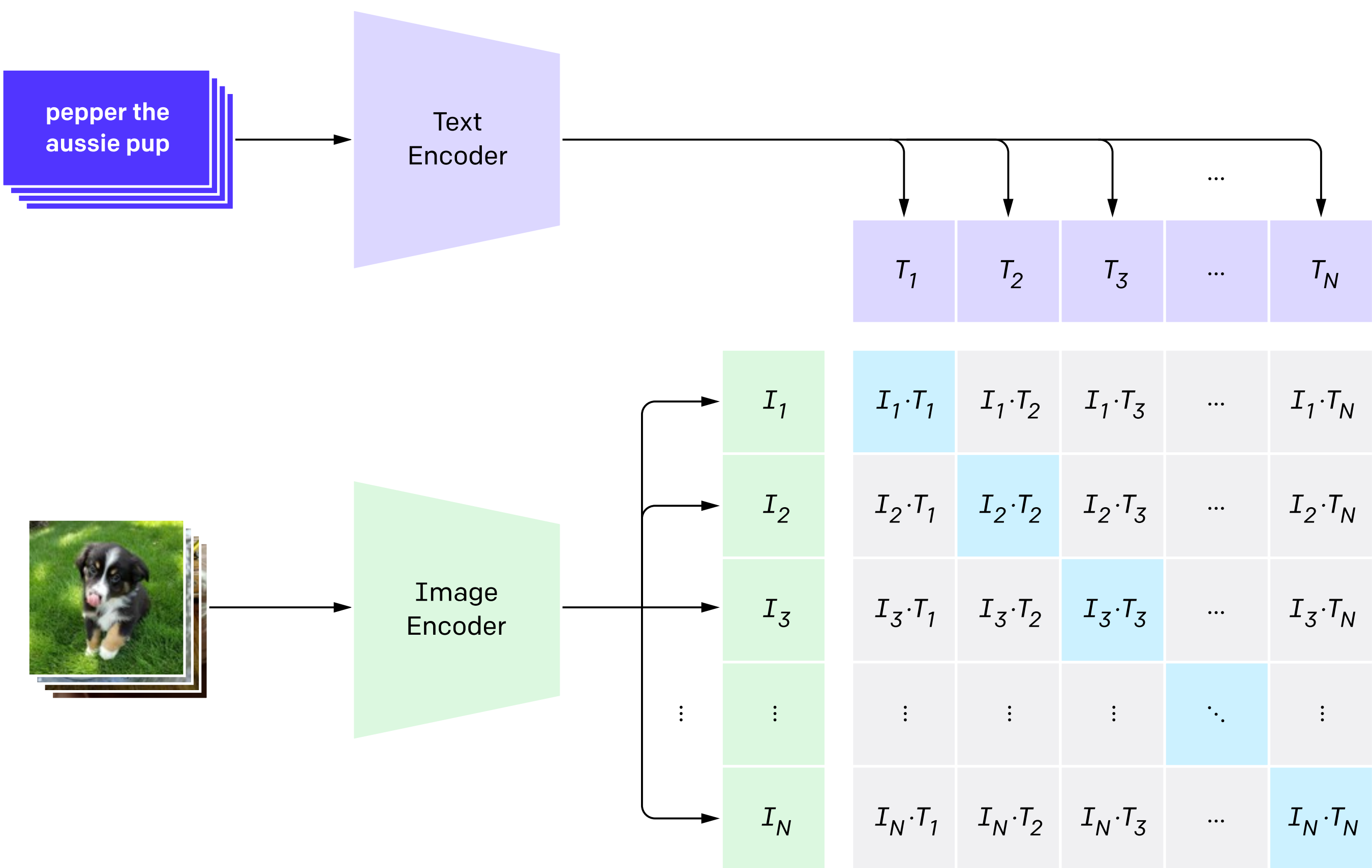# Synthesizing Images from Textual Descriptions

Xabier Lahuerta Vázquez
HAP/LAP - University of the Basque Country, Donostia, Basque Country
xlahuerta001@ikasle.ehu.eus

Legend

🔒 Frozen Parameters
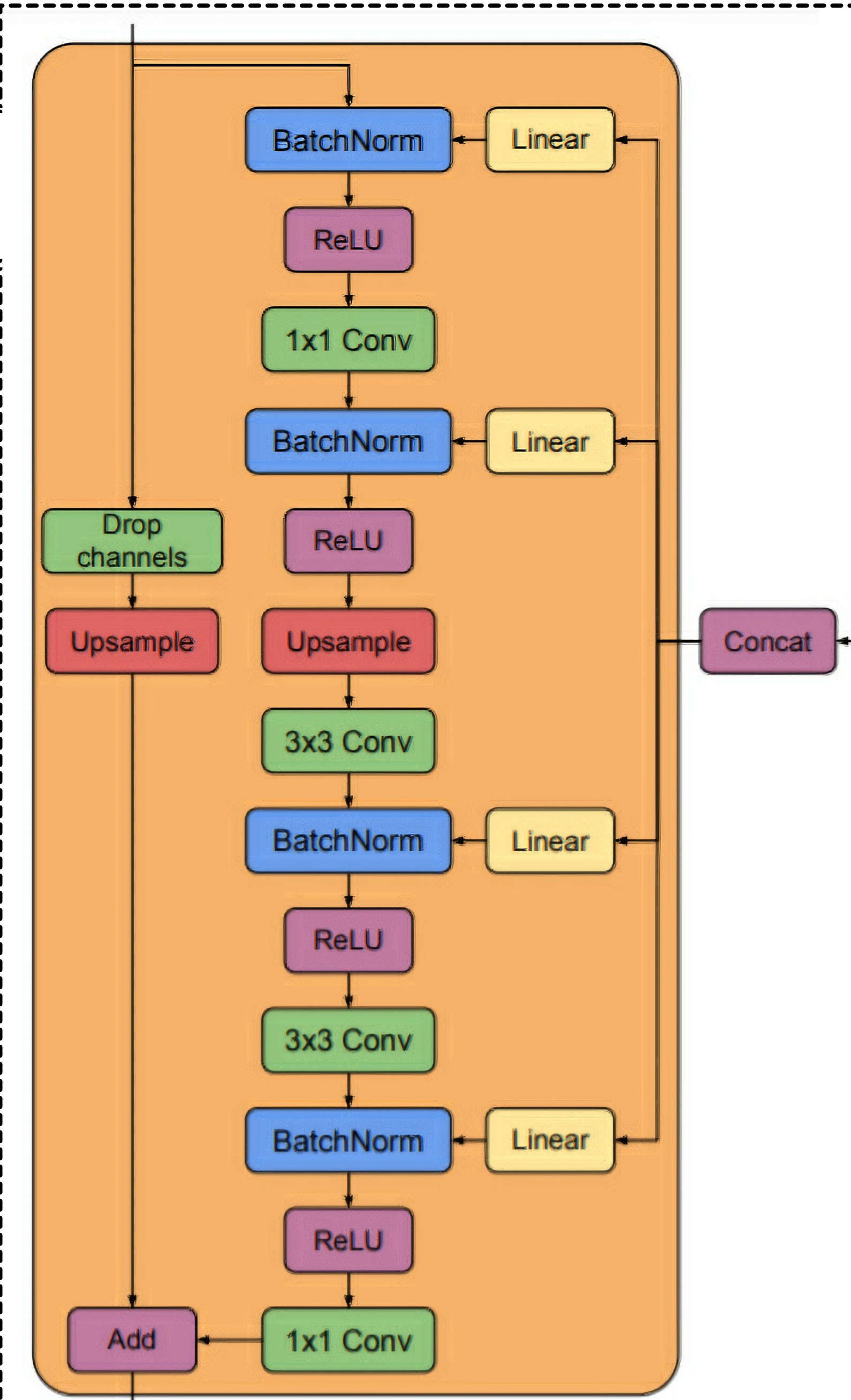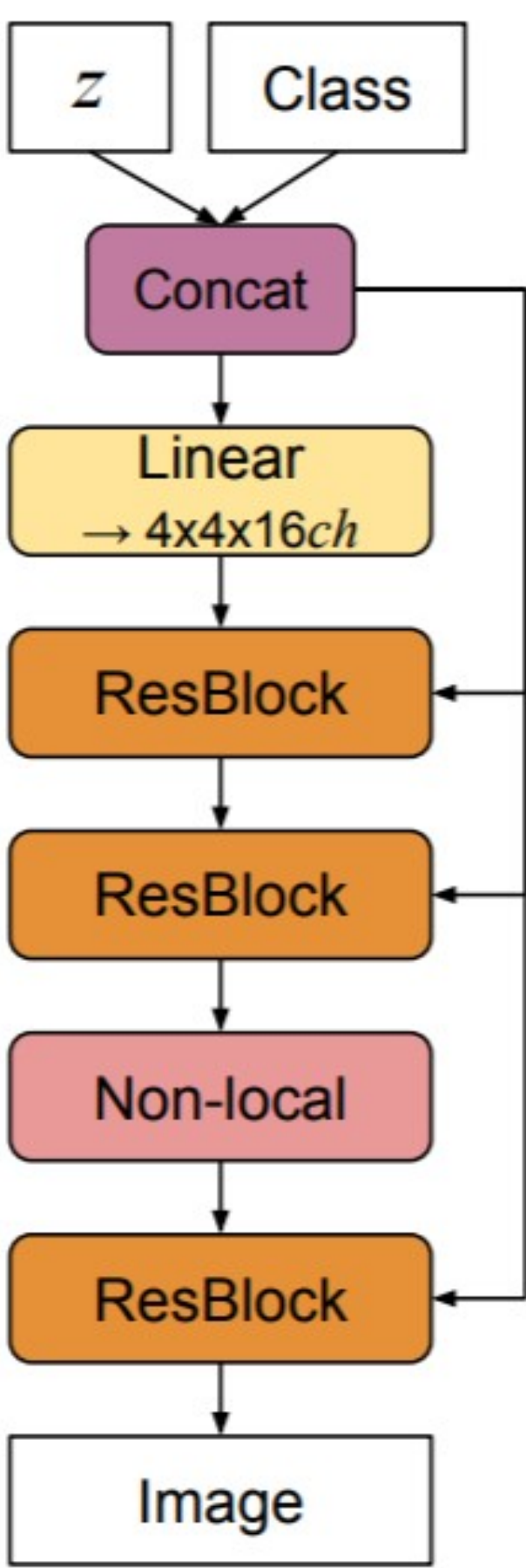→ Forward Pass
→ Backward Pass

Repeat Forward and Backward passes until convergence

Textual Description

CLIP Text Embedder 🔒

Latent vector

Generative Model (i.e. VQ-GAN, BigGAN) 🔒

CLIP Image Embedder 🔒

Embedding Similarity

CLIP: Connecting Text and Images

1. Contrastive pre-training

pepper the aussie pup → Text Encoder

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Image Encoder

BigGAN

Residual Block (ResBlock)

Architecture

z | Class

Concat

Linear → $4 \times 4 \times 16ch$

ResBlock

ResBlock

Non-local

ResBlock

Image

BatchNorm — Linear
ReLU
1x1 Conv
BatchNorm — Linear
ReLU
Drop channels
Upsample | Upsample
3x3 Conv
BatchNorm — Linear
ReLU
3x3 Conv
BatchNorm — Linear
ReLU
Add | 1x1 Conv

Concat

VQ-GAN: Vector Quantized Generative Adversarial Network

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	r	f	f
r	r	r	f
f	r	r	f

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

D CNN Discriminator
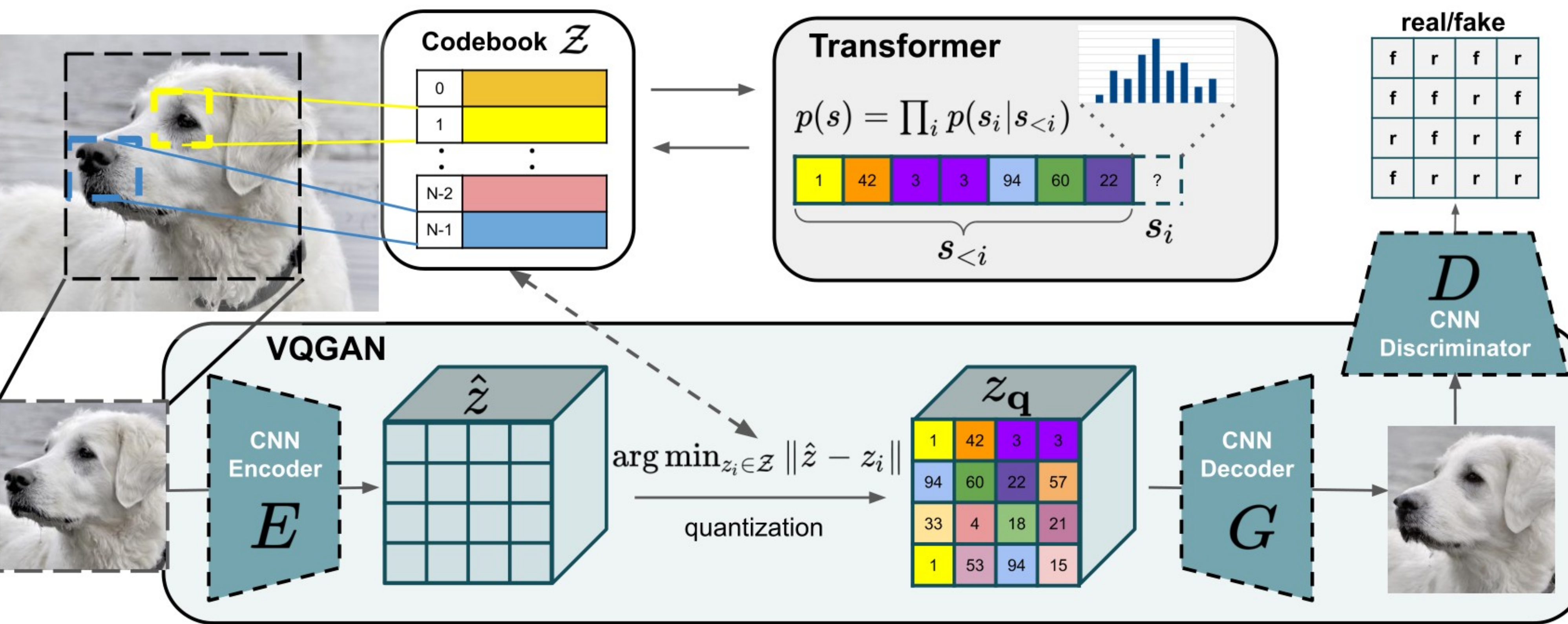
Generated Samples for Model Comparison

1: A photo of a flying penguin

2: A photo of a gold garden spider hanging on web

3: A photo of a dog with a cigarette

4: A photo of a blue bird with big beak

5: A photo of a lemon with hair and face

6: A photo of a polar bear under the blue sky

Try it yourself!

	1	2	3	4	5	6
VQ-GAN						
BigGAN						