

---

# Intelligent Agents

## Language-Vision-Models: DALL-E

Prof. Dr. Ralf Möller

Universität zu Lübeck

Institut für Informationssysteme



# Generate Images from Text – Naïve Approach

TEXT PROMPT

an armchair in the shape of an avocado [...]

AI-GENERATED IMAGES



1. Concatenate the set of text tokens with the unrolled set of pixel values in a corresponding image (typically unrolled top left to bottom right).
2. Given this sequence of text and pixel values, we can factor the distribution  $p(x|y)$  autoregressively:

$$p(x|y) = p(x_1, x_2, x_3, \dots | y) = p(x_1 | y) p(x_2 | x_1, y) p(x_3 | x_1, x_2, y) \dots$$

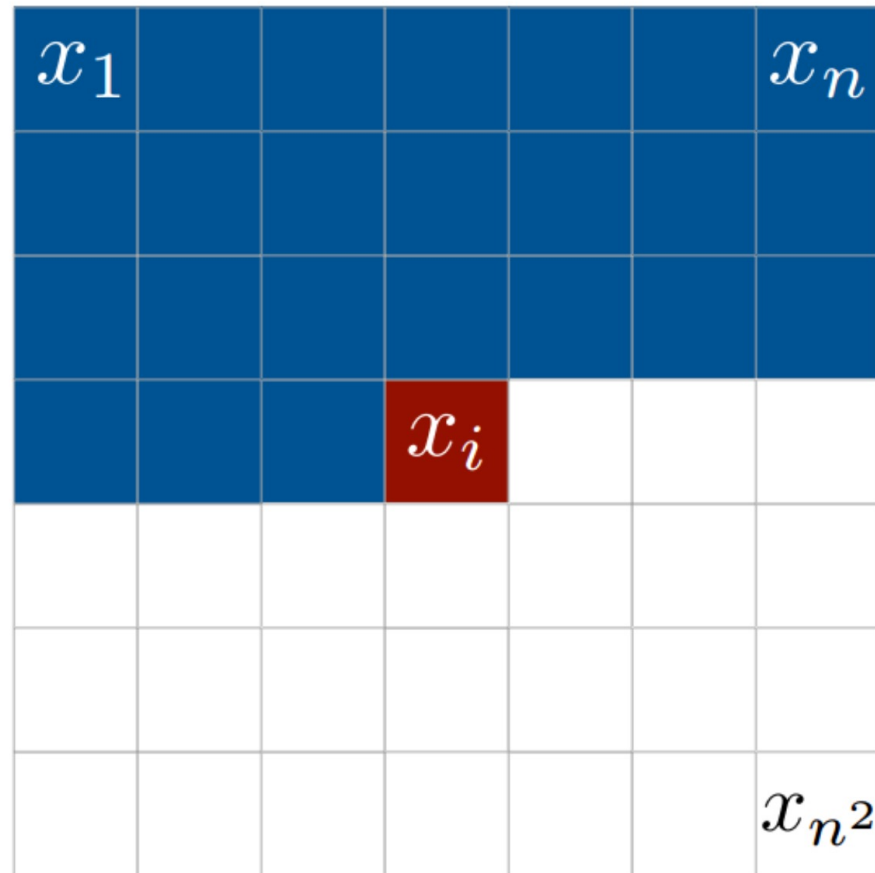
Here  $x_i$  is the  $i$ th pixel value in the unrolled image.

3. We now estimate  $p(x|y)$  by running maximum likelihood estimation on any autoregressive sequence model (e.g. LSTM or Transformer) over each of these  $p(x_i | x_{i-1}, x_{i-2}, \dots, x_2, x_1, y)$  factors.

That is to say, we want to train a model to predict the next pixel value in an image, given some text and all previous pixel values.

# Use RNN decoder to generate images??

An armchair in the shape of [...]



# Generate Images from Text – Naïve Approach

TEXT PROMPT

an armchair in the shape of an avocado [...]

AI-GENERATED IMAGES



1. Concatenate the set of text tokens with the unrolled set of pixel values in a corresponding image (typically unrolled top left to bottom right).
2. Given this sequence of text and pixel values, we can factor the distribution  $p(x|y)$  autoregressively:

$$p(x|y) = p(x_1, x_2, x_3, \dots | y) = p(x_1 | y) p(x_2 | x_1, y) p(x_3 | x_1, x_2, y) \dots$$

Here  $x_i$  is the  $i$ th pixel value in the unrolled image.

3. We now estimate  $p(x|y)$  by running maximum likelihood estimation on any autoregressive sequence model (e.g. LSTM or Transformer) over each of these  $p(x_i | x_{i-1}, x_{i-2}, \dots, x_2, x_1, y)$  factors.

That is to say, we want to train a model to predict the next pixel value in an image, given some text and all previous pixel values.

# Acknowledgements

---

## Zero-Shot Text-to-Image Generation

Authors: Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen,  
and Ilya Sutskever

Open AI (ICML2021)

Presentation from: Adam Kutchak, George Lu, Fernando Treviño, and Sarah Wilson

Zero-Shot Text-to-Image Generation. Aditya Ramesh, Mikhail Pavlov,  
Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya  
Sutskever Proceedings of the 38th International Conference on Machine  
Learning, PMLR 139:8821-8831, **2021**.



# Introduction

- Generate Images from text captions
- 12 billion parameters version of GPT-3
- Dataset comprised of 3.3 million text - image pairs
- Combine unrelated concepts

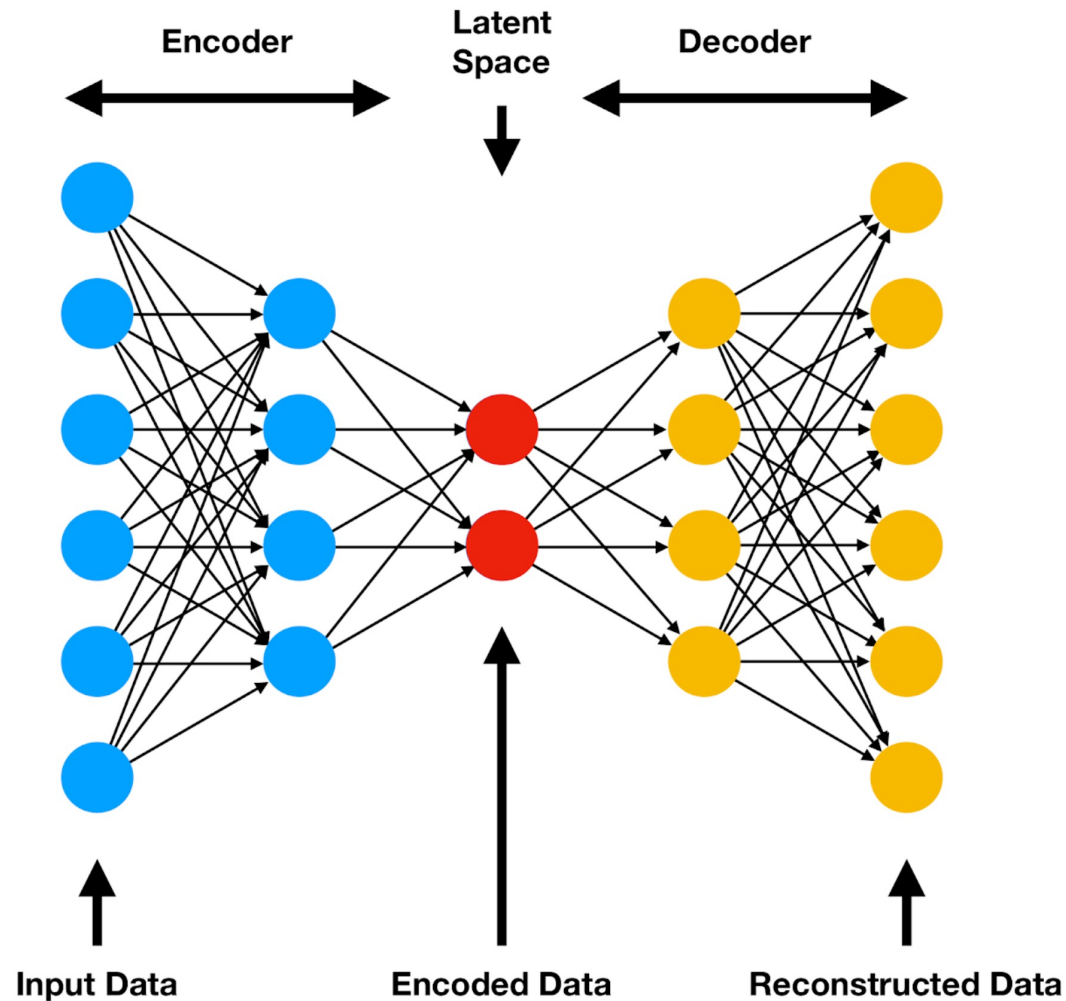


# Related Works

---

- Autoencoder - (encoder - decoder)
- Variational Autoencoders (continuous state space)
- Vector Quantized-Variational AutoEncoder VQ-VAE (discrete quantized state space)

# Related Work - Autoencoder





# Related Work - Autoencoder

## Encoder



image to  
discrete codes



56	73	67	23	81	19	...
----	----	----	----	----	----	-----

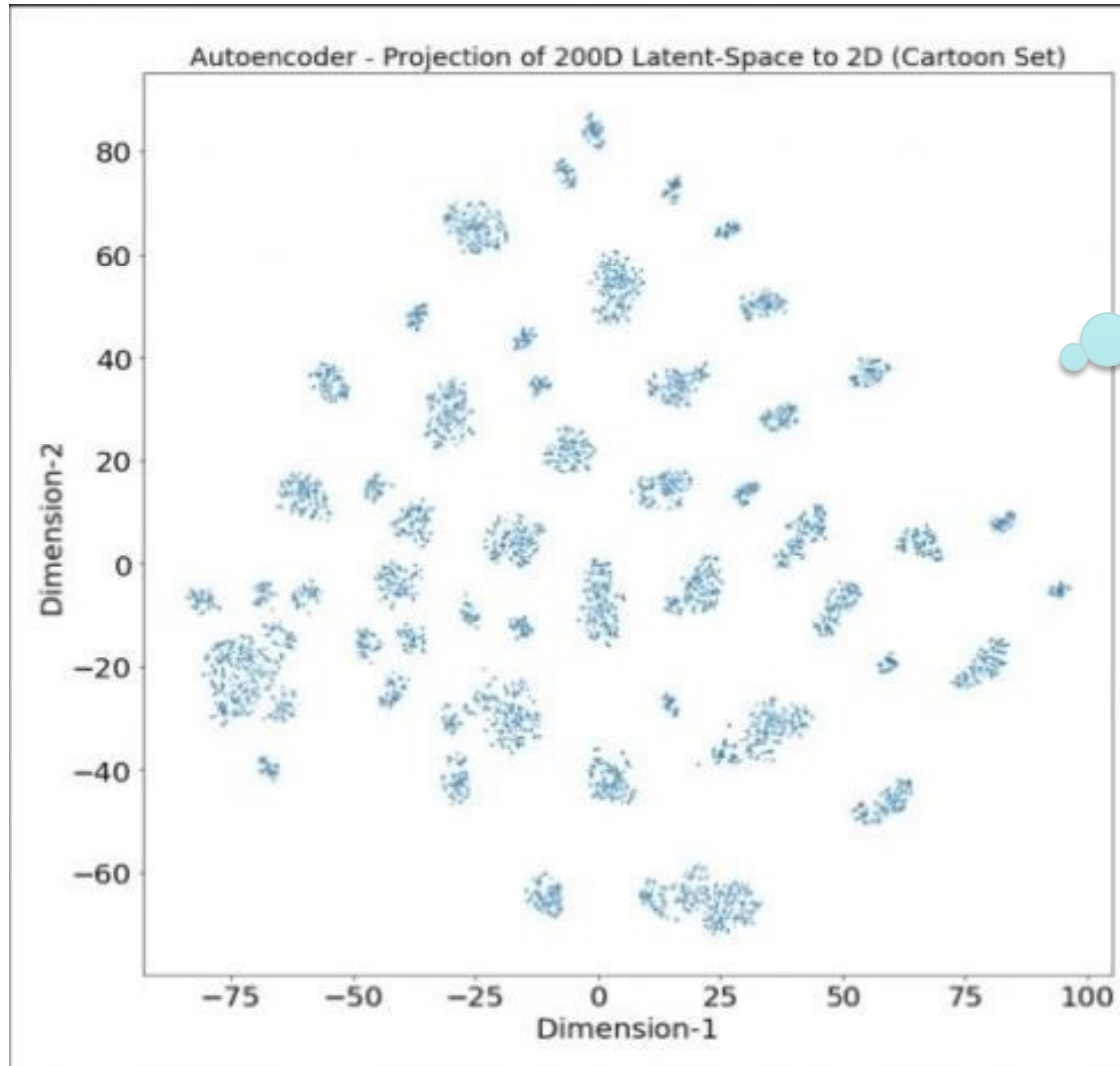
## Decoder

56	73	67	23	81	19	...
----	----	----	----	----	----	-----

discrete codes  
to image



# Related Work - Autoencoder problem



Continuous latent space, but...

# Related Work - Variational Autoencoder

---

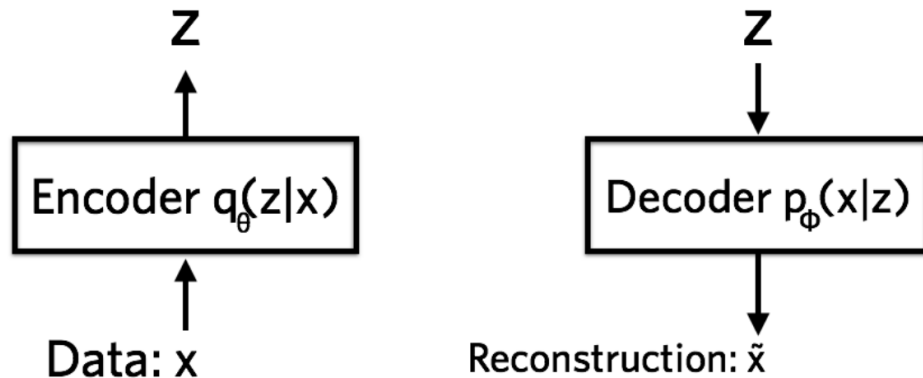
- Consider our latent space  $z$  as a random variable
- First let's enforce a **prior**  $p(z)$  on our latents, in most VAEs this is typically just a standard gaussian distribution  $\mathcal{N}(0, 1)$
- Given a raw datapoint  $x$ , we also define a **posterior** for the latent space as  $p(z | x)$
- The goal is to compute this posterior for the data, which we could express using Bayes' rule as

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

- But...  $p(x)$  is intractable, so this does not work directly
- Need approximation

# Related Work - Variational Autoencoder

- Restrict approximation of the posterior to a specific family of distributions: independent gaussians. Call this approximated distribution  $q(z|x)$



VAE: Add a prior to the autoencoder latent space: Approximate  $p(z)$  with  $q(z|x)$

Derive loss function:  $-E_{z \sim q(z|x)}[\log(p(x|z))] + KL(q(z|x) || p(z))$

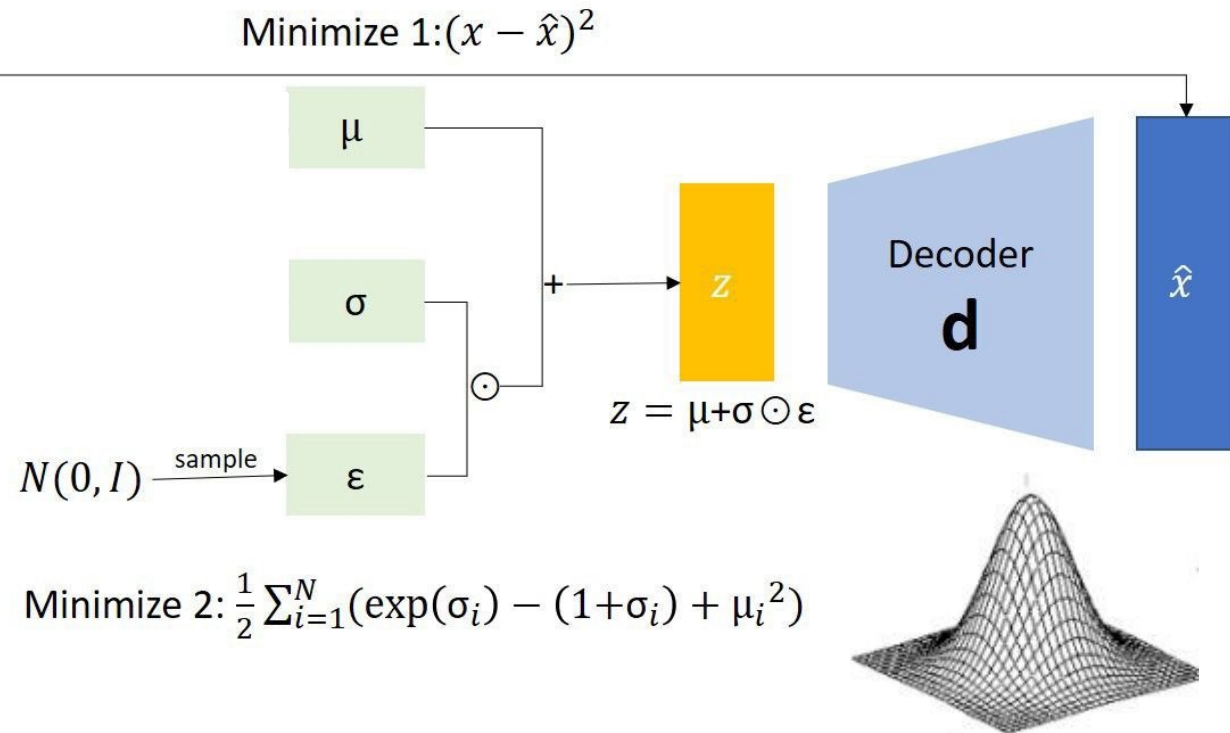
ELBO: See course on Probabilistic and Differential Programming

# KL-Divergence

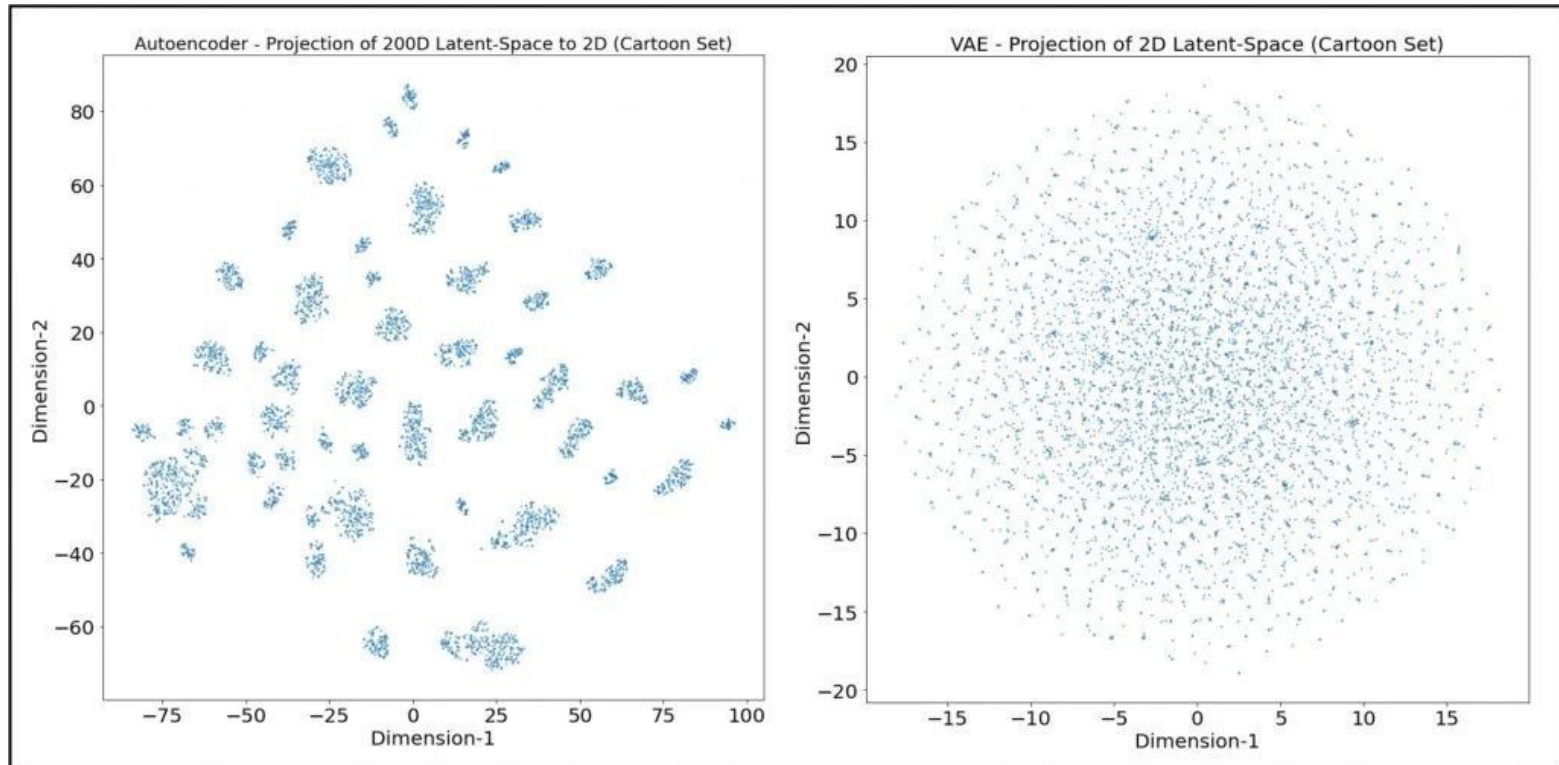
---

- From an information theory perspective,
- ... the Kullback-Leibler divergence indicates how much space per character is wasted on average
- ... when a character based on  $Q$ -based coding is applied to an information source
- ... that follows the actual distribution  $P$ .

# Variational Autoencoder as a Generator



# Related Work - Autoencoder vs. VAE

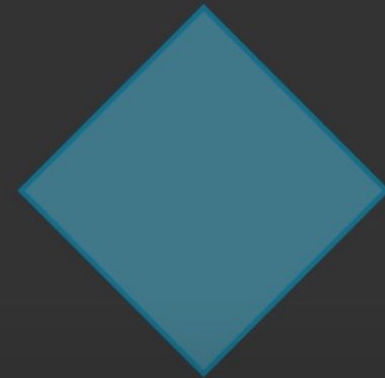


# Variational Autoencoder as a Generator

Latent space distribution  
after training



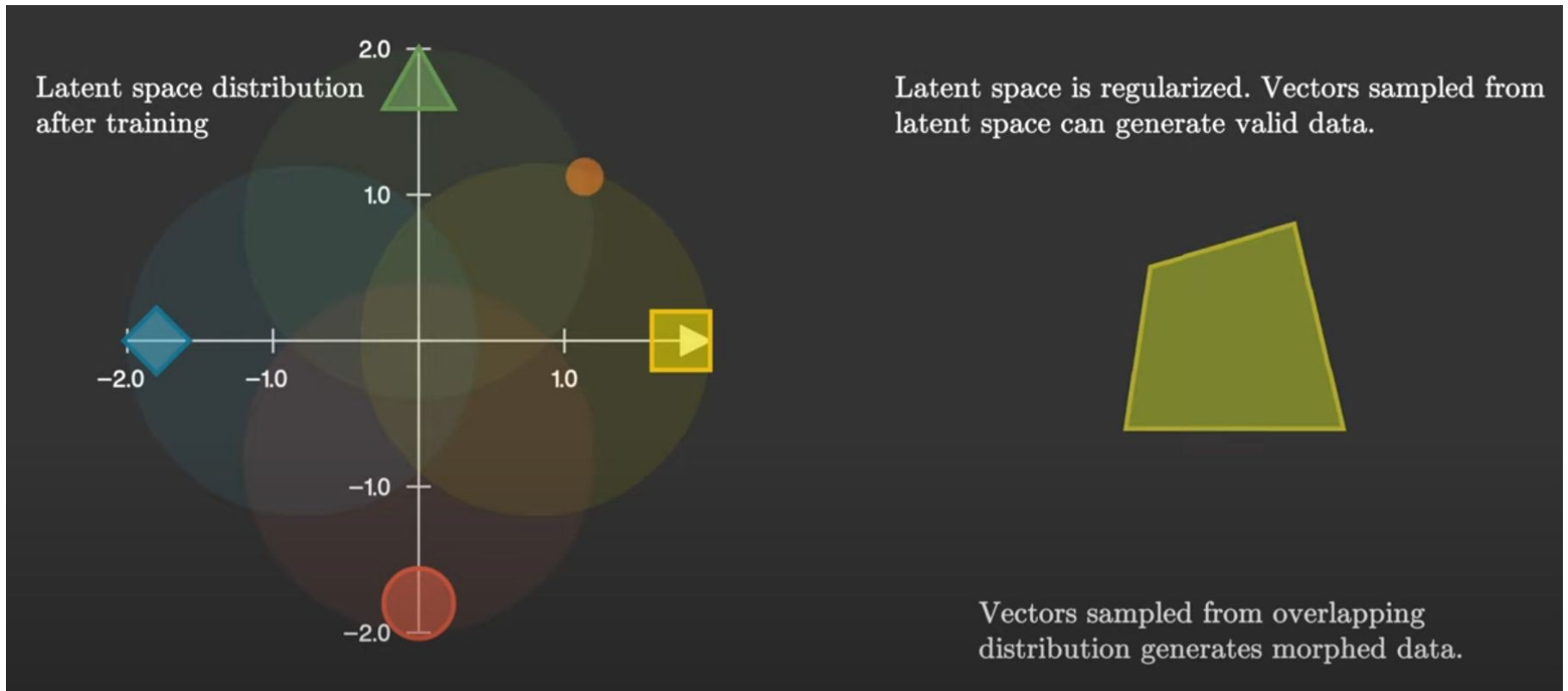
Latent space is regularized. Vectors sampled from latent space can generate valid data.



Vectors sampled from overlapping distribution generates morphed data.

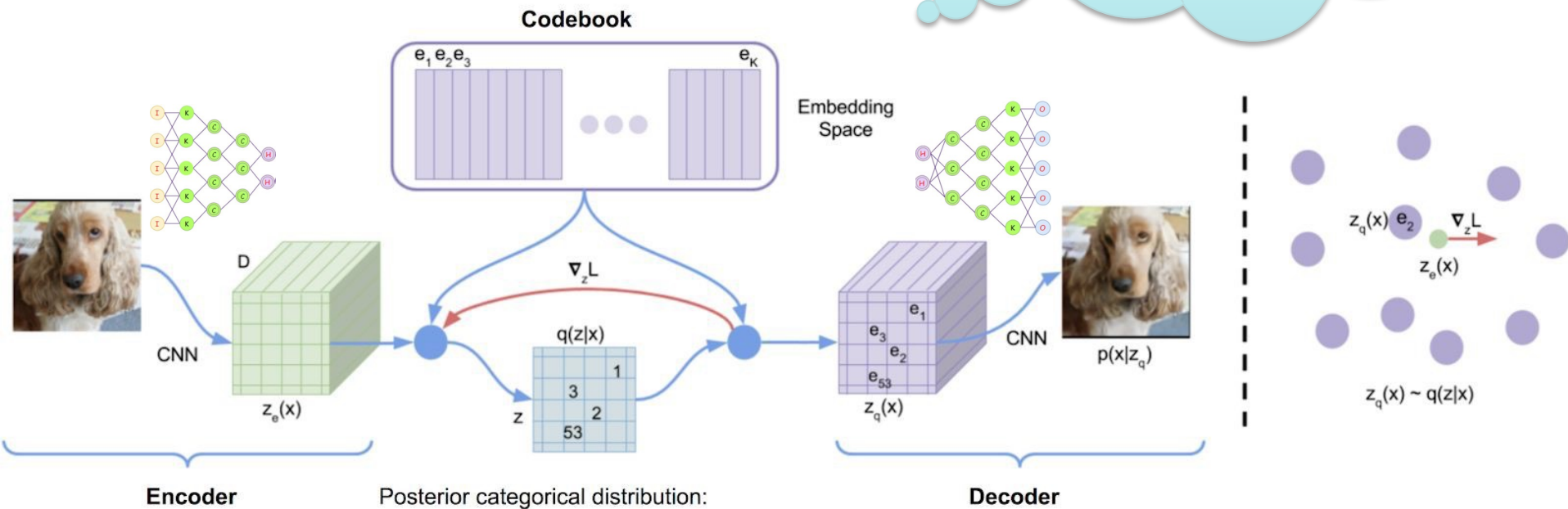


# Variational Autoencoder as a Generator

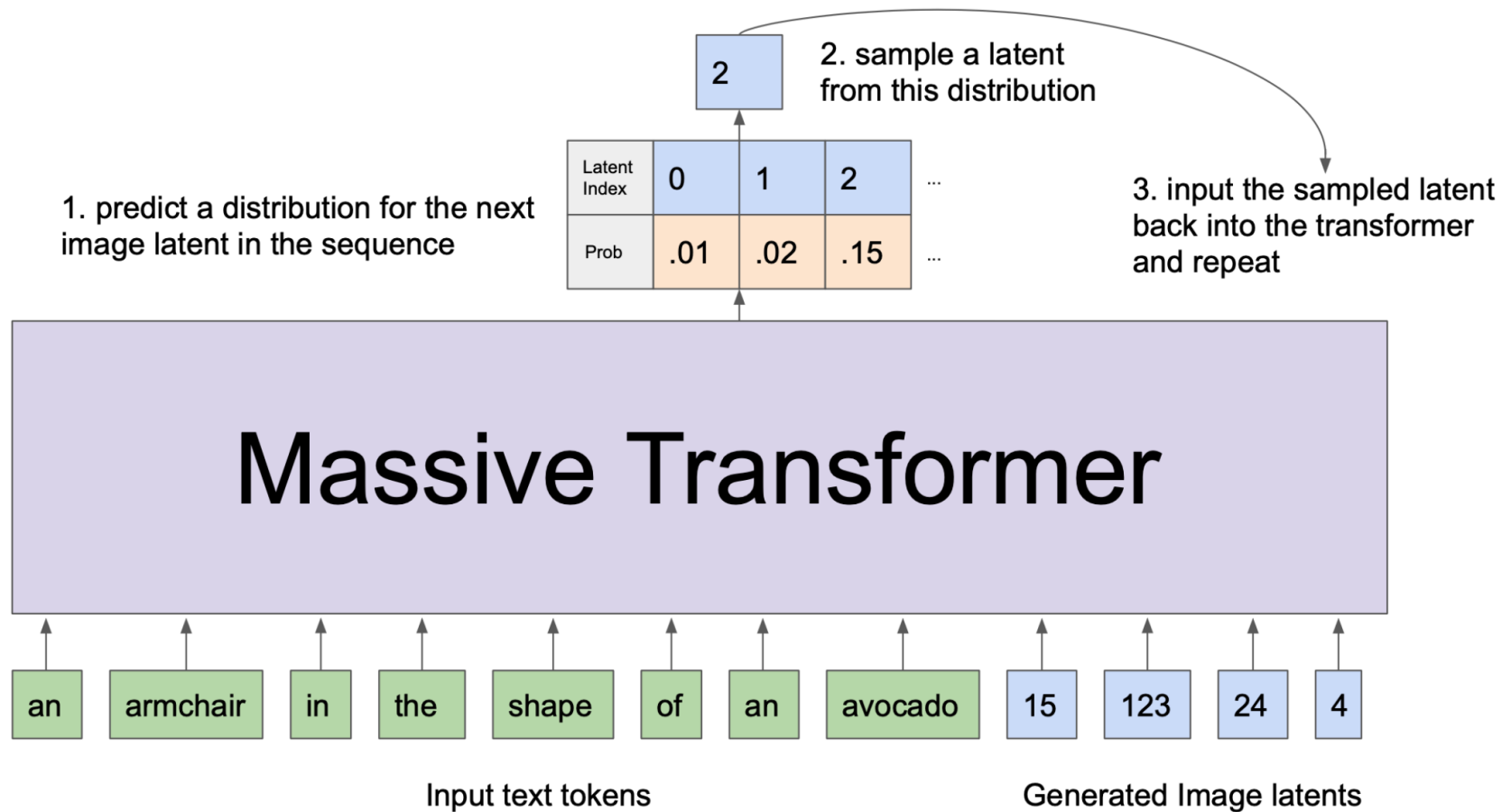


# Related Work - VQ-VAE

Want discrete latent space?  
Vector Quantized VAE

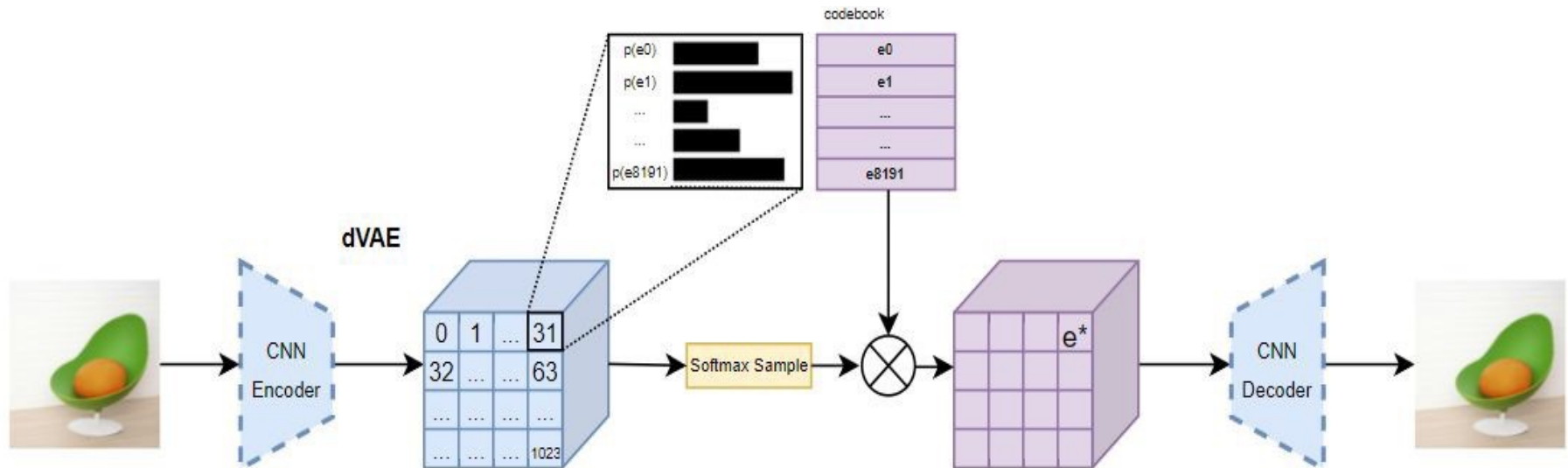


# DALL-E – Central Idea



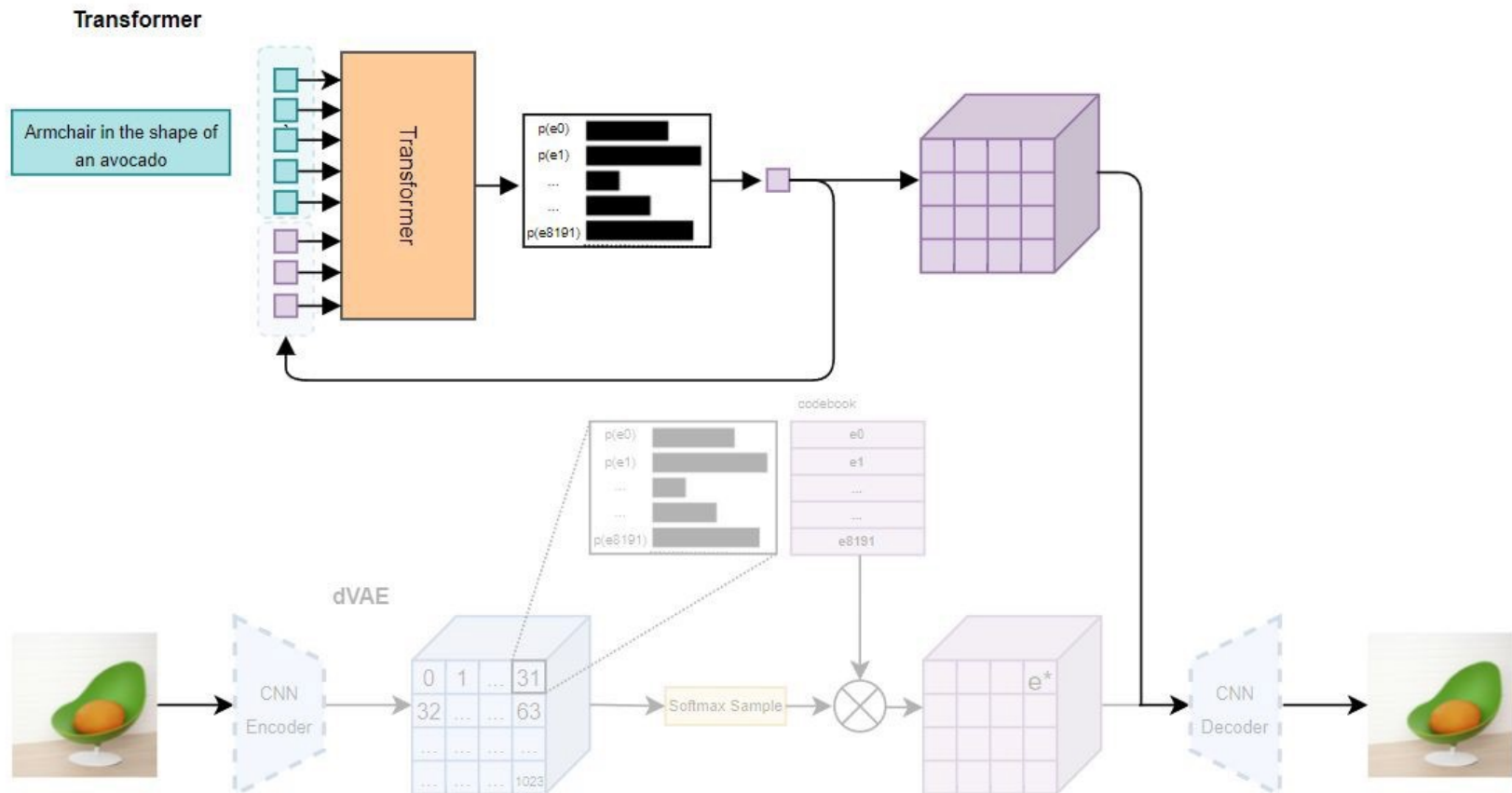
# Training

## 1. Stage: Visual Codebook



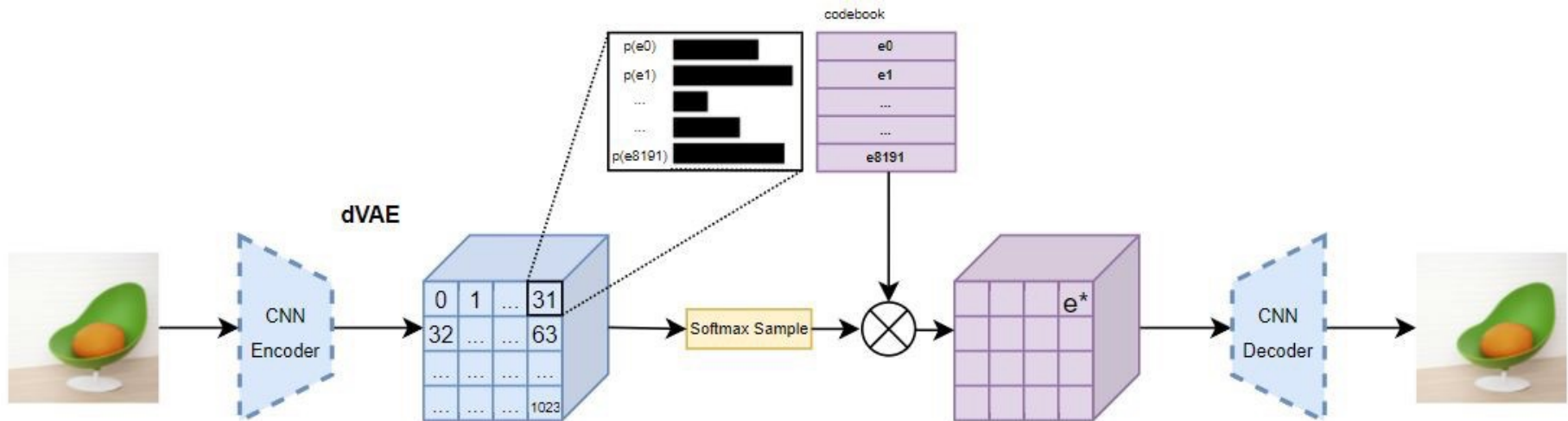
# Training

## 2. Stage: Learning of autoregressive generation of image codes



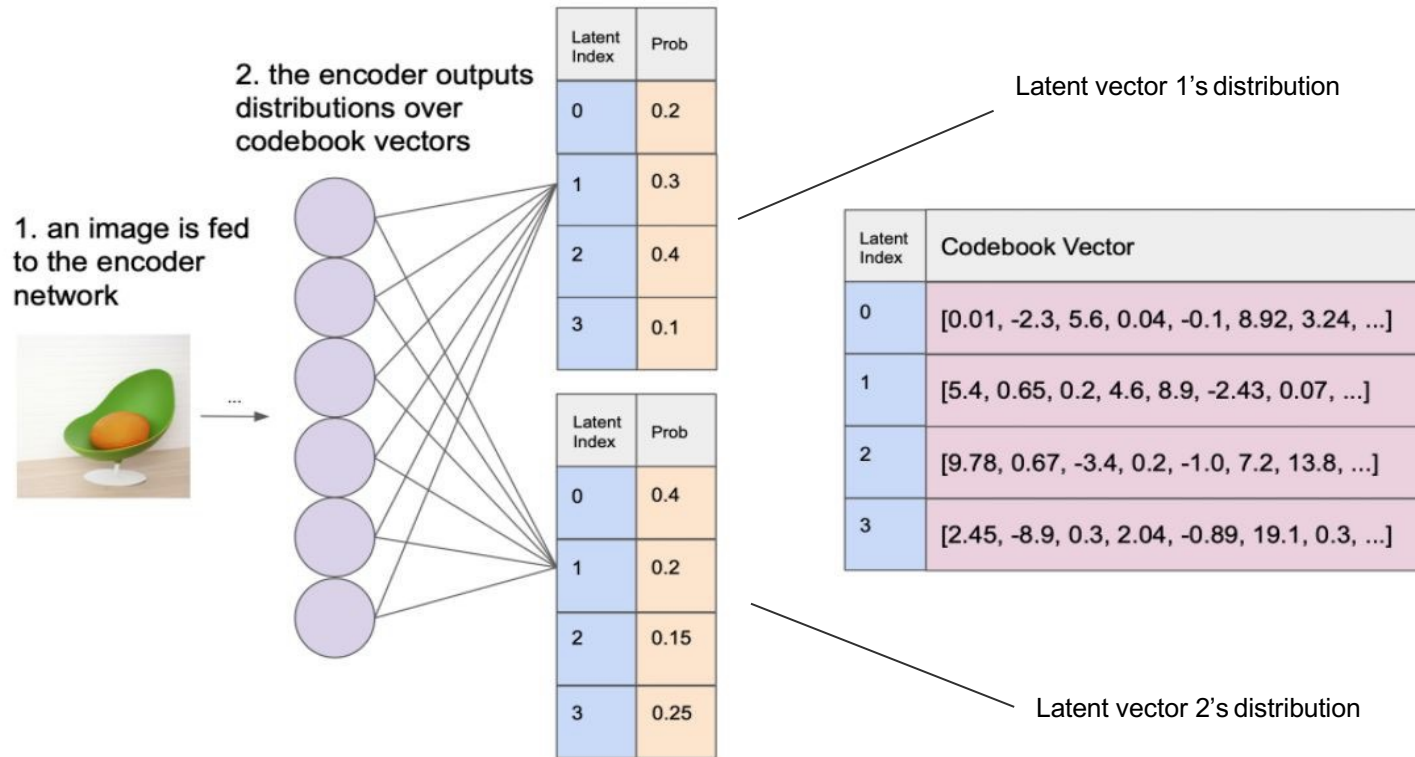
# Stage One: Learning the Visual Codebook

- Discrete Variational Autoencoder (dVAE)
  - Similar to VQ-VAE but uses distribution instead of nearest neighbor



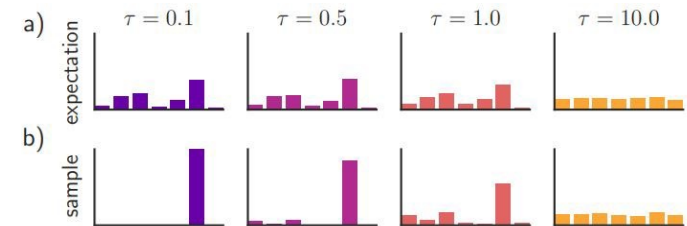
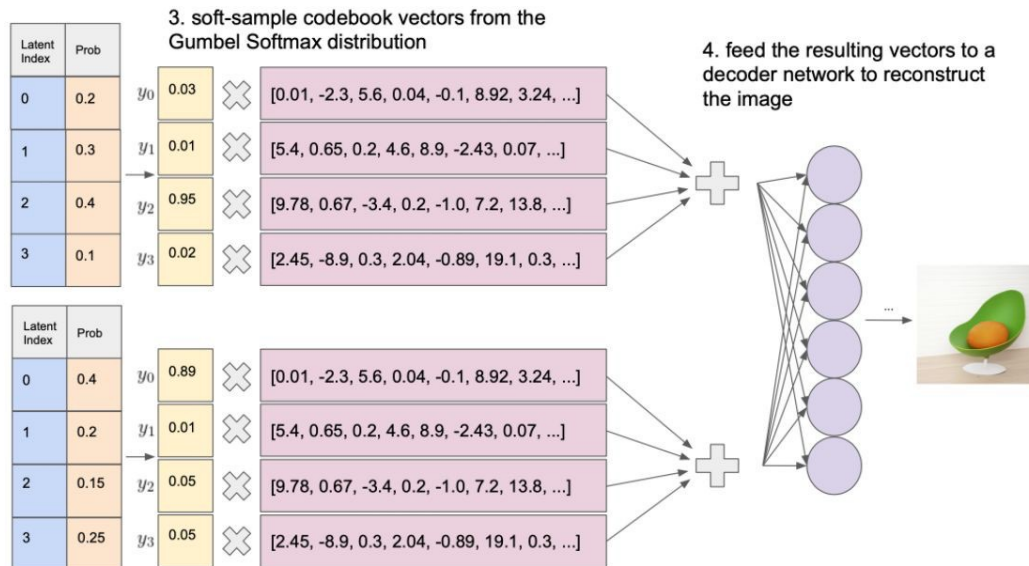
# Stage One: Learning the Visual Codebook

- Discrete Variational Autoencoder (dVAE) encoder



# Stage One: Learning the Visual Codebook

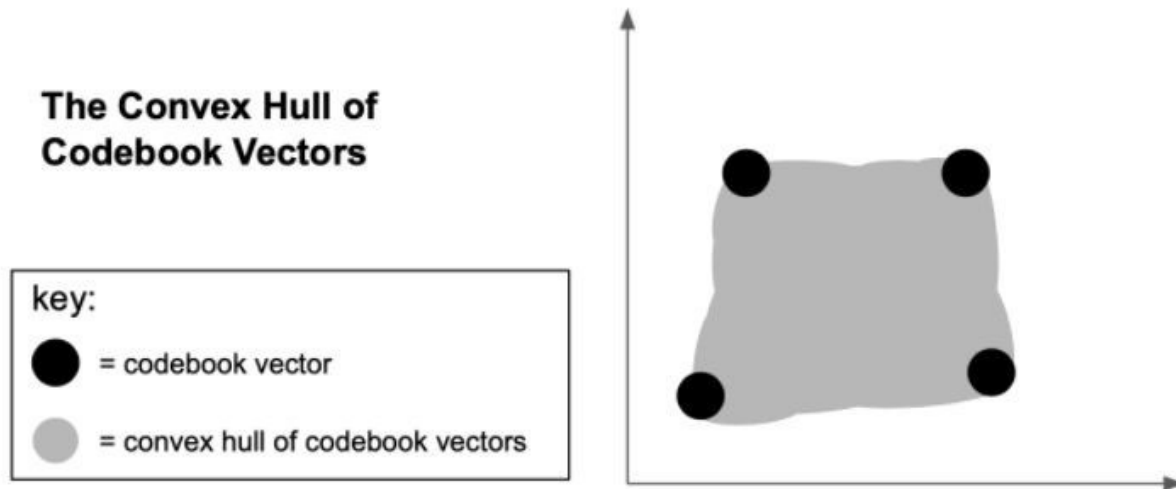
- Discrete Variational Autoencoder (dVAE) decoder
  - Gumbel softmax distribution becomes categorical over training schedule





# Stage One: Learning the Visual Codebook

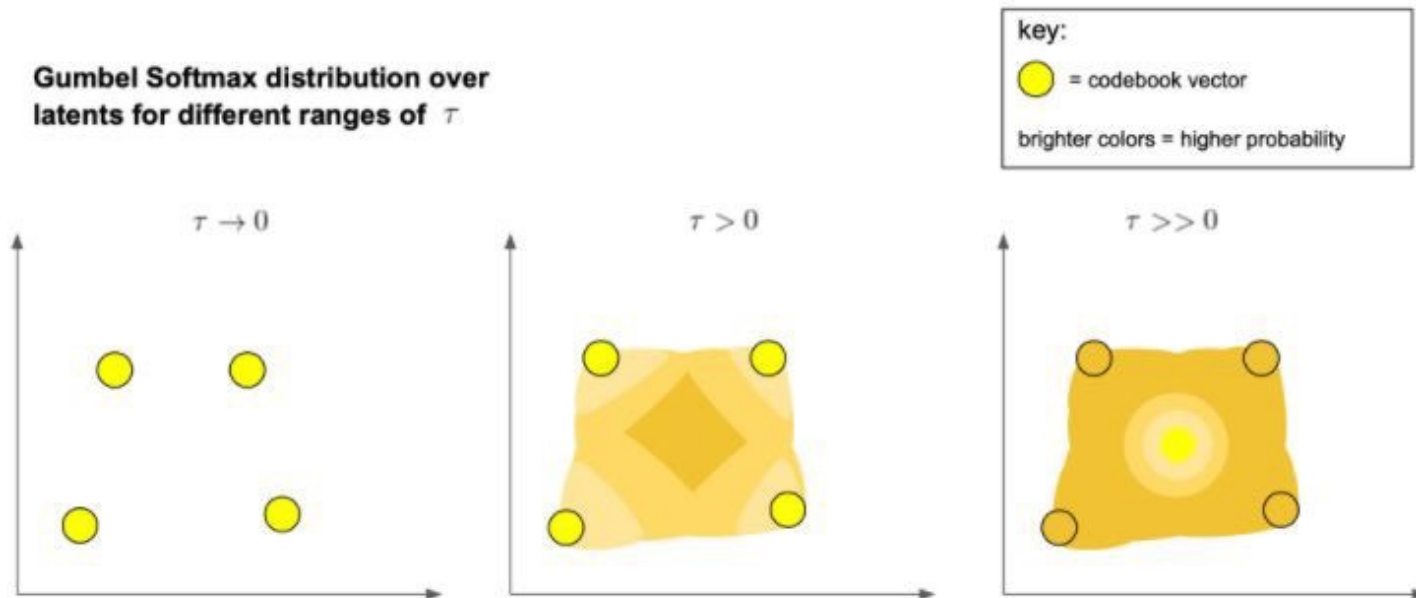
- Discrete Variational Autoencoder (dVAE) encoder
  - Issue: Can't differentiate backprop through category distribution of the bottleneck
  - Solution: Relax the bottleneck to include vectors from convex hull of set of codebook vectors



# Stage One: Learning the Visual Codebook

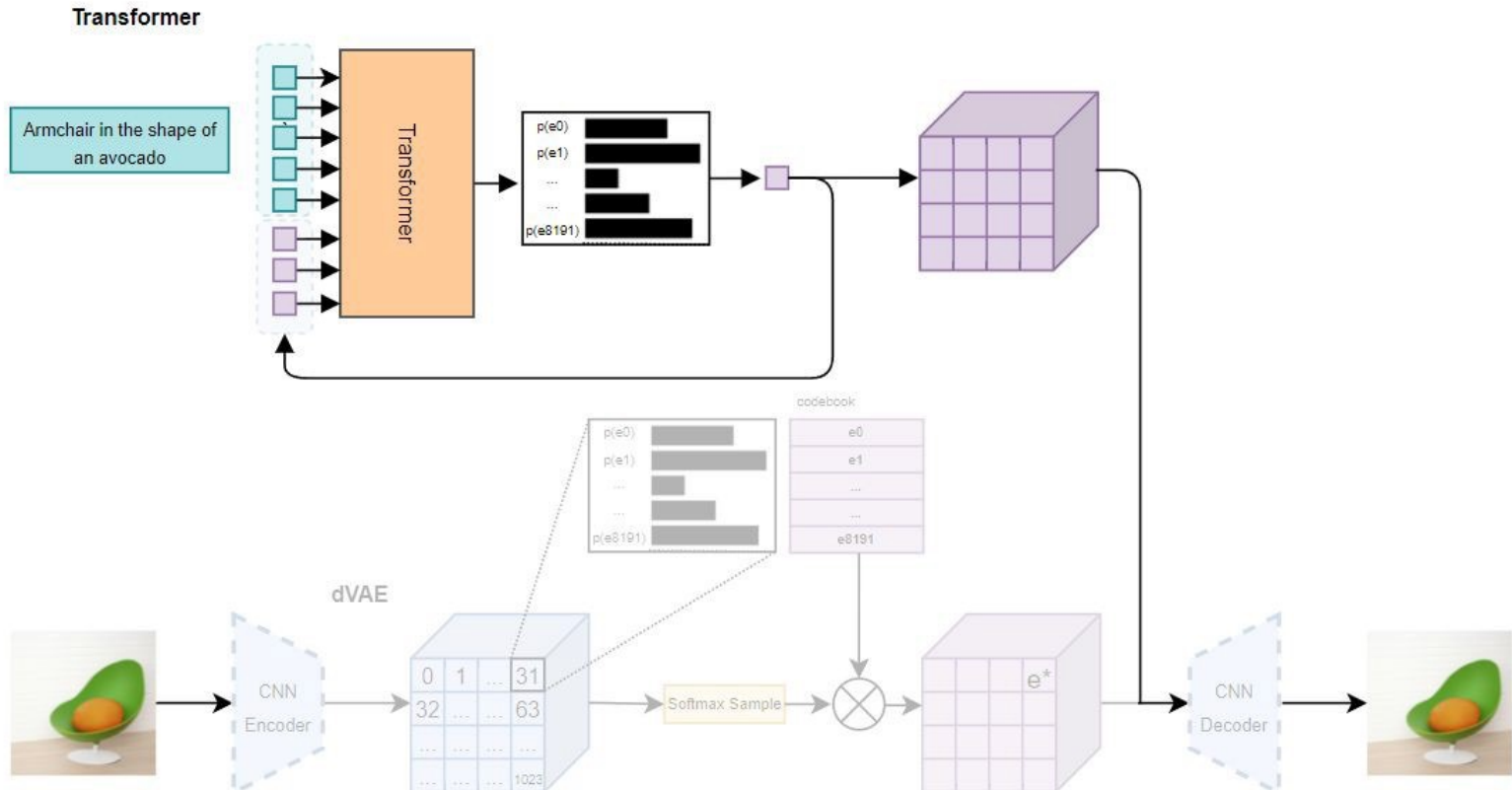
- Gumbel Softmax Relaxation
  - Sample:  $z = \text{codebook}[\text{argmax}_i [g_i + \log(q(e_i|x))]]$ 
    - Gives weights  $y_i$
    - Sampled latent vector is the sum of the weighted codebook vectors
  - Differentiable
  - Relaxation temperature annealing schedule for hyperparameter  $\tau$

**Gumbel Softmax distribution over latents for different ranges of  $\tau$**



# Stage Two: Learning Prior Distribution

- Transformer
  - Predict distribution for next token
  - Sample distribution and repeat until 1024 image tokens

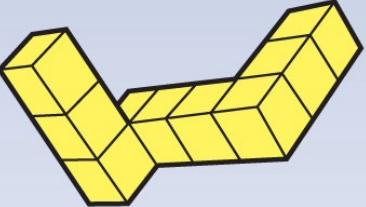
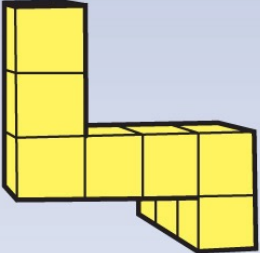
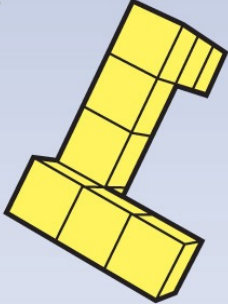
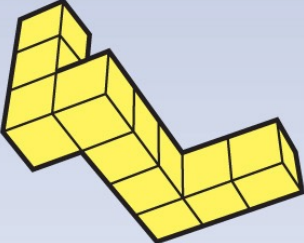
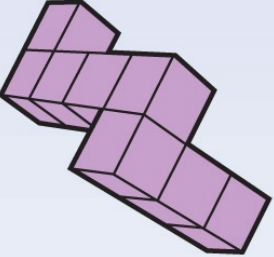
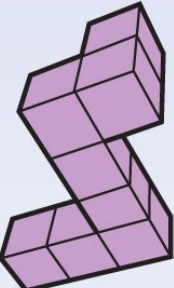
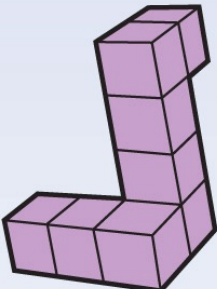
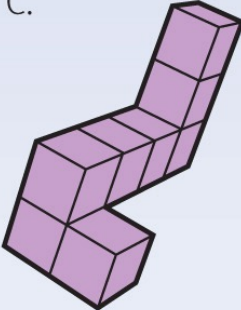


# Nice Applications. But...

---

- How can agents use text-descriptive image generation?
- Agent could generate “internal images” and interpret them to optimally carry out tasks
  - Planning might get easier with “mental imagery”
- Mental imagery (aka visual imagery) has a long tradition in cognitive psychology
- → Imagery Debate
  - Propositional or visual/“perceptive” reasoning
  - “Pylyshyn vs. Kosslyn”

# Mental Rotation

Standard	Comparison shapes		
1. 	A. 	B. 	C. 
2. 	A. 	B. 	C. 

# Mental Scanning



Island stimulus for mental scanning used in (Kosslyn, Ball, & Reiser, 1978). The island contains different locations that differ in their distance to each other. In the lower left corner a hut, a well, a lake, and a tree are visible. On the top is a rock and further locations include grass and a beach.

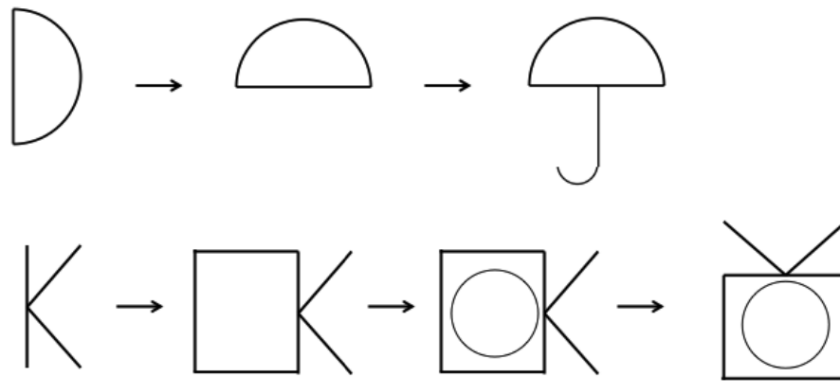
Kosslyn, S. M., Ball, T. M., & Reiser, B. J., Visual images preserve metric spatial information: evidence from studies of image scanning. *Journal of experimental psychology: Human Perception and Performance*, 4(1), 47, 1978.

# Mental Scanning

---

- Using their mental image, participants are asked to shift their attention from one entity in the image to another entity.
- It turned out that participants take significantly longer for attention shift between, for example, the hut and the rock, ...
- ... than they do for a shift between the hut and the well
  
- Strong linear correlation between the time it takes to scan between two entities in the mental image and the distance between these two entities in the original stimulus

# Mental Reinterpretation



(a)



(b)

(a) The first figure in each line is described to the participants verbally who then mentally transform their mental images according to verbal instructions so that the depicted intermediate figures should result.

(b) The respectively left one is briefly shown to the participants who then have to find an alternative interpretation of just the right side of the stimulus using their mental image.




# Dual coding theory

---


- Human cognition divided into two processing systems: visual and verbal.
  - The visual system deals with graphical information processing and the verbal system deals with linguistic processing
  - These two systems are separate and are activated by different information
- GenAI: Text → Image/Video
- GenAI: Image/Video → Caption as text
- Use CLIP principles for mental images?
- YOLO for mental videos?

Sadoski, Mark; Paivio, Allan, "A Dual Coding Theoretical Model of Reading", *Theoretical Models and Processes of Reading*, DE: International Reading Association, *Cognitive psychology*. Cengage Learning. pp. 1329–1362, 2016.

# Counterfactuals and Causality



**Associations**  
Seeing  
What if I see ... ?



**Interventions**  
Doing  
What would I do ... ?  
How?

