Image Language Modeling

Prof. Bryan Pardo
AI Faculty
Head of the Interactive Audio Lab
Co-director HCI+Design Center
Northwestern University

MIDJOURNEY

PROMPT: a Singapore gen z at a thrift shop in the style of Roy Lichtenstein

Tilly Norwood is a character created using generative artificial intelligence in 2025 by Xicoia, the AI division of Particle6 Group, a production company founded by Eline Van der Velden. "AI Commissioner", the first project to feature the Norwood character, was criticised by reviewers for *The Guardian*, *PC Gamer*, and *The A.V. Club*. A press release that talent agencies expressed interest in representing the character attracted strong criticism from Hollywood actors and firms, prompting allegations of personality rights violations and arguments over the impact of the character on production costs in the media industry.

History [edit]

Tilly Norwood

A 2025 AI-generated image of Norwood

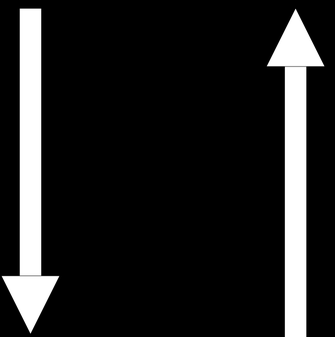
First appearance	"AI Commissioner" 2025
Created by	Eline Van der Velden

How do they work?

Let’s talk Language Models…

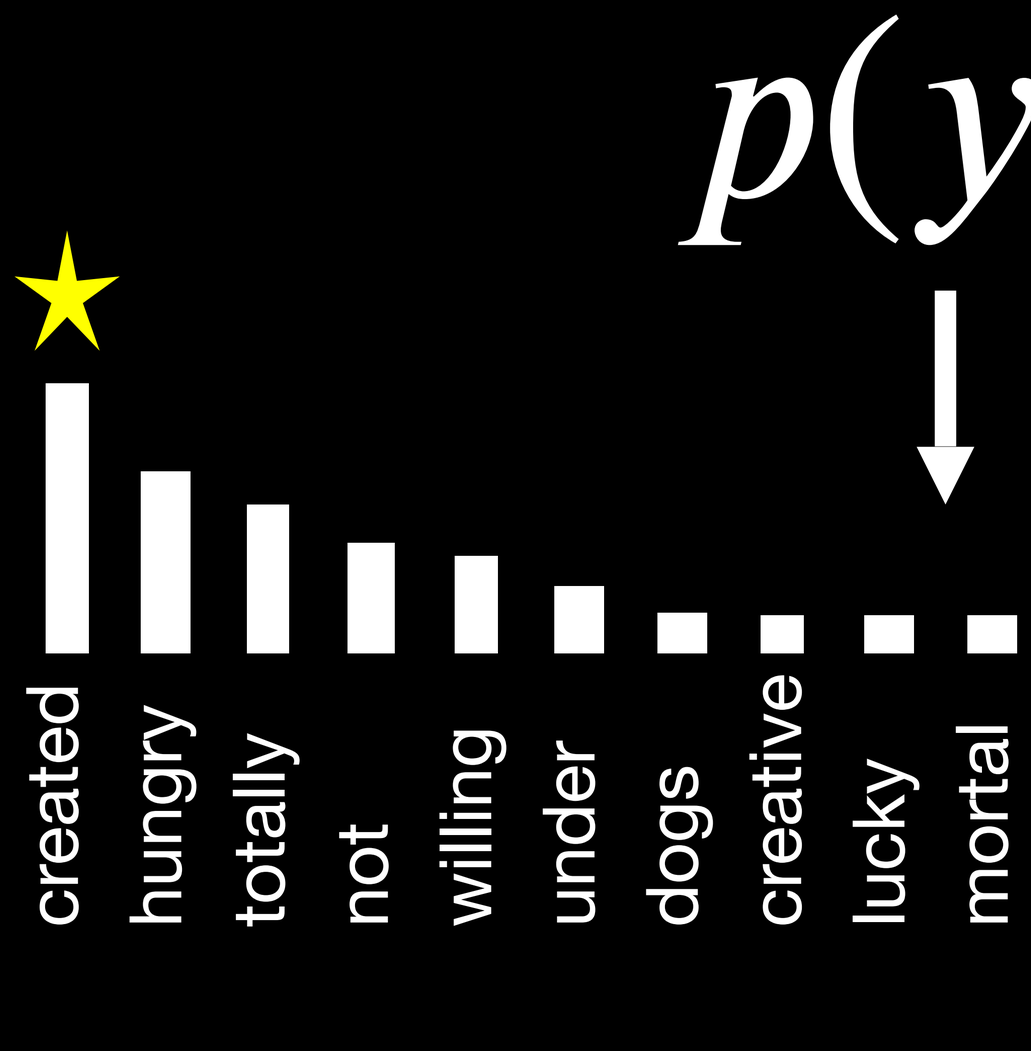
(We won’t be discussing: Diffusion, Flow matching,
latent diffusion, structured state space)

$$p(y \mid x)$$



The output

The input
(aka “conditioning”)



We hold these truths to be self–evident, that all men are…

$$p(y \mid x)$$

We hold these truths to be self–evident, that all men are created…

Labels? We don’t need no stinking labels

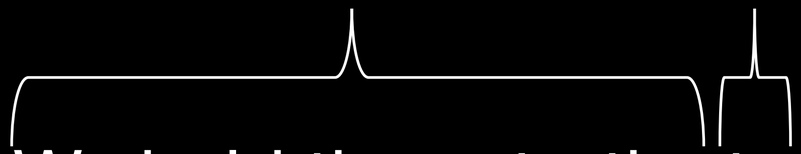
$$p(y \mid x)$$

x y

We hold these truths to be self–evident, that all men are created equal

$$p(y \mid x)$$

x y



We hold these truths to be self–evident, that all men are created equal

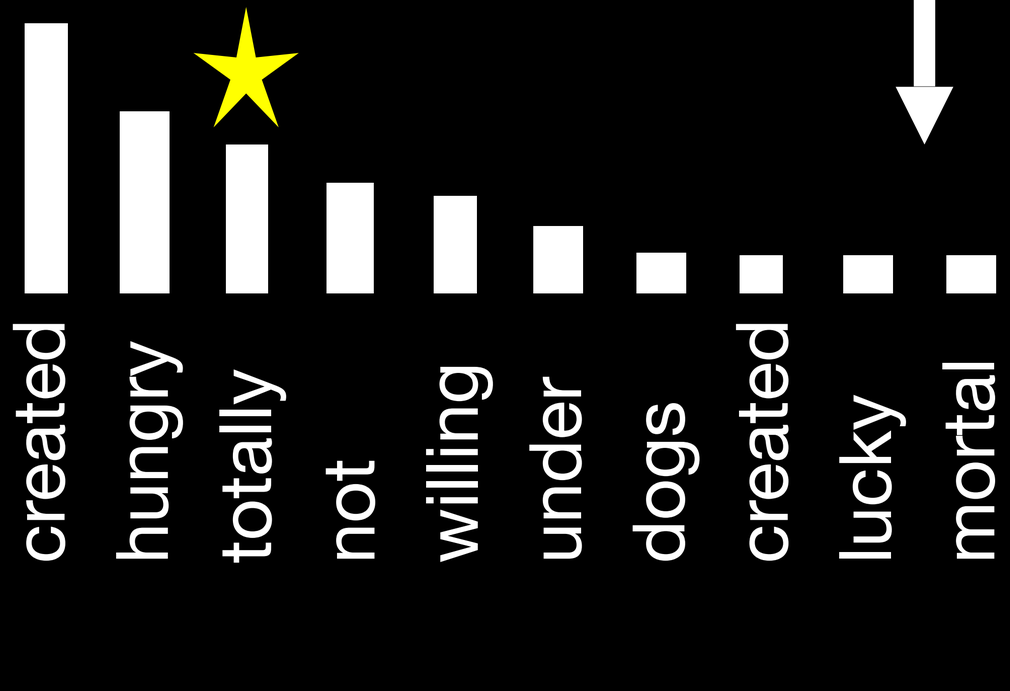
$$p(y \mid x)$$

x y

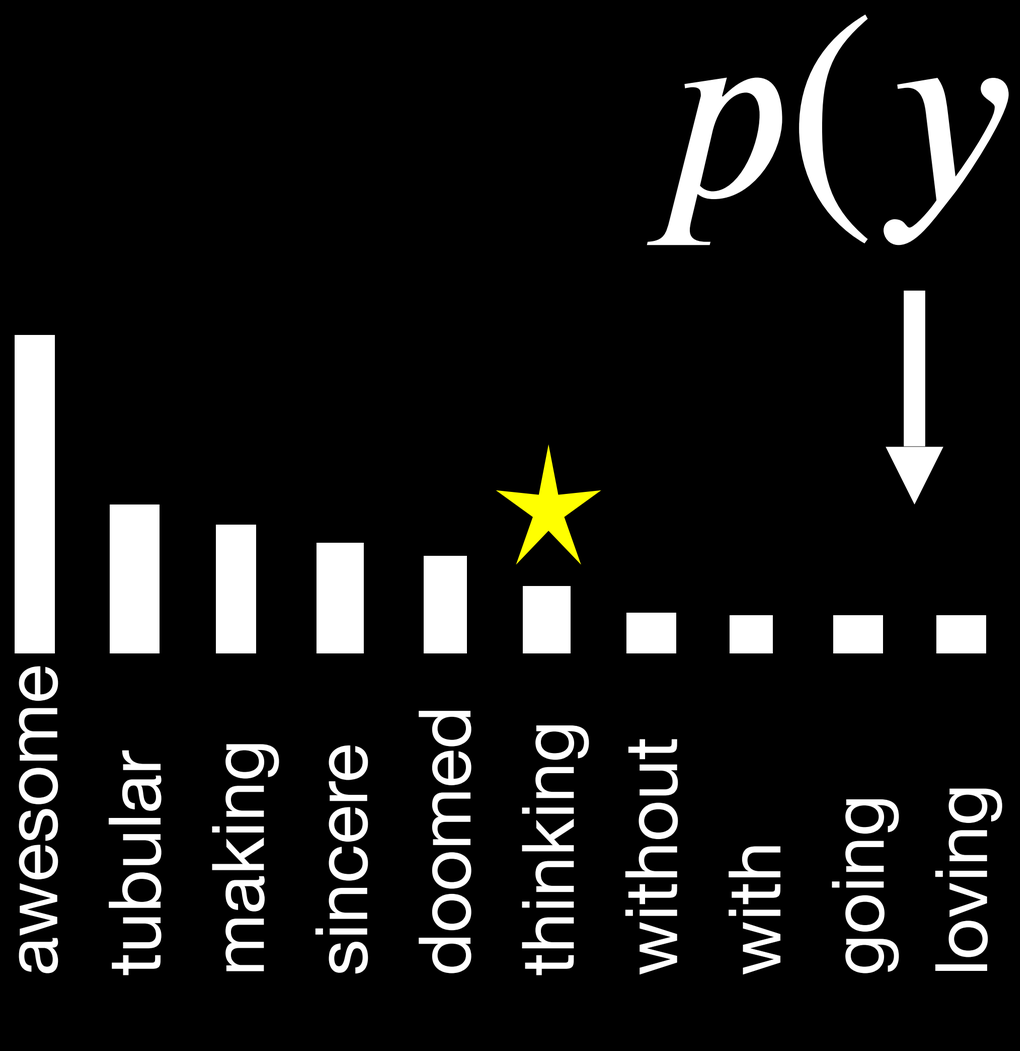
We hold these truths to be self–evident, that all men are created equal

Random walk = creativity?

$$p(y \mid x)$$

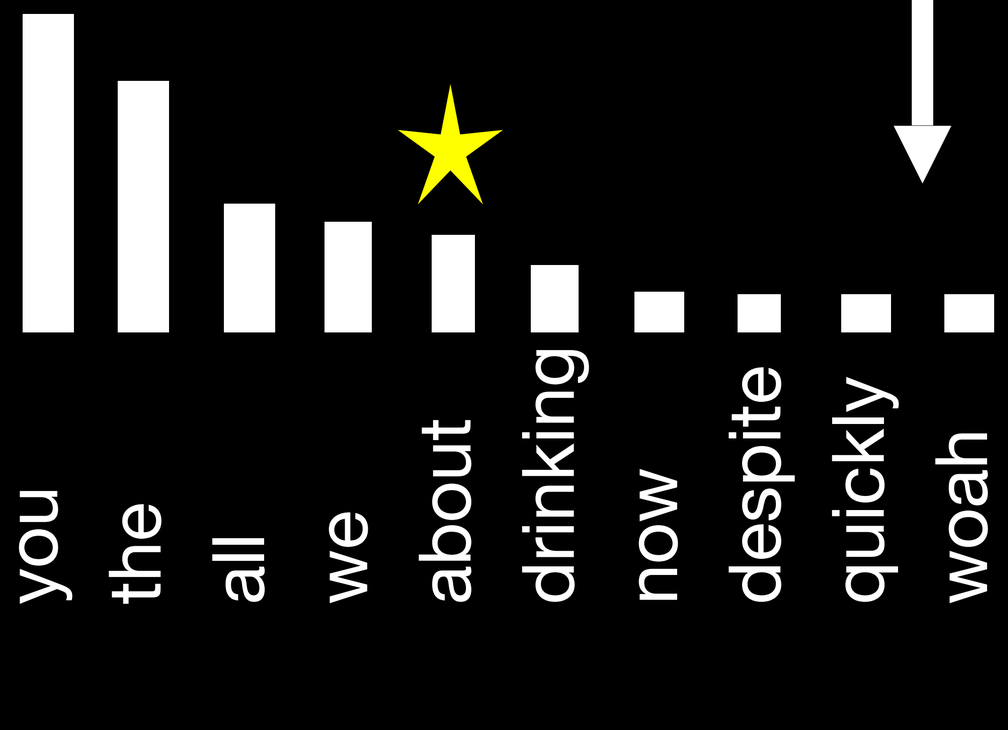


We hold these truths to be self–evident, that all men are…

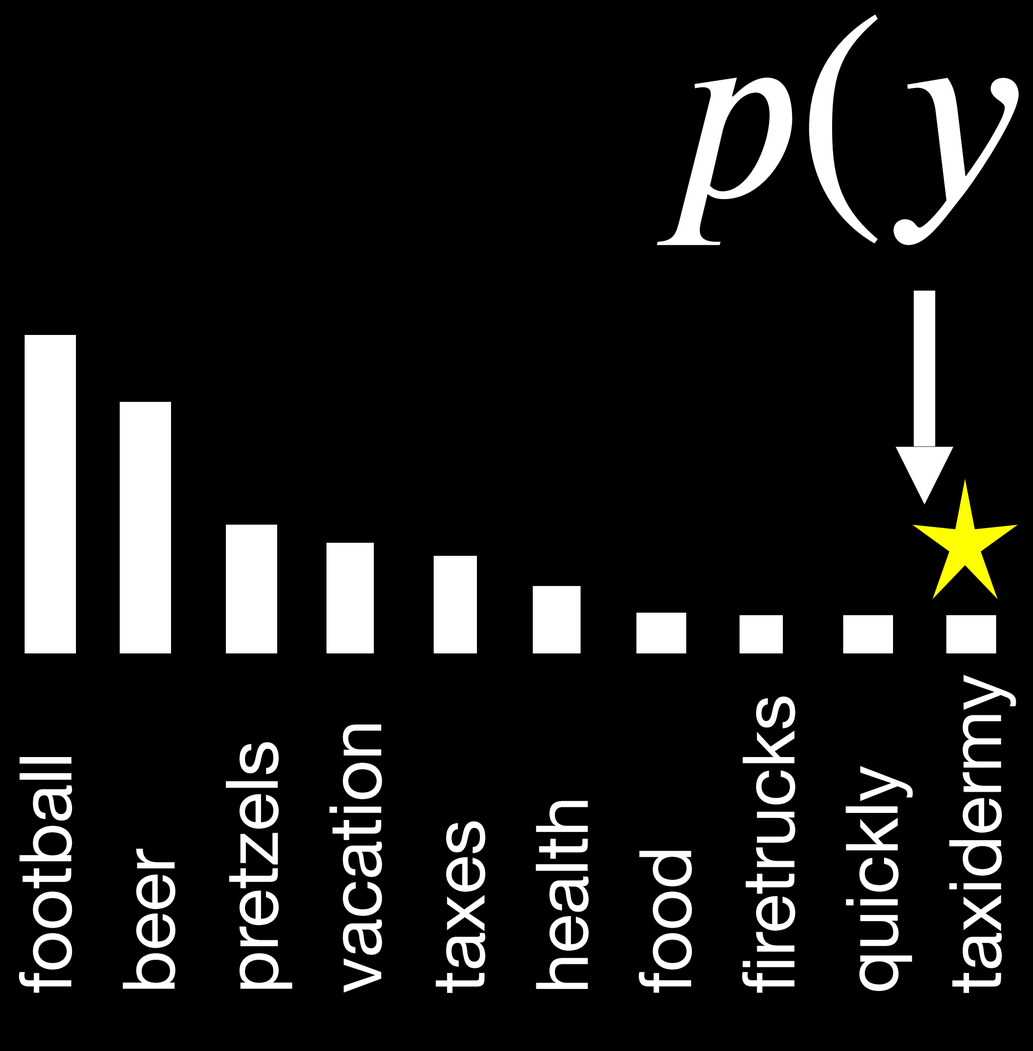


We hold these truths to be self–evident, that all men are totally…

$$p(y \mid x)$$



We hold these truths to be self–evident, that all men are totally thinking…



We hold these truths to be self–evident, that all men are totally thinking about…

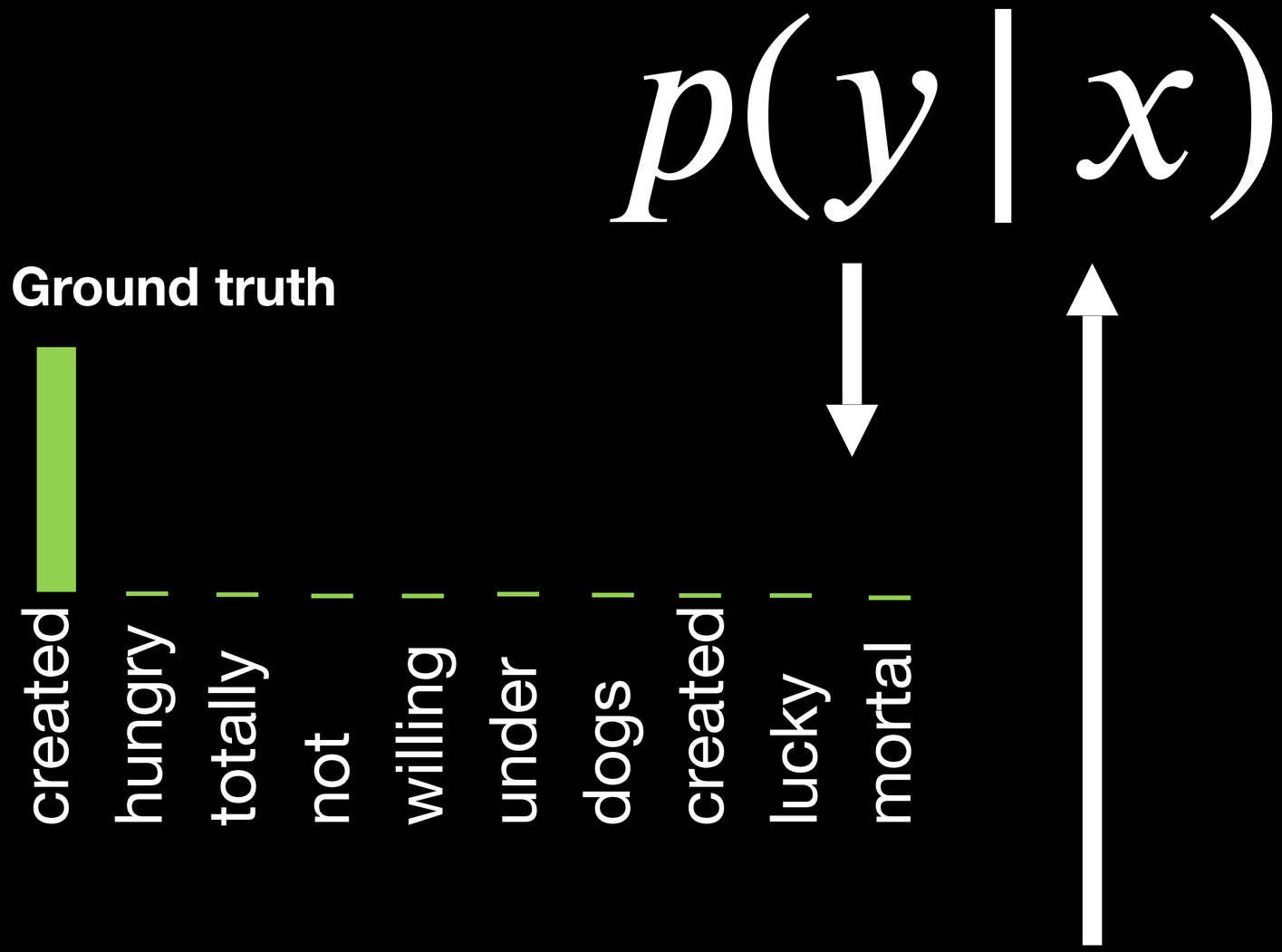
We hold these truths to be self–evident, that all men are totally thinking about **taxidermy.**

Their loss is our gain

$$p(y \mid x)$$

Initial distribution

We hold these truths to be self–evident, that all men are…


$p(y \mid x)$
Ground truth
created
hungry
totally
not
willing
under
dogs
created
lucky
mortal
We hold these truths to be self–evident, that all men are…

Teacher forcing

Teacher forcing

- Don’t penalize for past mistakes
- Measure loss as if everything up to this token was correct
- Allows parallelization of training
  - Current result doesn’t depend on past success/failure
  - Take any sentence and make it into a training batch

Take any sentence and make it into a training batch

- p(y|x) = p(any | take)
- p(y|x) = p(sentence | take any)
- p(y|x) = p(and | take any sentence)
- p(y|x) = p(make | take any sentence and)
- p(y|x) = p(it | take any sentence and make)
- p(y|x) = p(into | take any sentence and make it)
- p(y|x) = p(a | take any sentence and make it into)
- p(y|x) = p(training | take any sentence and make it into a)
- p(y|x) = p(batch | take any sentence and make it into a training)

We don’t actually want to learn perfectly

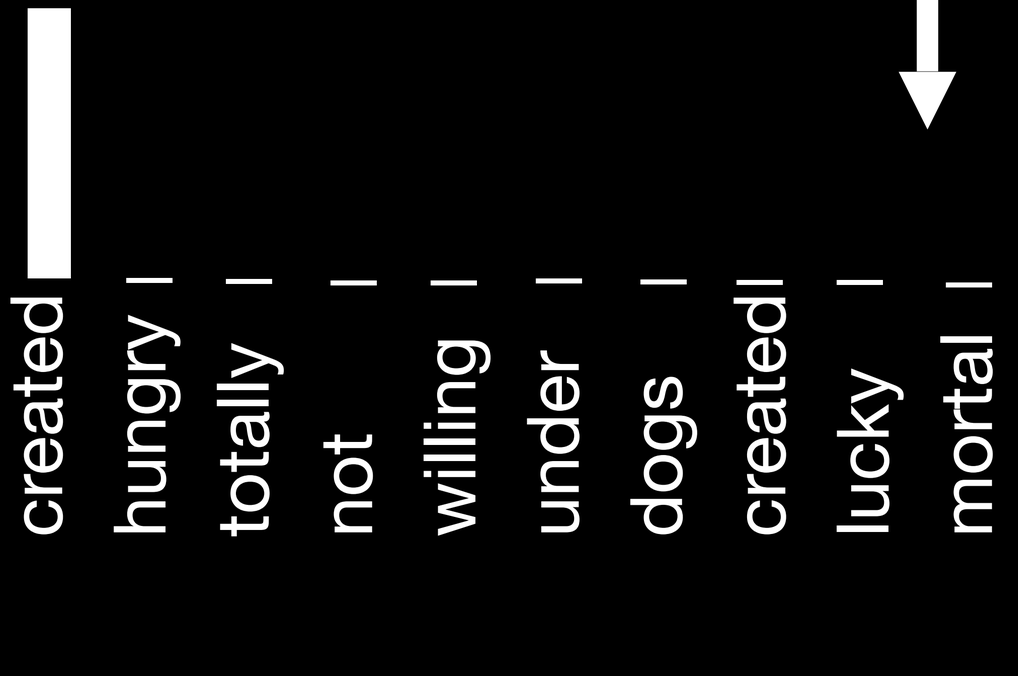
$$p(y \mid x)$$

Learned distribution

We hold these truths to be self–evident, that all men are…

$$p(y \mid x)$$

Overfit (mode collapse)



We hold these truths to be self–evident, that all men are…

It is a balance…

We hold these truths to be self–evident, that all men are…

Gibberish: …but are not but are but are not but are…

Creative: …totally thinking about taxidermy.

Copyright violating: ...that all men are created equal, but some are more equal than others.

Correct?: …created equal, that they are endowed, by their Creator, with certain unalienable rights, that among these are Life, Liberty, and the pursuit of Happiness.

It’s a balance

- Learn the distribution too well and only reproduce your training data
- Learn it too poorly and your output is gibberish
- Sample with argmax and violate copyright
- Sample more randomly and risk hallucination

A token example

What if we replace words with “tokens”?

- Consider all variants: “talk”, “talking”, “talks”, “talked”
- Break text into frequently-occurring parts
- These are tokens

Can we tokenize other things?

Language Modeling Images

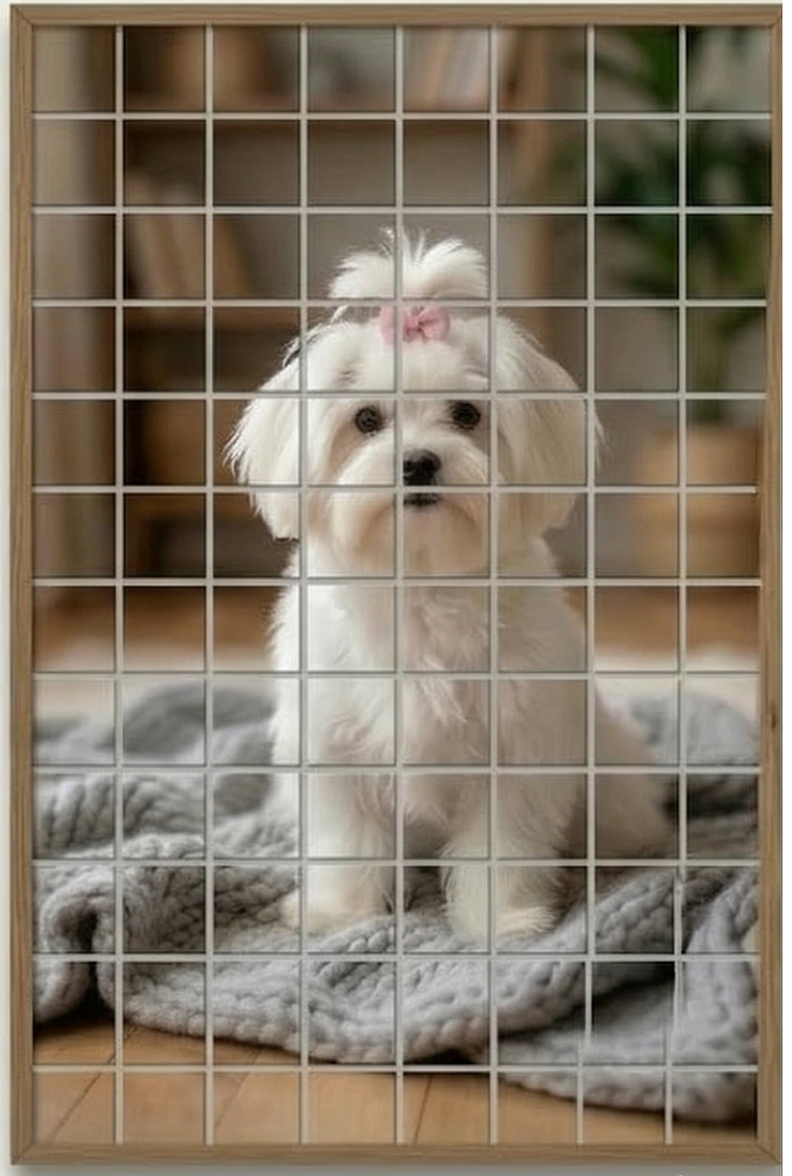
Big White Dog

Poodle Running

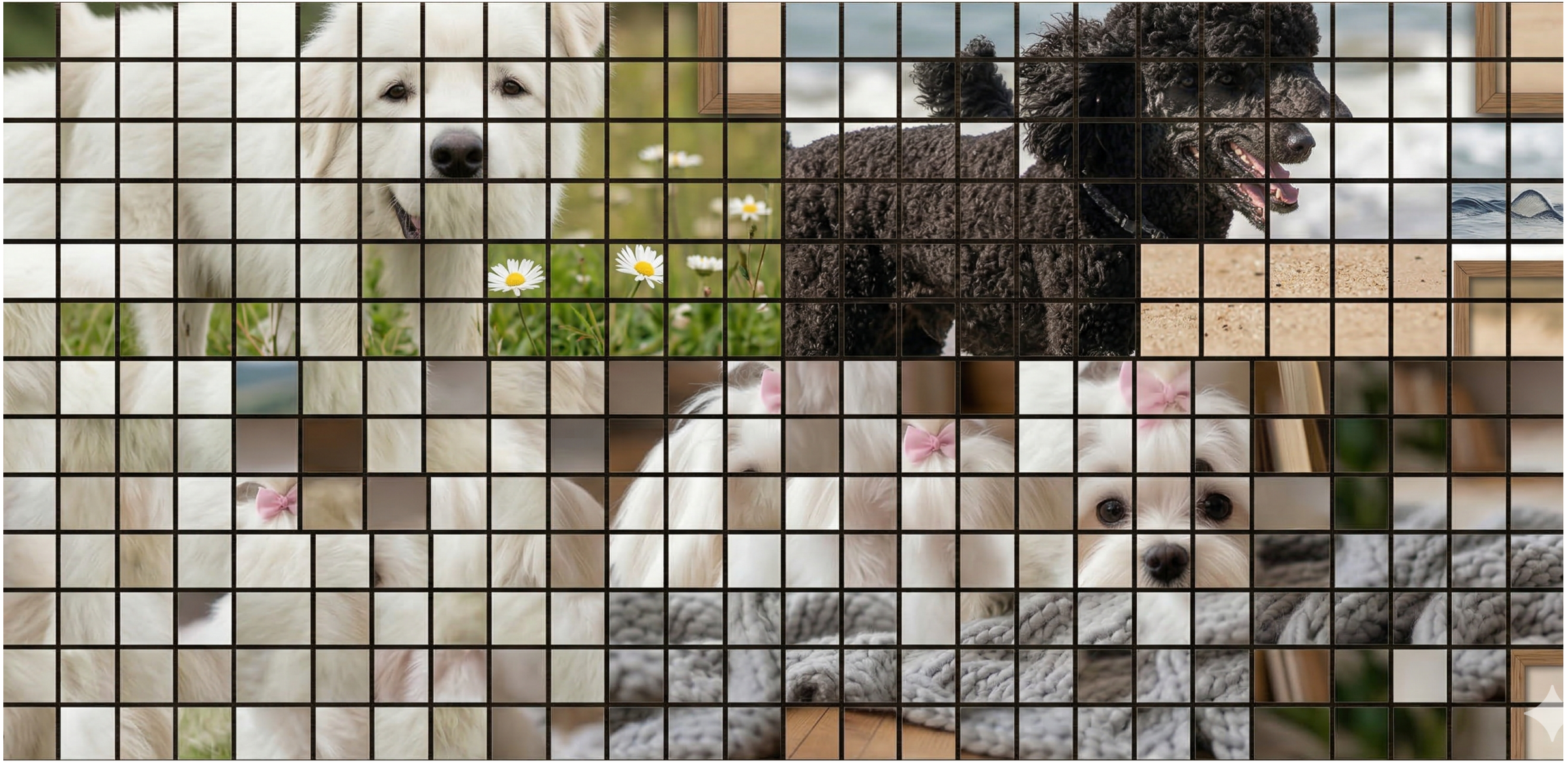
Maltese Dog Sitting

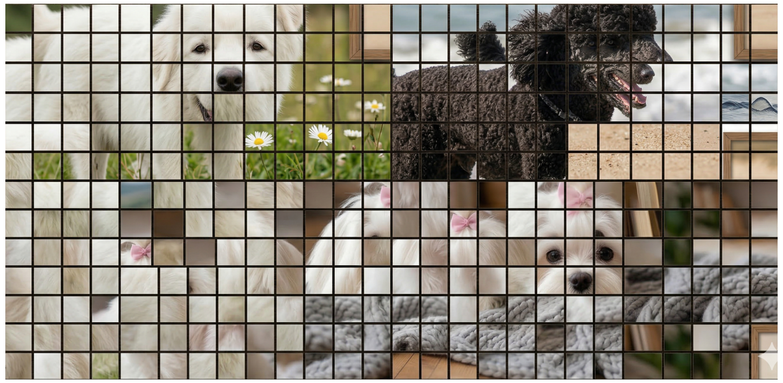
Big White Dog

Poode Running

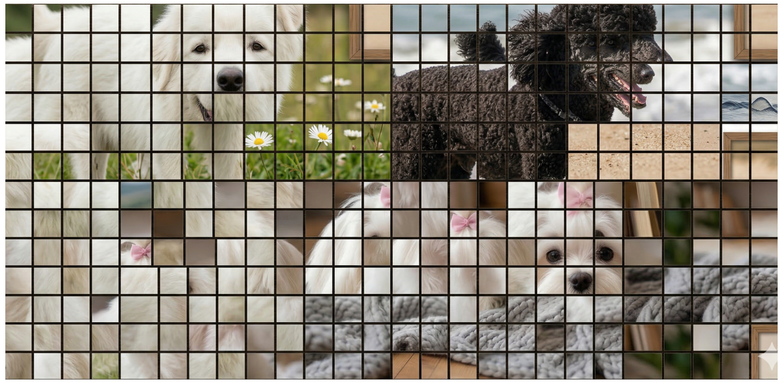


Maltese Dog Sitting

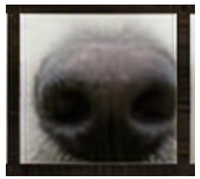


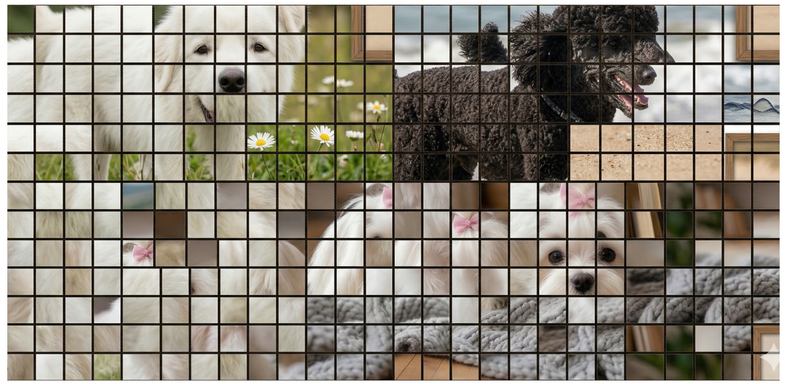


$x \sim p_x$ Sample from dataset



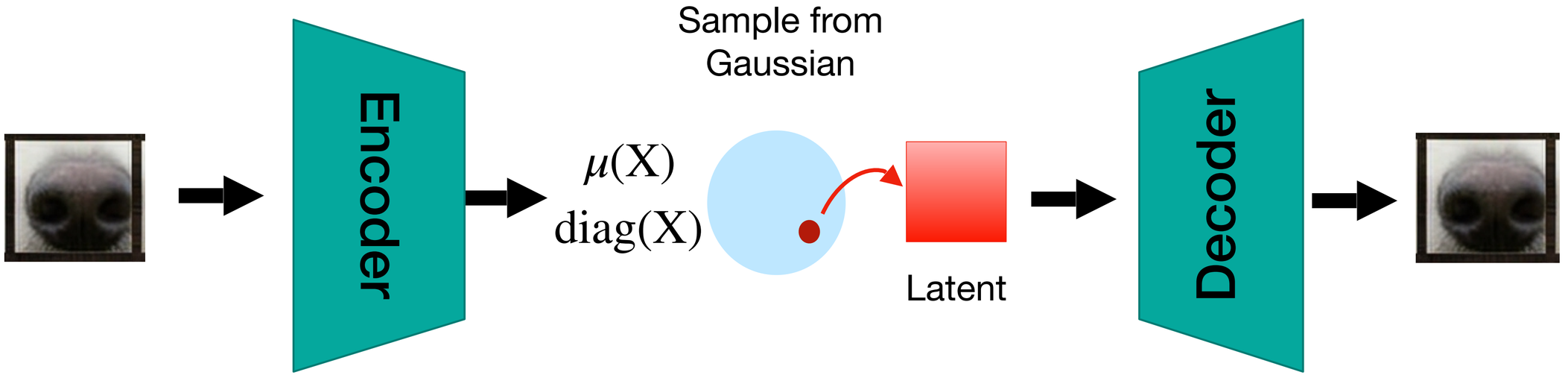
$x \sim p_x$ Sample from dataset



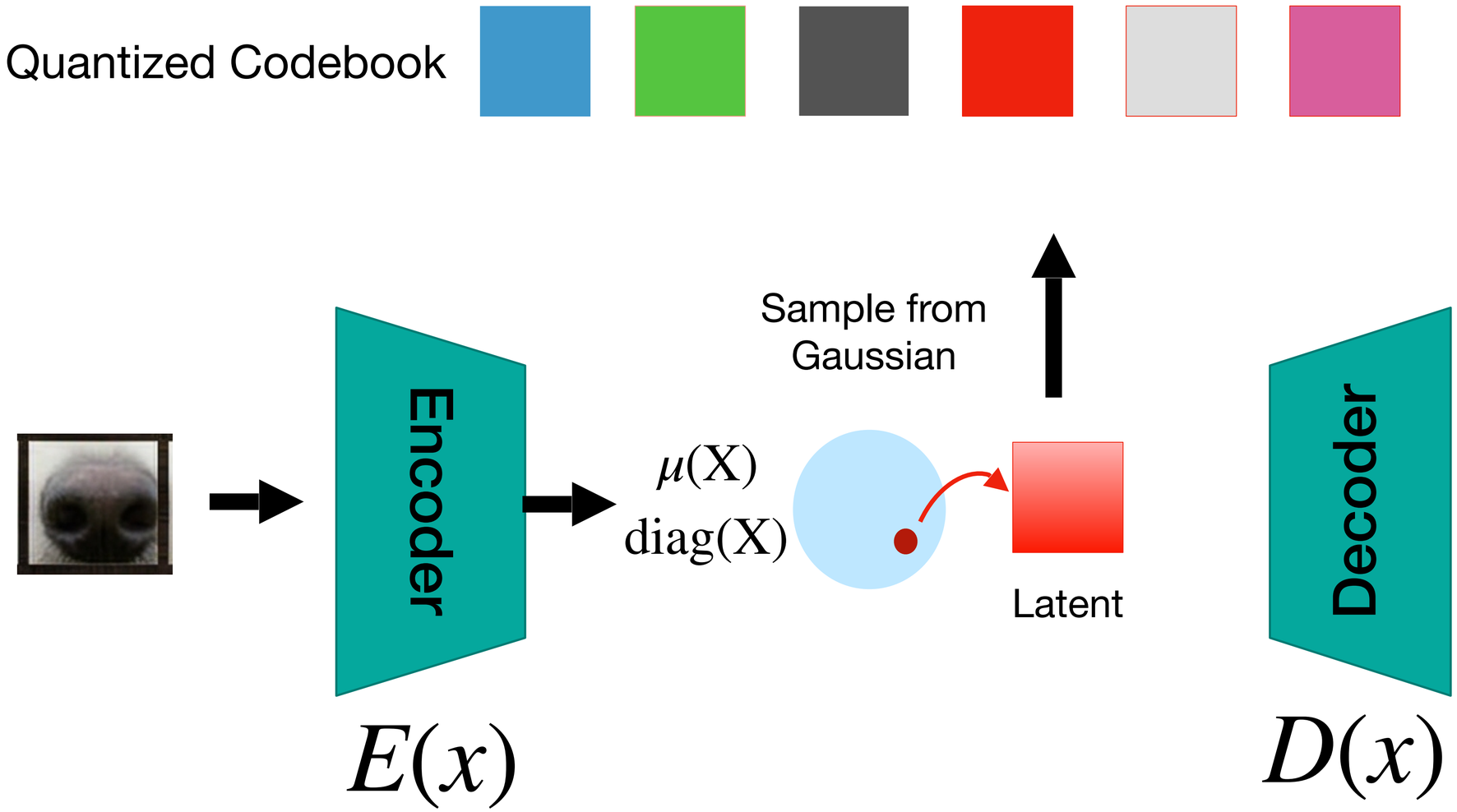


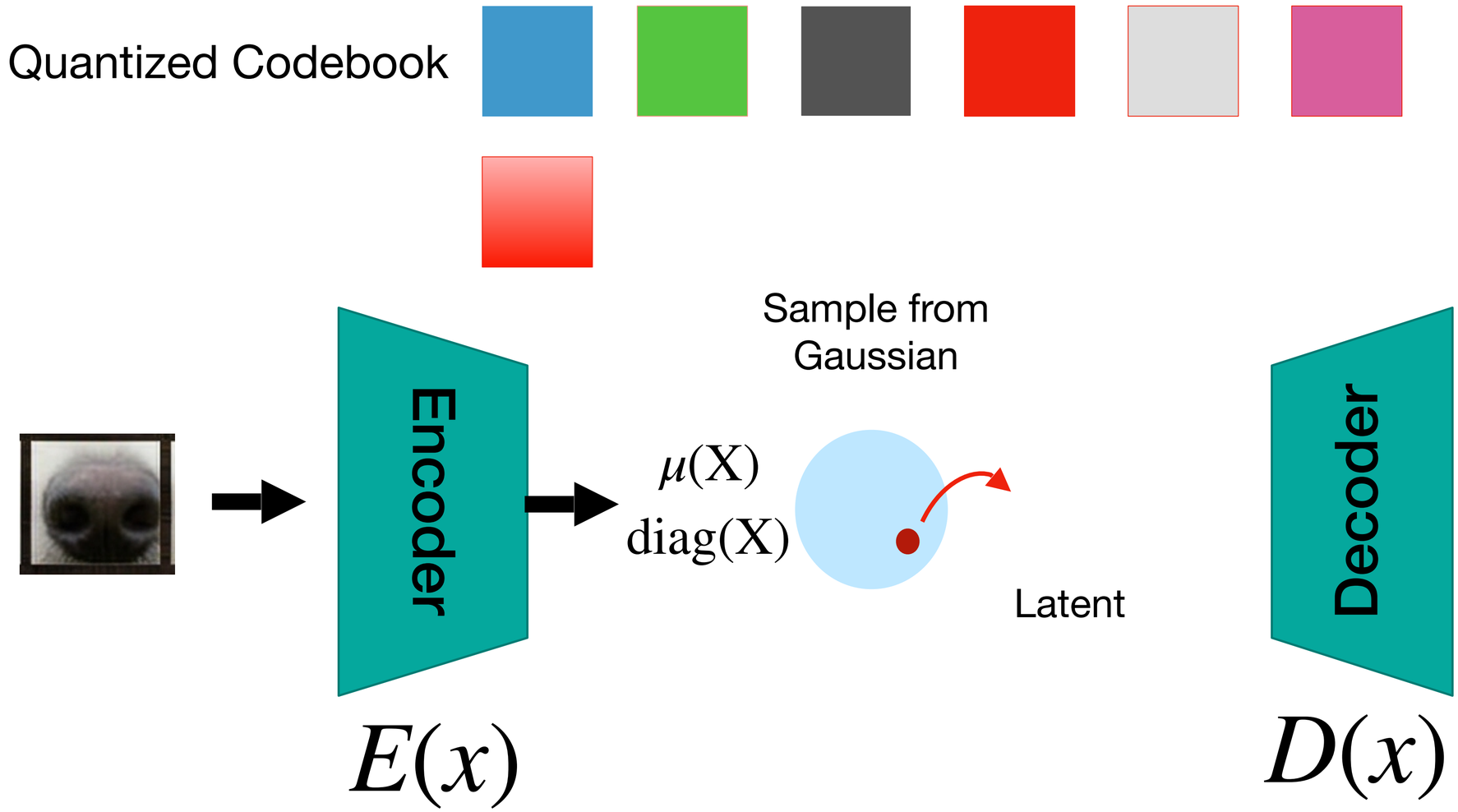
$$\text{loss } \mathscr{L} = \left\| x - D(E(x)) \right\|^2$$

$x \sim p_x$ Sample from dataset



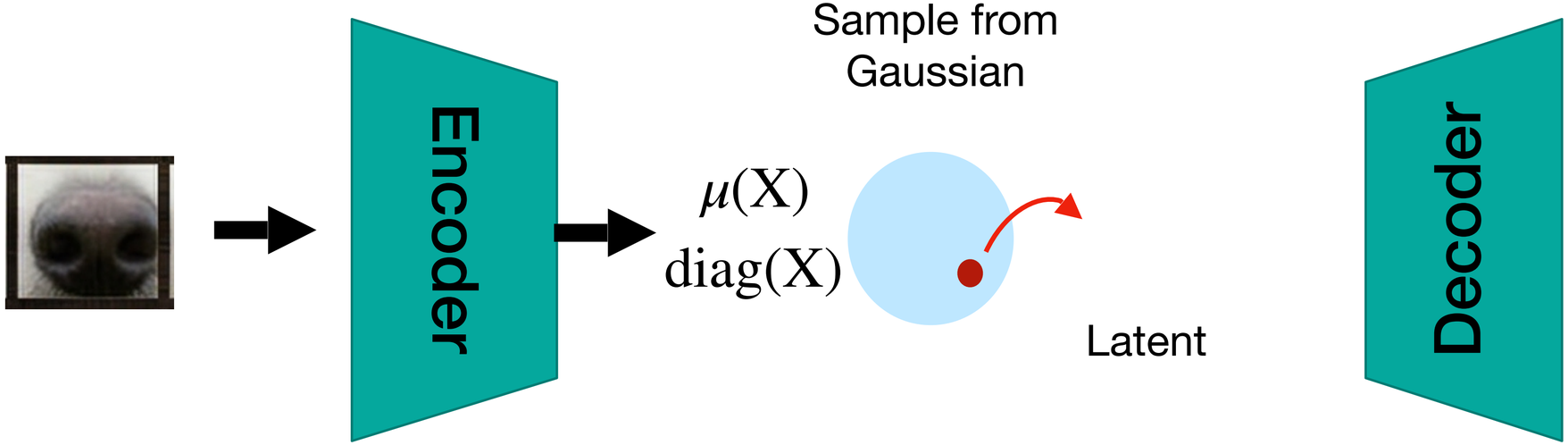
$E(x)$ $D(x)$


Quantized Codebook
Sample from
Gaussian
Encoder
$\mu(X)$
diag(X)
Latent
Decoder
$E(x)$
$D(x)$


Quantized Codebook
Sample from
Gaussian
Encoder
$\mu(X)$
diag(X)
Latent
Decoder
$E(x)$
$D(x)$

Quantized Codebook

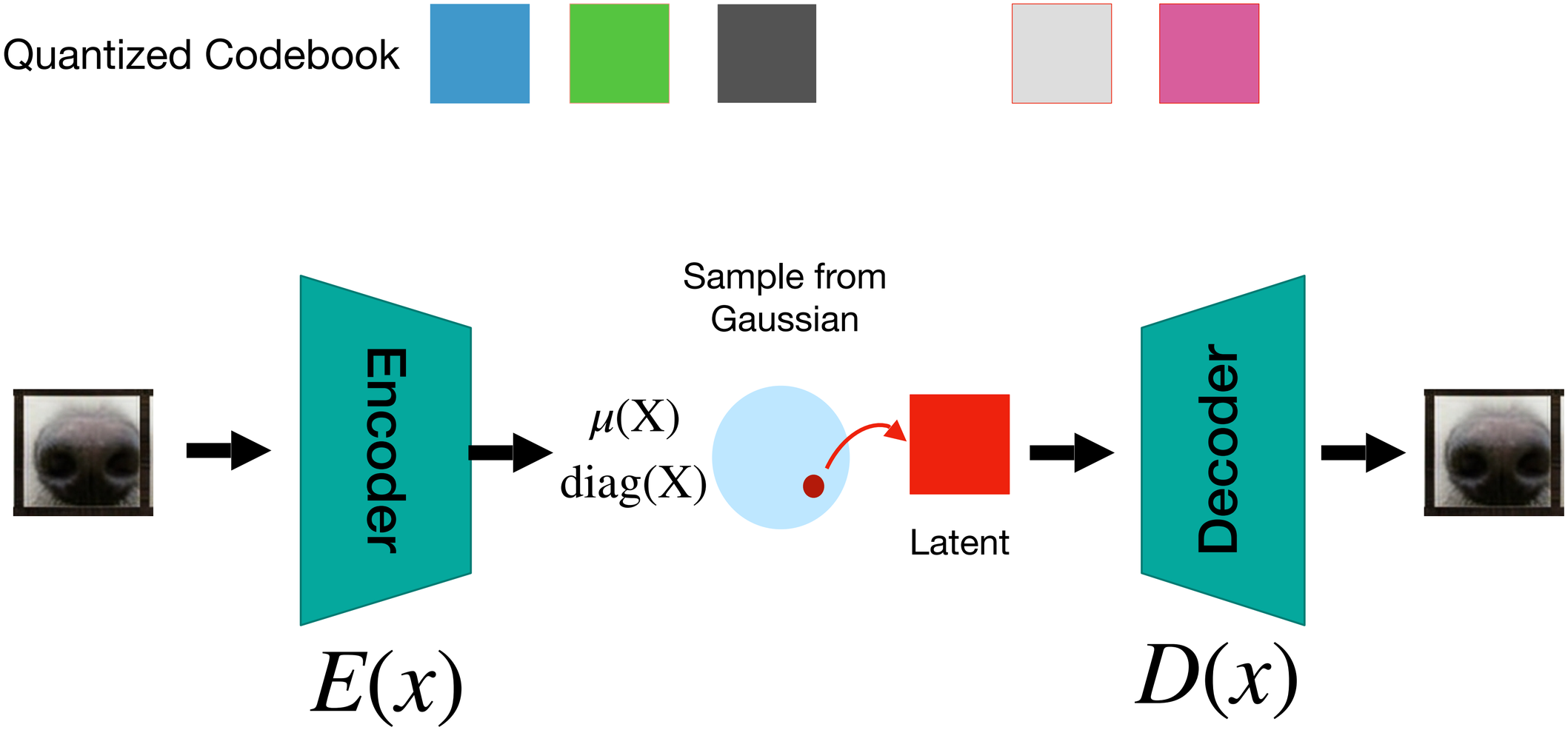
Sample from
Gaussian

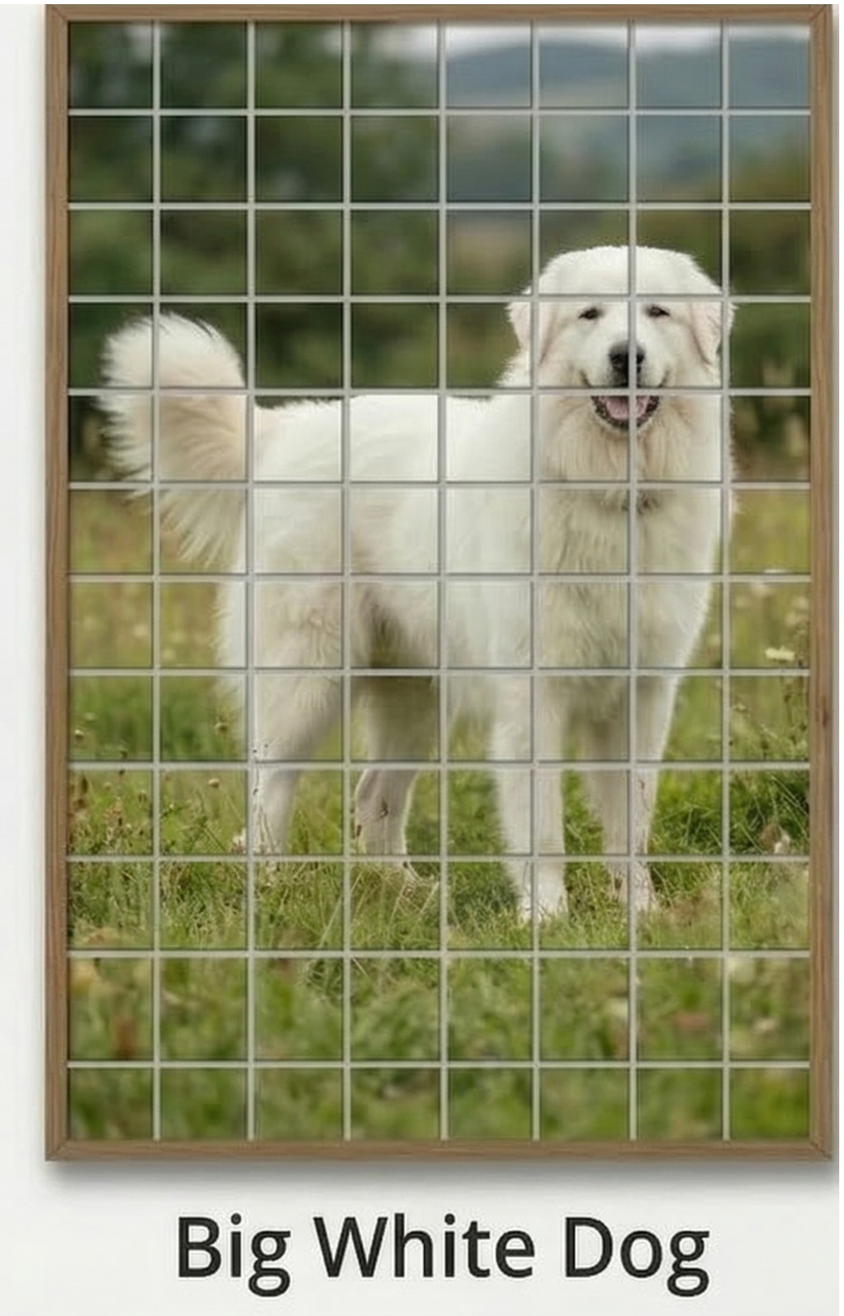


$E(x)$

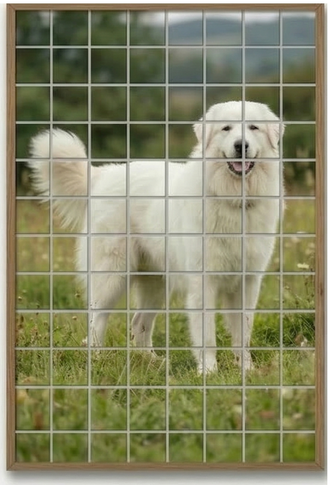
$D(x)$

Quantized Codebook

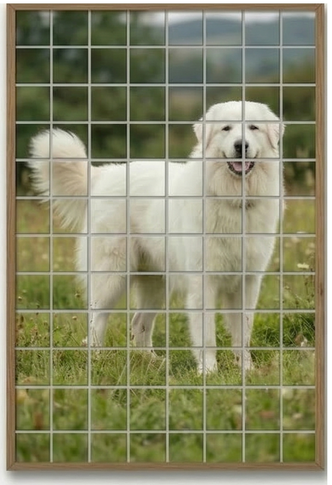



Big White Dog

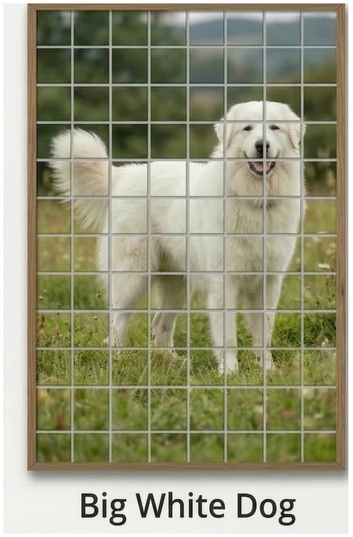
Big White Dog



Big White Dog



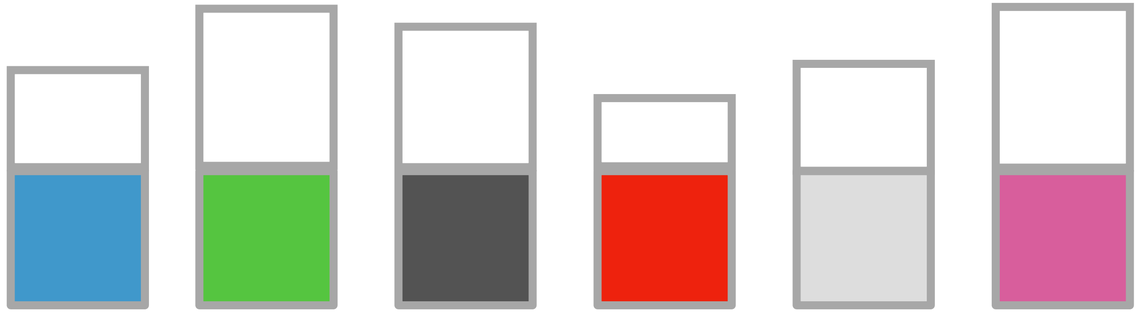
Big White Dog

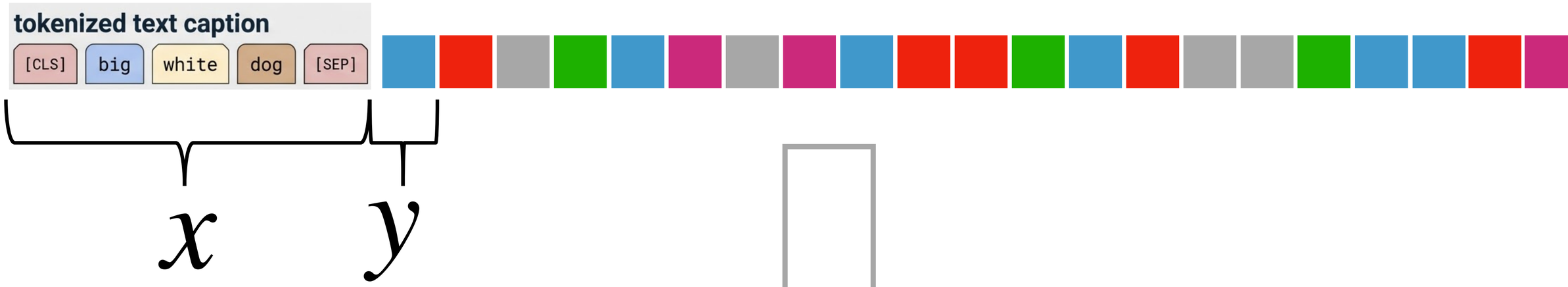


Big White Dog

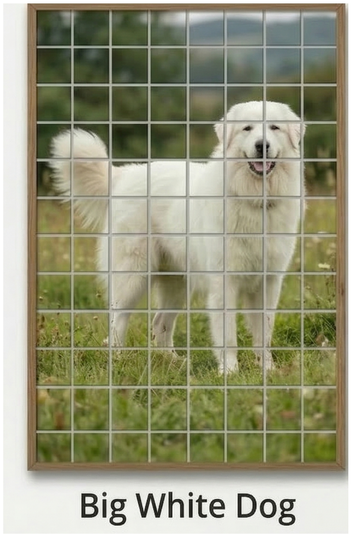
$$p(y \mid x)$$

Model output





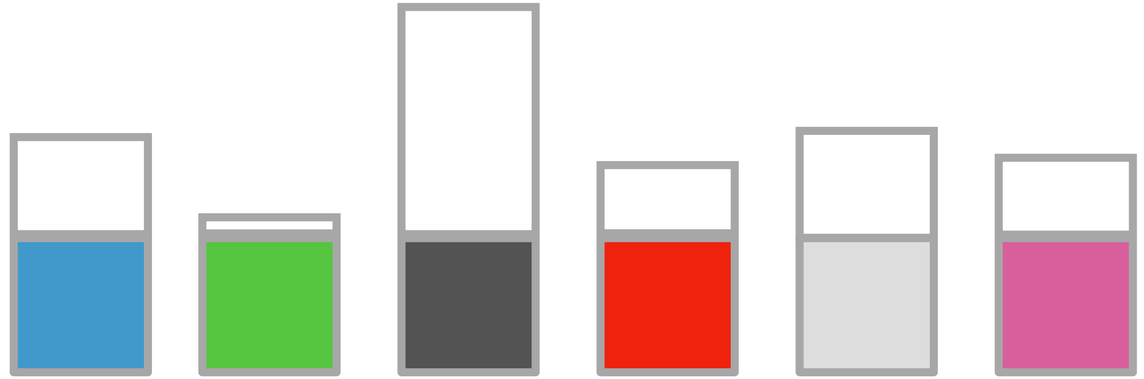
Ground truth



Big White Dog

$$p(y \mid x)$$

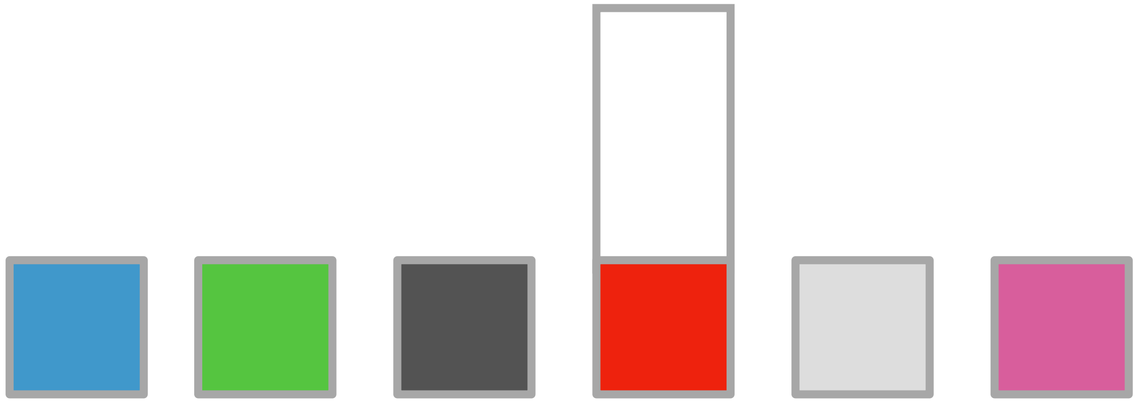
Model output

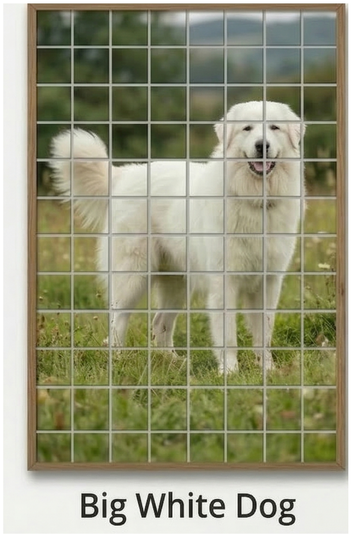


tokenized text caption

x y

Ground truth

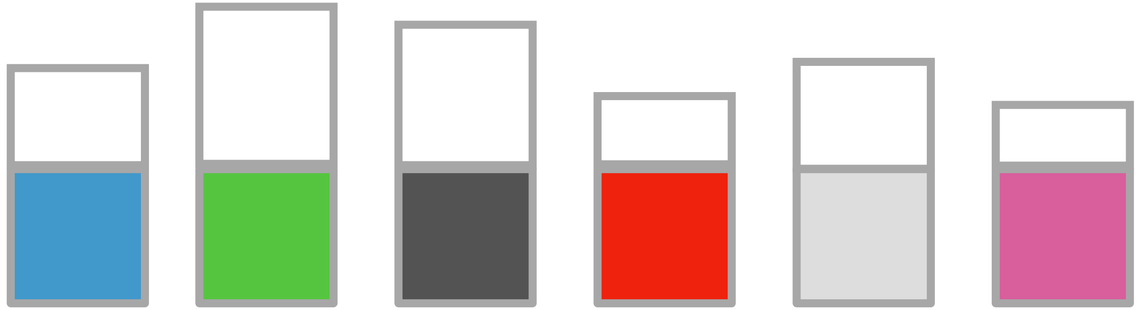




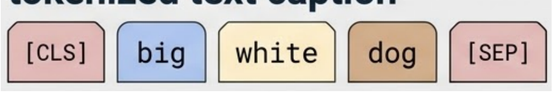
Big White Dog

$$p(y \mid x)$$

Model output

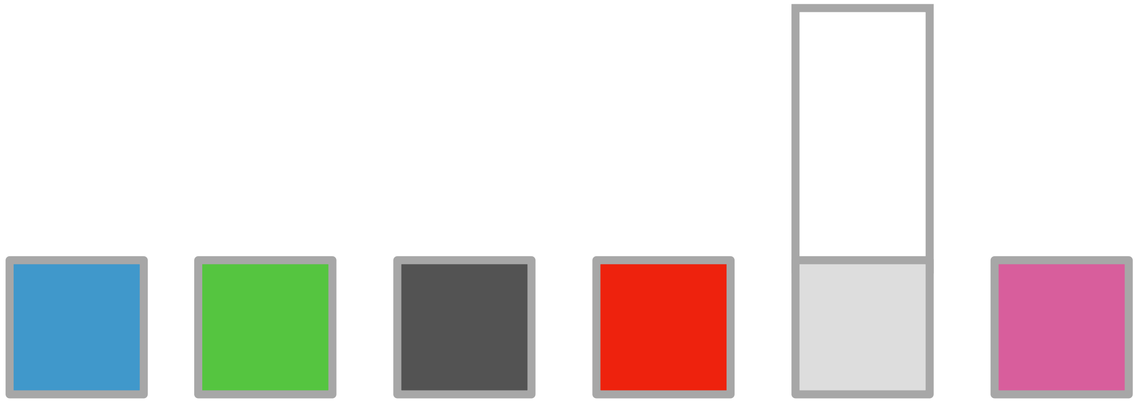


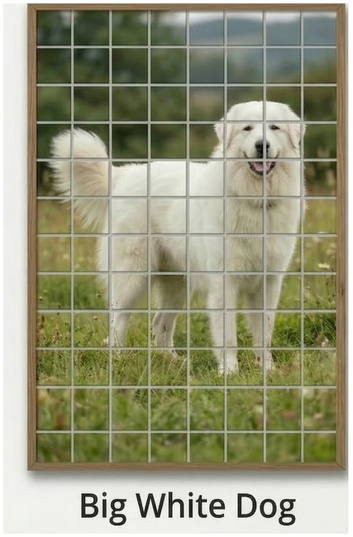
tokenized text caption



x y

Ground truth



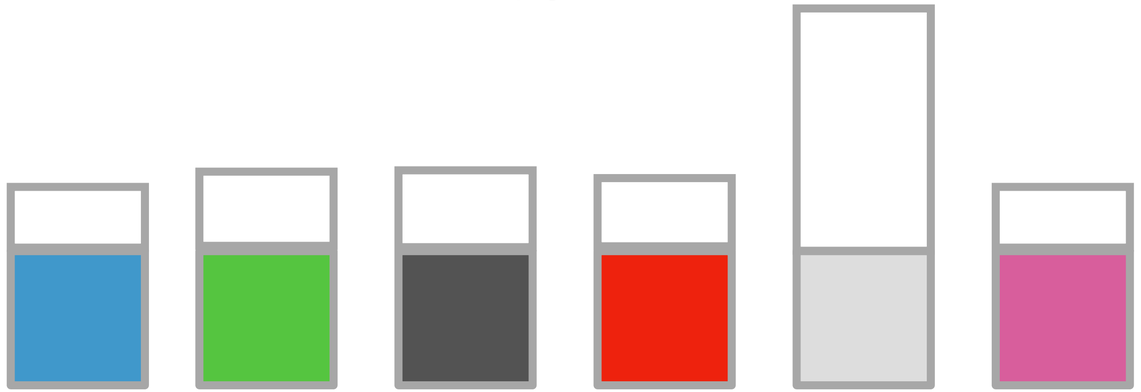


Big White Dog

$$p(y \mid x)$$

Update with cross entropy loss

Model output



Ground truth

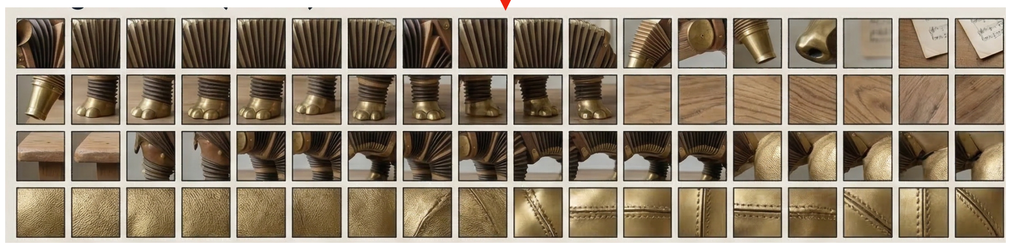
tokenized text caption

[CLS] a tapir made of accordion [SEP]

Trained language model

$$p(y \mid x)$$

$D(x)$ Decoder



Ramesh, Aditya, et al. "Zero-shot text-to-image generation." *International conference on machine learning*, 2021.