

Multimodal AI

Lecture 6.1 – Multimodal Generation

Paul Liang

Assistant Professor

MIT Media Lab & MIT EECS



<https://pliang279.github.io>

ppliang@mit.edu

 [@pliang279](https://twitter.com/pliang279)



Assignments for This Coming Week

HW3 released, due 3/17

Midterm exam next Thursday 3/19. Midterm review next Tuesday 3/17.

Project mentors released. Try to meet as often as you can. Meet with me today.

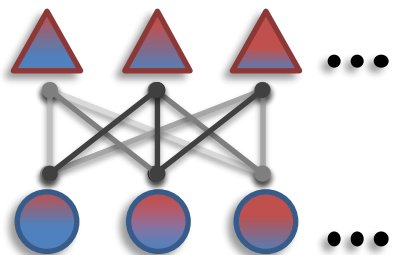
Today's lecture

- 1 Text-to-image generation
- 2 Introduction to generative models
- 3 GMMs and VAEs

Lecture outline

Part 1: Multimodal foundation model representations of text, video, audio

*It's just a privilege to
watch your mind at work.*

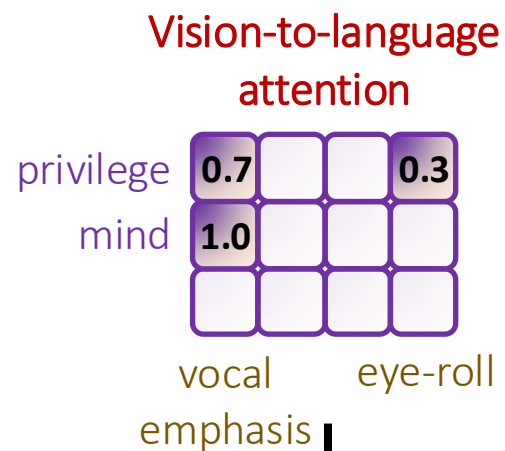
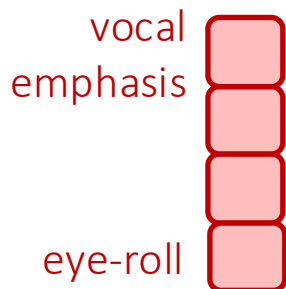
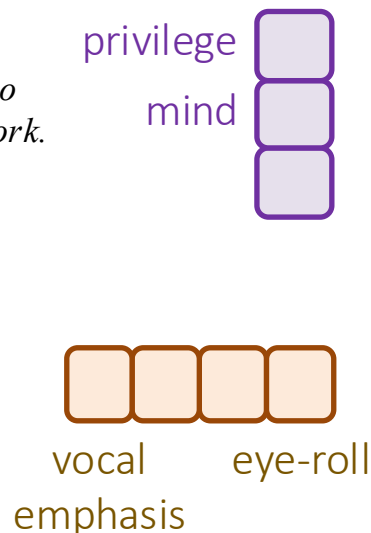


Multimodal
representation



Multimodal Transformers

It's just a privilege to watch your mind at work.



privilege
mind

New **language** representation
contextualized with vision

(row) normalize to 0-1

$$h = \text{softmax} \left(\frac{X_1 W_q W_k^T X_2^T}{\sqrt{d}} \right) X_2 W_v$$

3 x d d x 4 4 x d

normalize wrt dimension d

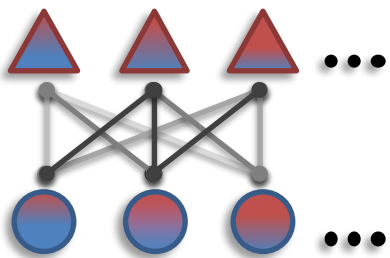
3 x 4
(weighted) outer product

Sarcasm

Lecture outline

Part 2: Adapting large language models for multimodal text generation

*It's just a privilege to
watch your mind at work.*



Multimodal
representation



*This person is being sarcastic.
They seem to be close friends.*

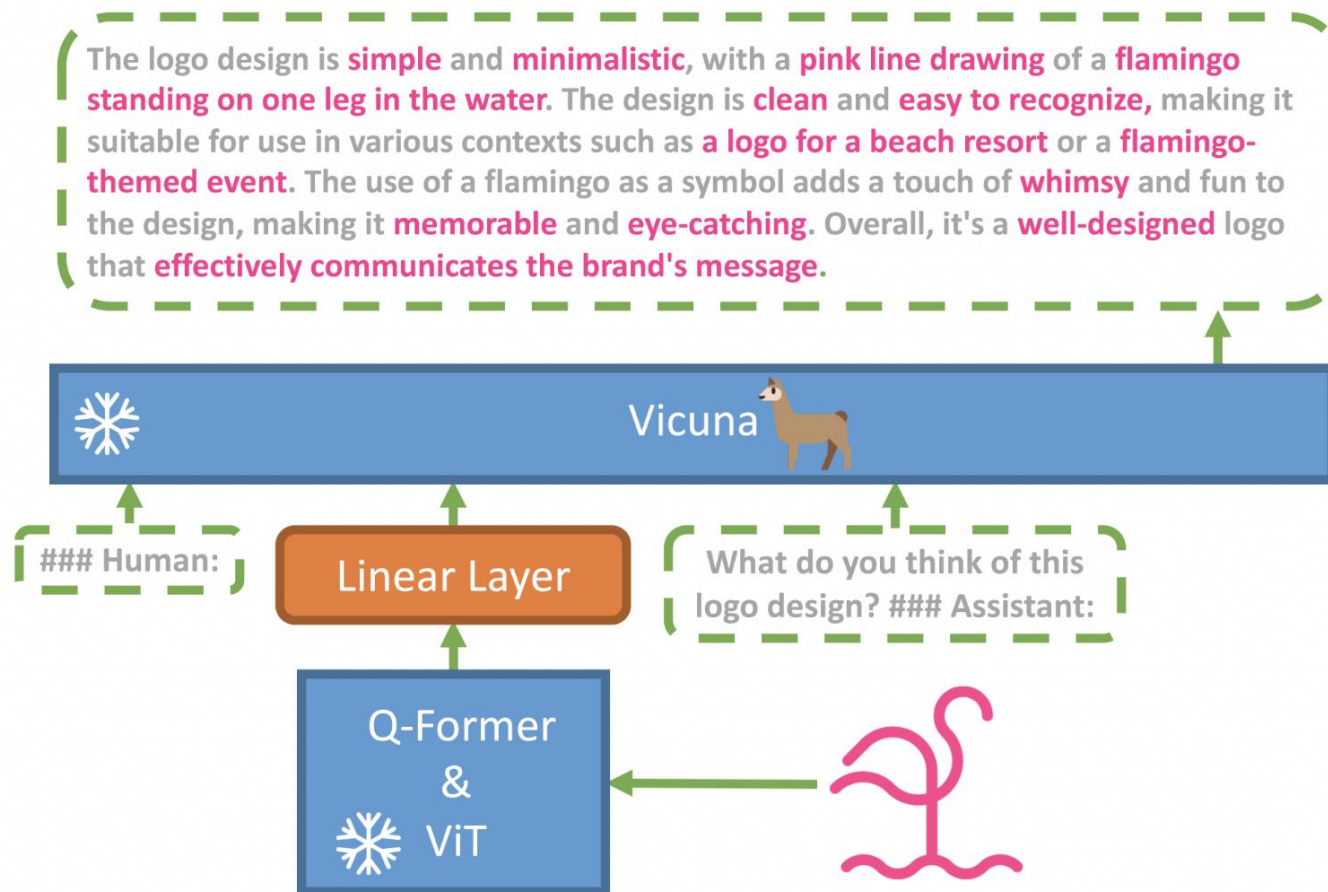


Adapting Large Language Models

Mini-GPT4

Stage 1: **Alignment** using paired image-text data.

Stage 2: **Instruction tuning** using image + text instructions and example completions.

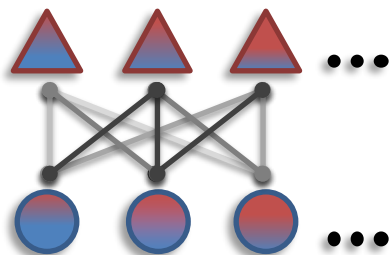


The architecture of MiniGPT-4.

Lecture outline

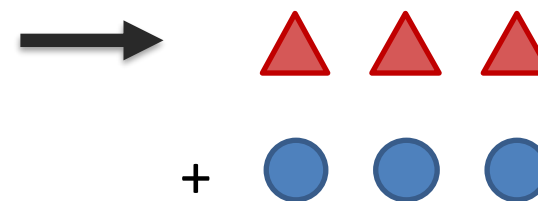
Part 3: Enabling text and image generation

*It's just a privilege to
watch your mind at work.*



Multimodal
representation

*This person is being sarcastic.
They seem to be close friends.*



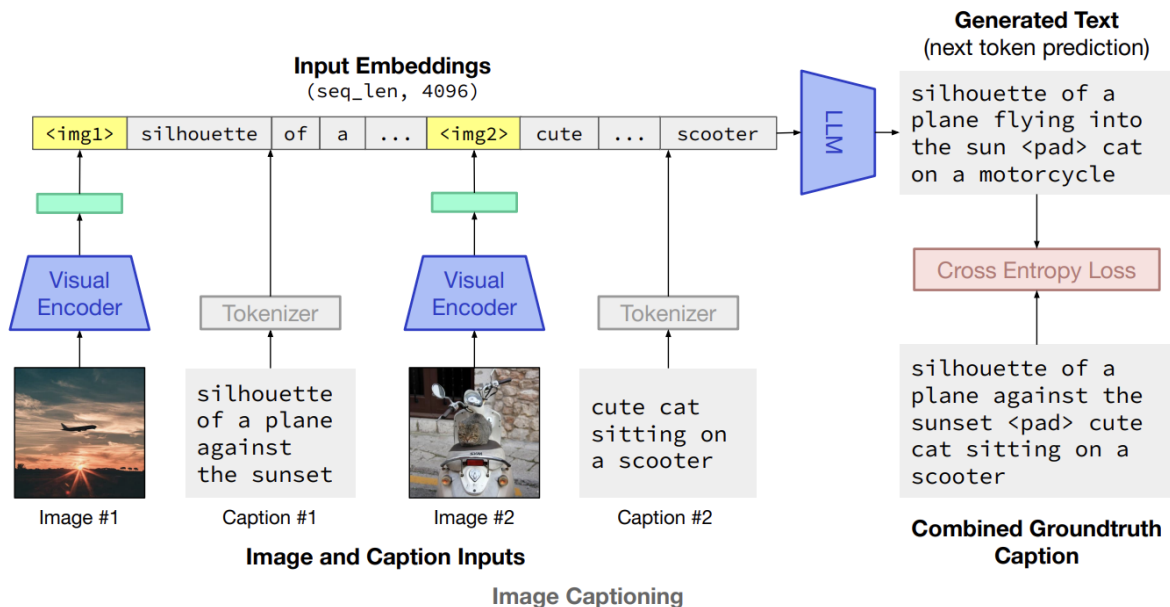
*(quote previous episodes)
(highlight multimodal information)*

1. Text-to-image retrieval
2. Text-to-image generation

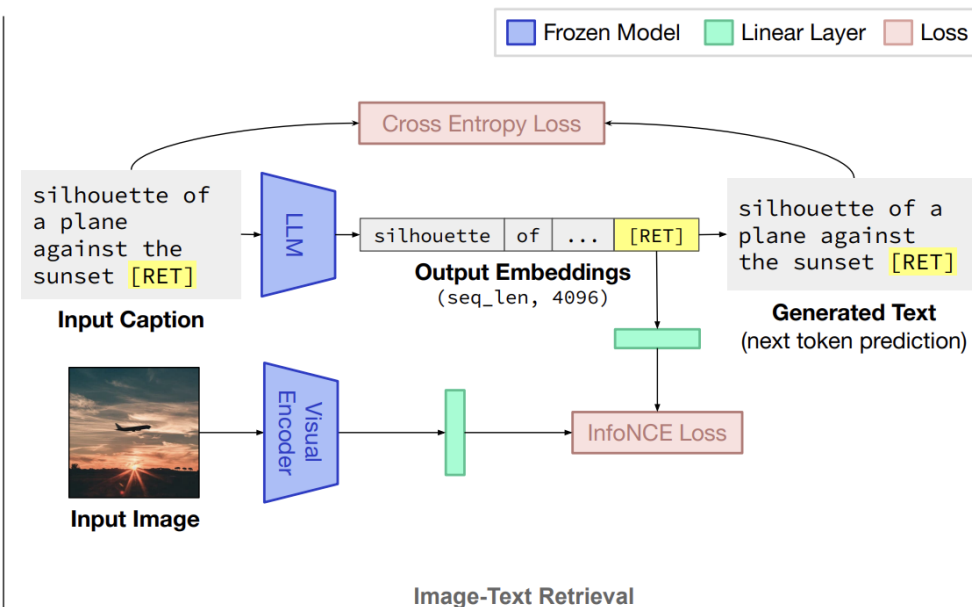
Grounding LMs for Image Retrieval

LIMBeR + CLIP. Trainable in 1 day on 1 GPU


Interleaved images and text





CLIP, with a frozen LLM





Grounding LMs for Image Retrieval


I am thinking of an animal. It is brown and furry, and has a tail. What is it? 


A squirrel. 


No, the animal I am thinking of lives in the river. 


A beaver. 


That's right, what do beavers look like? 


They have a tail. 


Show me an illustration of a beaver 



A beaver. 





Yes, what about a pencil drawing of one? 





What about a photo of one? 






What is this? 


This is a house with a unique design. 


What is its shape? 


The house is shaped like a dome. 


Can you show me more houses like this? 


The house is designed by architect. 





That's nice, any others? 


Yes, there are many houses like this in the world. 





Provide instructions on how to cook. 


How do I make macarons? 


Macarons are a French dessert that are made by combining egg whites and sugar. The egg whites are beaten until stiff and then folded into the sugar. 


What does the stiff egg white look like? 

The egg white should be stiff and glossy. 

Show me a picture of what that looks like 



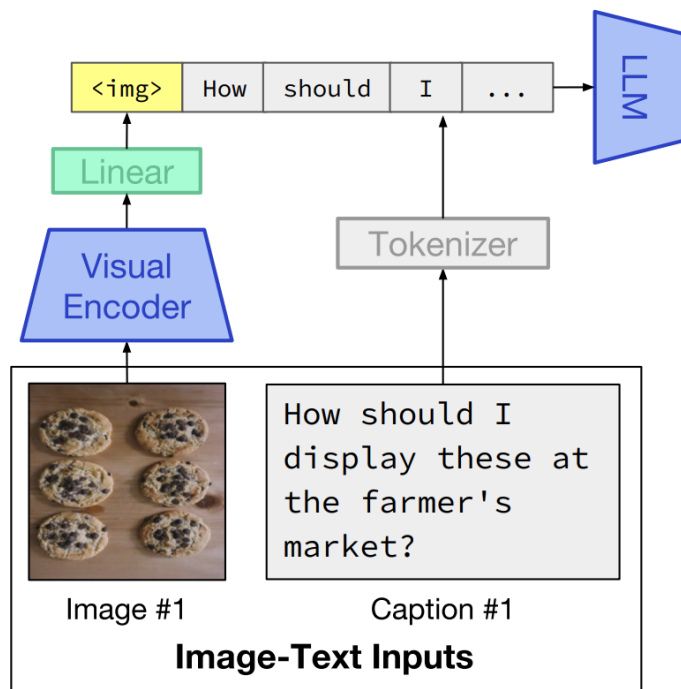
After this, what do I do? 

The macaron is then baked in a hot oven. 

Grounding LMs for Multimodal Generation

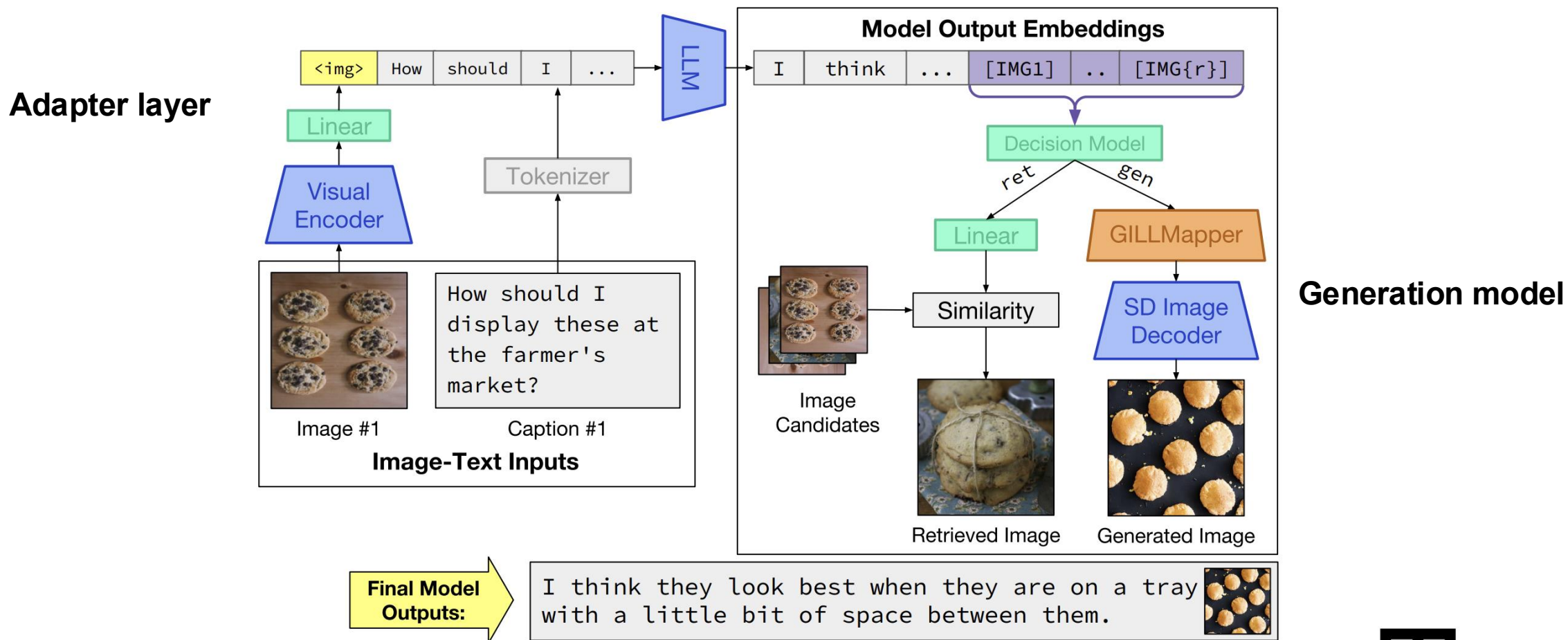
Large multimodal models with image generation

Adapter layer
(part 2)



Grounding LMs for Multimodal Generation


Large multimodal models with image generation



Grounding LMs for Multimodal Generation


Visual Storytelling

Our model can condition on interleaved image-and-text inputs to generate more relevant images compared to non-LLM based text-to-image generation models.



How can I make this more nutritious?


You can add vegetables to your ramen noodles, but you should be careful not to overdo it.




What are some vegetables I can add to it?

Broccoli, carrots, and green beans are all good choices.

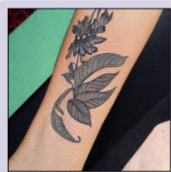
I'm looking for some ideas for a new tattoo. What do you think looks good on a petite person?




I want a tattoo of a sunflower, but I'm not sure what style would look best on me.




That's beautiful. Show me one that would work well as a complementary piece






How should I publicize these at the market?

I would suggest a simple sign with the name of the business and a picture of the cupcakes.




Hi, I'm looking for ideas for a customised cake design for my daughter's birthday.

I'm not sure what you mean by customised.




I want a cake with a unicorn theme, with pastel colors and a whimsical feel.

I think



Can you also show me some decorations that would work well for the party?

I can show you some ideas



Multimodal Dialogue

Our model can generate multimodal dialogue, weaving together text, retrieved images, and generated images.

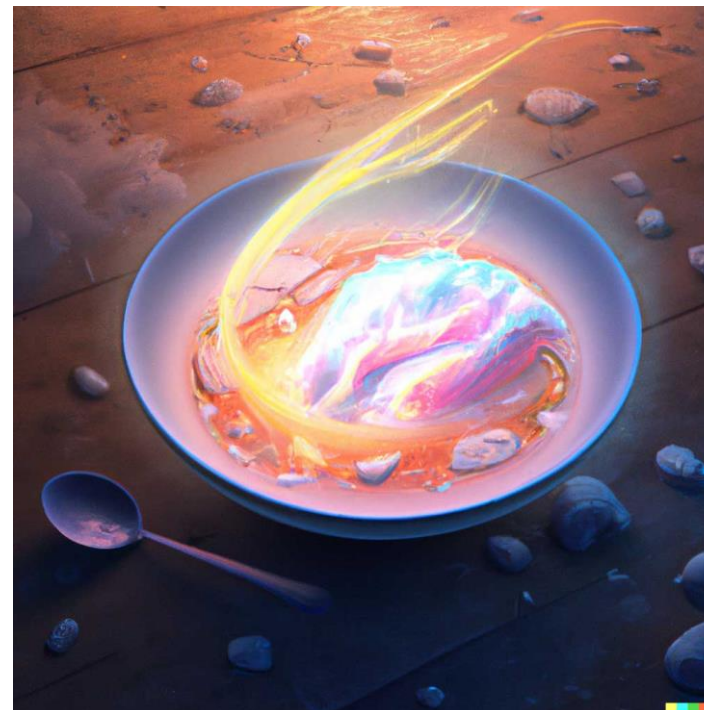
User prompts
 Retrieved
 Generated

Text to Image Generation

An astronaut riding a horse in the style of Andy Warhol.

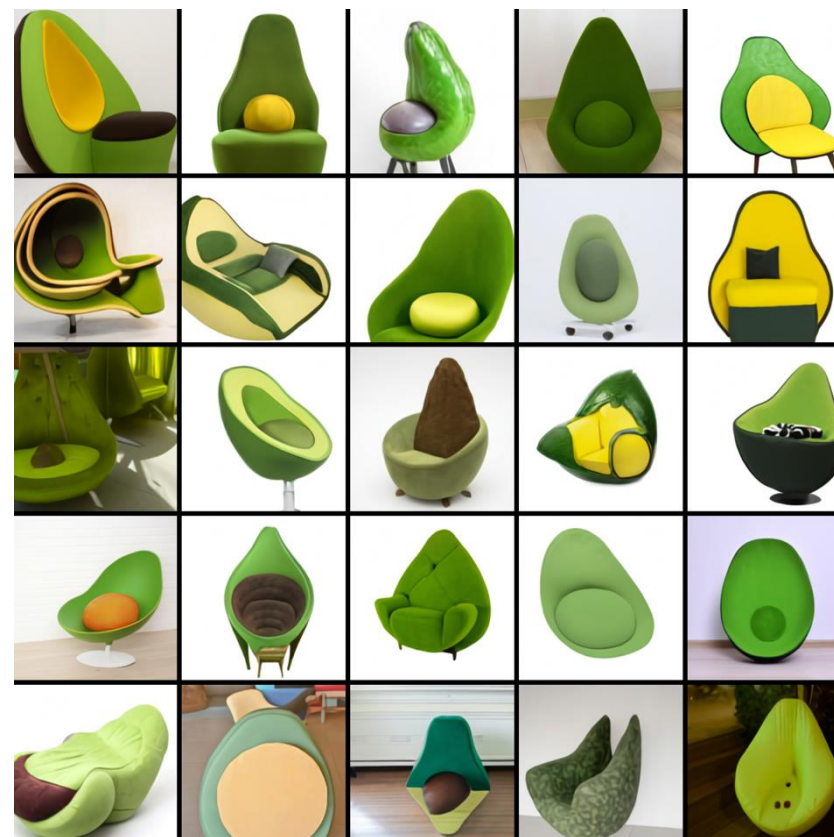


A bowl of soup that is a portal to another dimension as digital art



Text to Image Generation

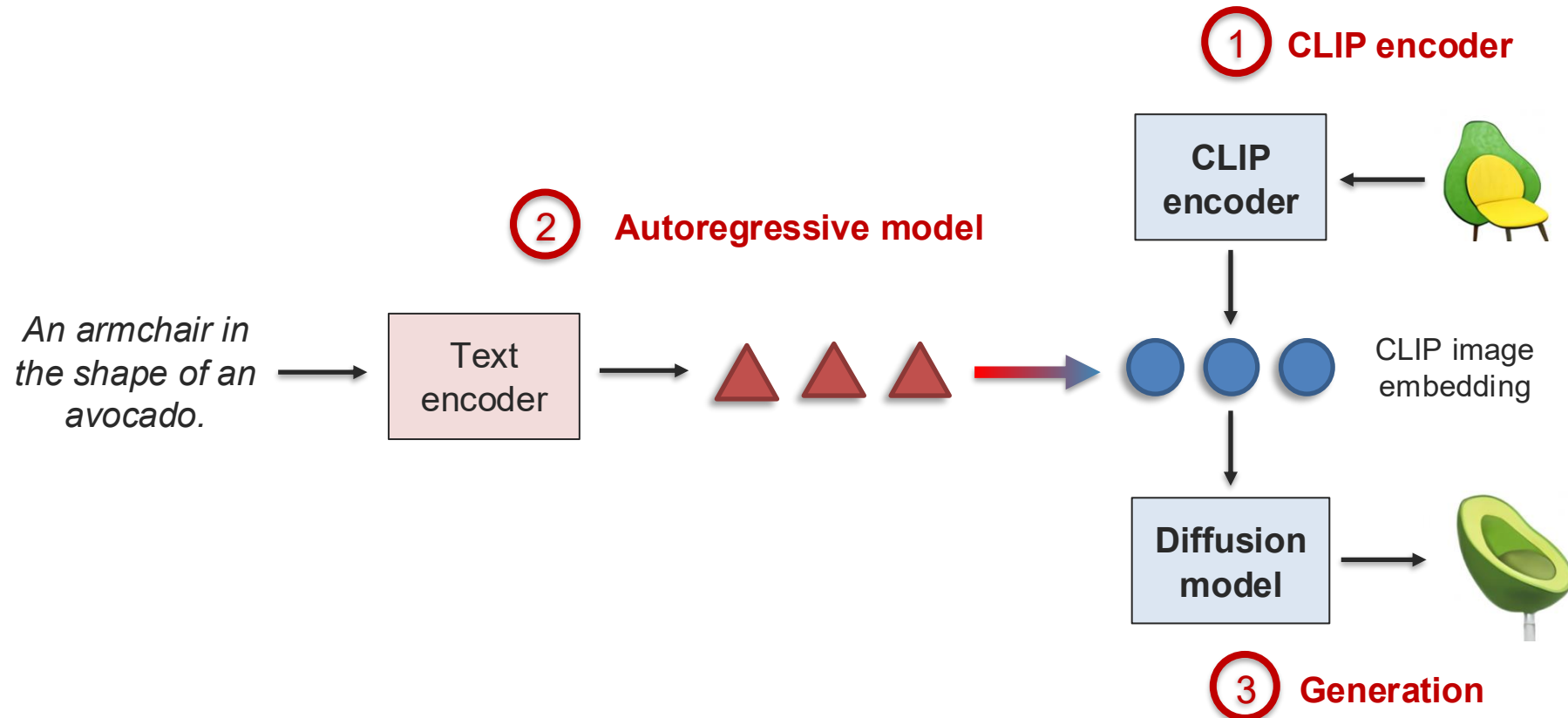
An armchair in the shape of an avocado



From Text to Multimodal Generation

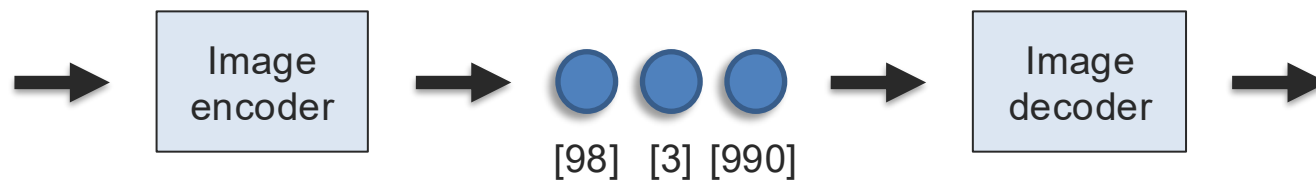
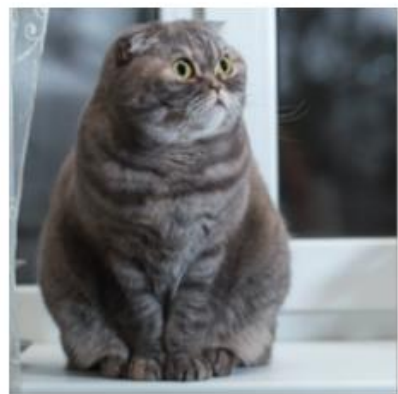
Directly training diffusion models with conditional information

Conditional latent variables are pretrained CLIP embeddings, then diffusion model to generate image.



From Text to Multimodal Generation

(1) Using a discrete variational autoencoder to learn discrete visual tokens

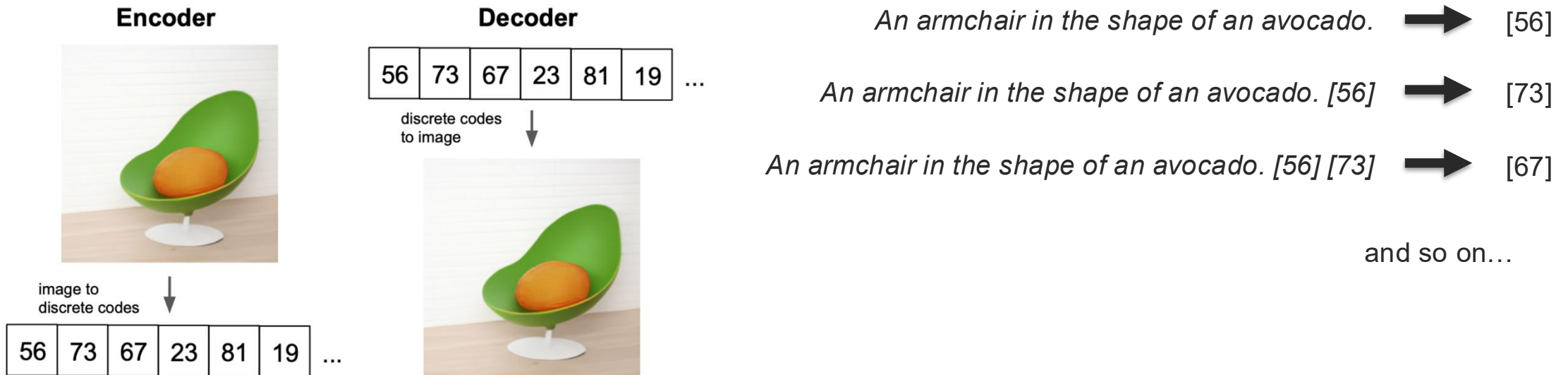


List of digits, [0... 8192]
Each digit is a “visual token”



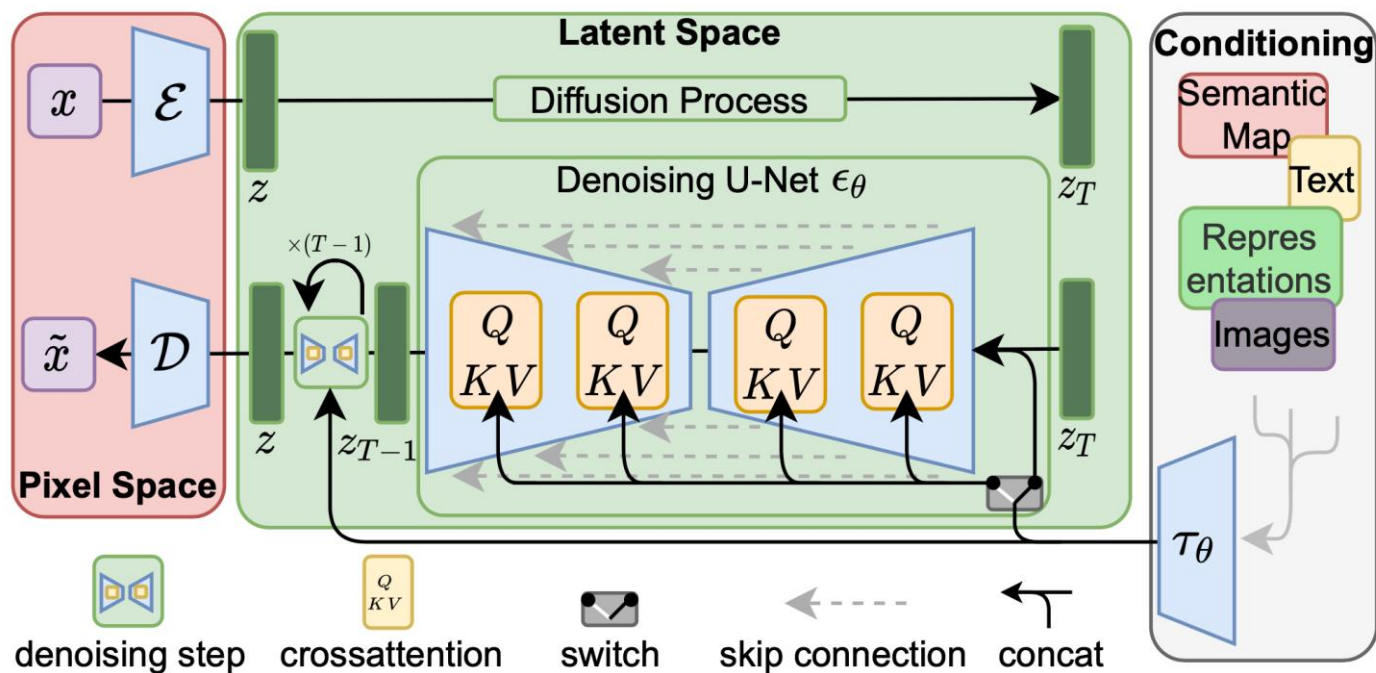
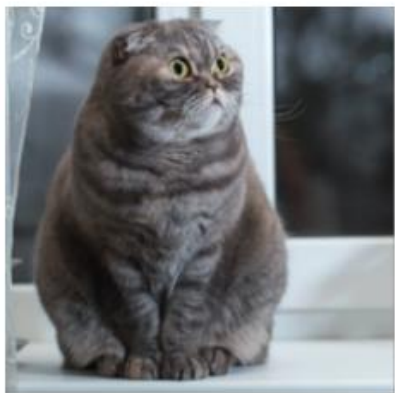
From Text to Multimodal Generation

(2) Autoregressive generation of discrete visual tokens



From Text to Multimodal Generation

Nowadays: replacing the VAE with a diffusion model



Crash Course on Modern Generative Models

Learn to model $p(x)$ where x = text, images, videos, multimodal data (by maximizing likelihood)

- Given x , **evaluate** $p(x)$ - realistic data should have high $p(x)$ and vice versa
- **Sample** new x according to $p(x)$ - sample realistic looking images
- Unsupervised **representation** learning - we should be able to learn what these images have in common, e.g., ears, tail, etc. (features)



INPUT (x)	RECONSTRUCTION (AUTR)	RECONSTRUCTION (Gen-RNN)
unable to stop herself, she briefly, gently, touched his hand.	unable to stop herself, she leaned forward, and touched his eyes.	unable to help her , and her back and her into my way.
why didn't you tell me?	why didn't you tell me?	why didn't you tell me?"
a strange glow of sunlight shines down from above, paper white and blinding, with no heat.	the light of the sun was shining through the window, illuminating the room.	a tiny light on the door, and a few inches from behind him out of the door.
he handed her the slip of paper.	he handed her a piece of paper.	he took a sip of his drink.

Crash Course on Modern Generative Models

Sometimes we also care about $p(x|c)$ - **conditional generation**

- c is a category (e.g. faces, outdoor scenes) from which we want to generate images

We might also care about $p(x_2|x_1,c)$ - **style transfer**

- c is a stylistic change e.g. negative to positive



From negative to positive

consistently slow .
 consistently good .
 consistently fast .

my goodness it was so gross .
 my husband 's steak was phenomenal .
 my goodness was so awesome .

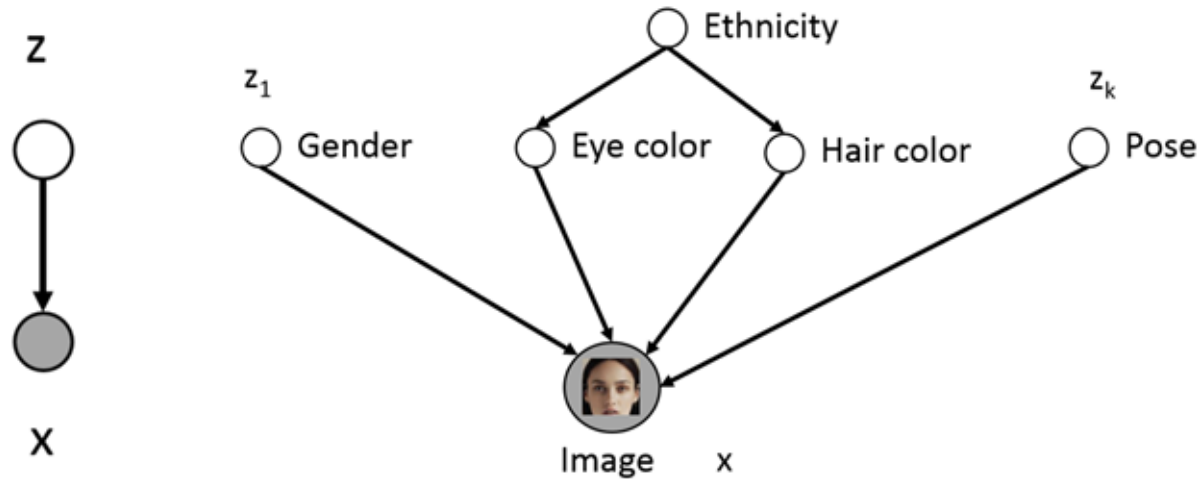
it was super dry and had a weird taste to the entire slice .
 it was a great meal and the tacos were very kind of good .
 it was super flavorful and had a nice texture of the whole side .

Latent Variable Generative Models

- Lots of variability in images \mathbf{x} due to gender, eye color, hair color, pose, etc.
- However, unless images are annotated, these factors of variation are not explicitly available (latent).
- Idea: explicitly model these factors using latent variables \mathbf{z}



Latent Variable Generative Models



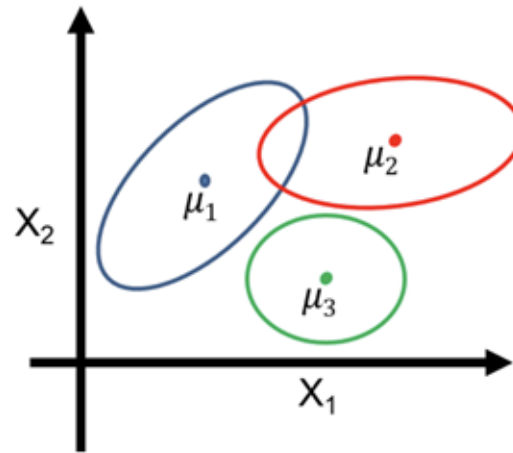
- Only shaded variables x are observed in the data
- Latent variables z are unobserved - correspond to high-level features
 - We want z to represent useful features e.g. hair color, pose, etc.
 - But very difficult to specify these conditionals by hand and they're unobserved
 - Let's **learn** them instead

Mixture of Gaussians

Mixture of Gaussians (Bayes network $z \rightarrow x$)

$$\mathbf{z} \sim \text{Categorical}(1, \dots, K)$$

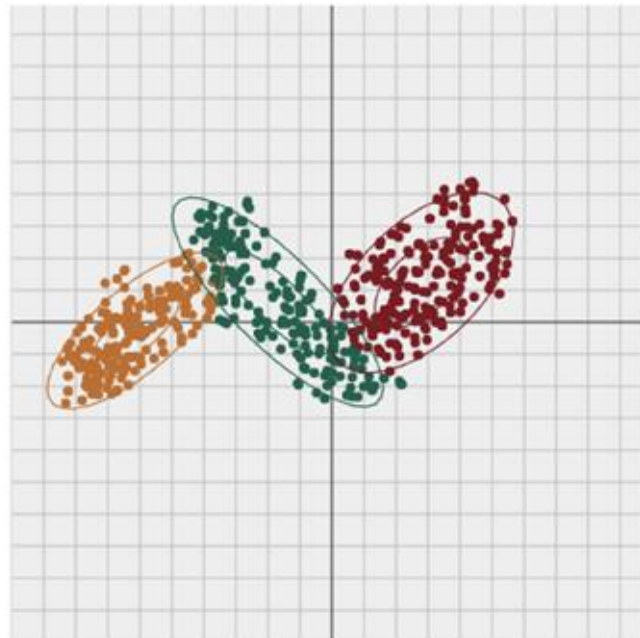
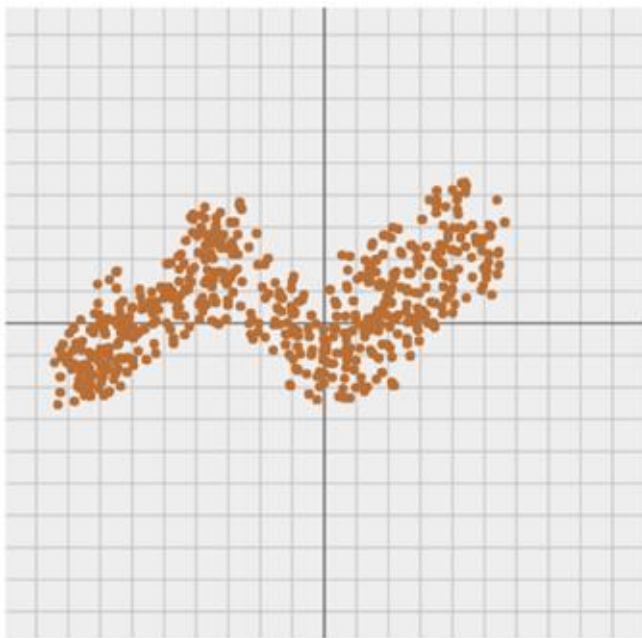
$$p(\mathbf{x} \mid \mathbf{z} = k) = \mathcal{N}(\mu_k, \Sigma_k)$$



Generative process

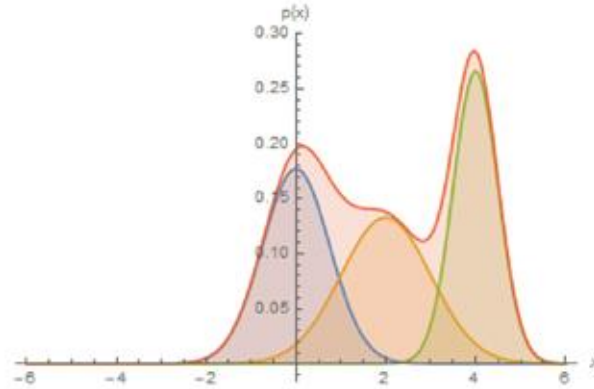
1. Pick a mixture component by sampling z
2. Generate a data point by sampling from that Gaussian

Mixture of Gaussians



Mixture of Gaussians

Combining simple models into more expressive ones



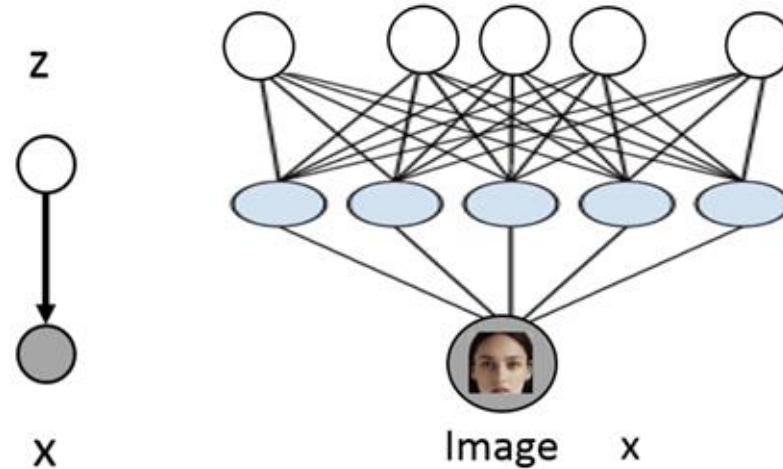
$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x} | \mathbf{z}) = \sum_{k=1}^K p(\mathbf{z} = k) \underbrace{\mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)}_{\text{component}}$$

can solve using expectation maximization

Expectation: use mean and variance to estimate $p(\mathbf{z}=k)$

Maximization: use estimate $p(\mathbf{z}=k)$ to update mean and variance

From GMMs to VAEs



- Put a prior on z $\mathbf{z} \sim \mathcal{N}(0, I)$
 $p(\mathbf{x} | \mathbf{z}) = \mathcal{N}(\mu_{\theta}(\mathbf{z}), \Sigma_{\theta}(\mathbf{z}))$ where $\mu_{\theta}, \Sigma_{\theta}$ are neural networks
- Hope that after training, z will correspond to meaningful latent factors of variation - useful features for unsupervised representation learning
- Given a new image x , features can be extracted via $p(z|x)$
- Given a random z , a new x can be generated => controllable if z is interpretable
- Even though $p(x|z)$ is simple, marginal $p(x)$ is much richer/complex/flexible

Learning Parameters of VAEs

- Learning parameters of VAE: we have a joint distribution $p(\mathbf{X}, \mathbf{Z}; \theta)$
- We have a dataset \mathbf{D} where for each datapoint the \mathbf{x} variables are observed (e.g. images, text) and the variables \mathbf{z} are not observed (latent variables)
- We can try maximum likelihood estimation:

$$\log \prod_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}; \theta) = \sum_{\mathbf{x} \in \mathcal{D}} \log p(\mathbf{x}; \theta) = \sum_{\mathbf{x} \in \mathcal{D}} \log \underbrace{\sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta)}$$

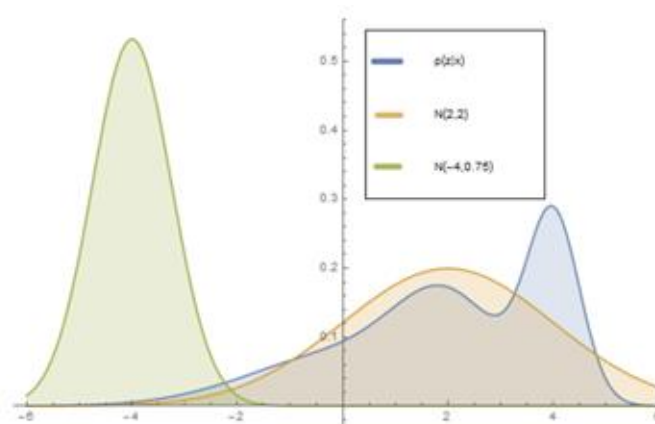
Need cheaper approximations to optimize for VAE parameters

intractable :-)

- if \mathbf{z} binary with 30 dimensions, need sum 2^{30} terms

- if \mathbf{z} continuous, integral is impossible

Variational Inference



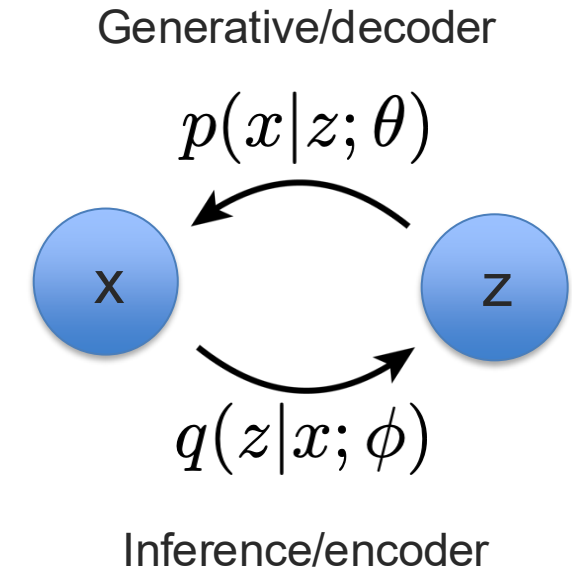
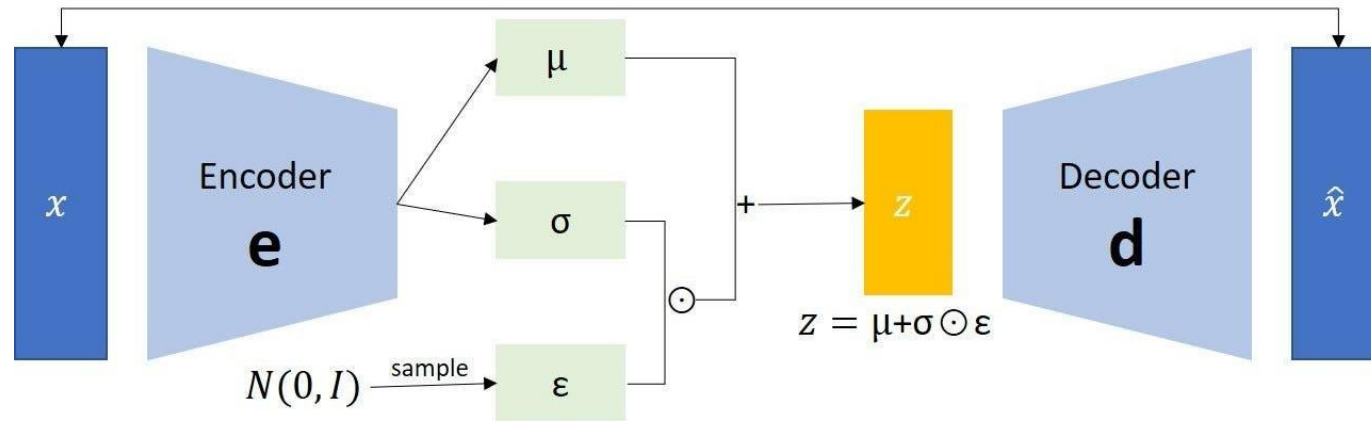
Suppose $q(\mathbf{z}; \phi)$ is a (tractable) probability distribution over the hidden variables parameterized by ϕ (variational parameters)

- For example, a Gaussian with mean and covariance specified by ϕ

$$q(\mathbf{z}; \phi) = \mathcal{N}(\phi_1, \phi_2)$$

- Variational inference: optimize variational parameters so that $q(\mathbf{z}; \phi)$ is **as close as possible** to $p(\mathbf{x}, \mathbf{z}; \theta)$ while being **simple** to compute
- E.g. in figure, posterior (in blue) is better approximated by orange Gaussian than green

Learning Parameters of VAEs

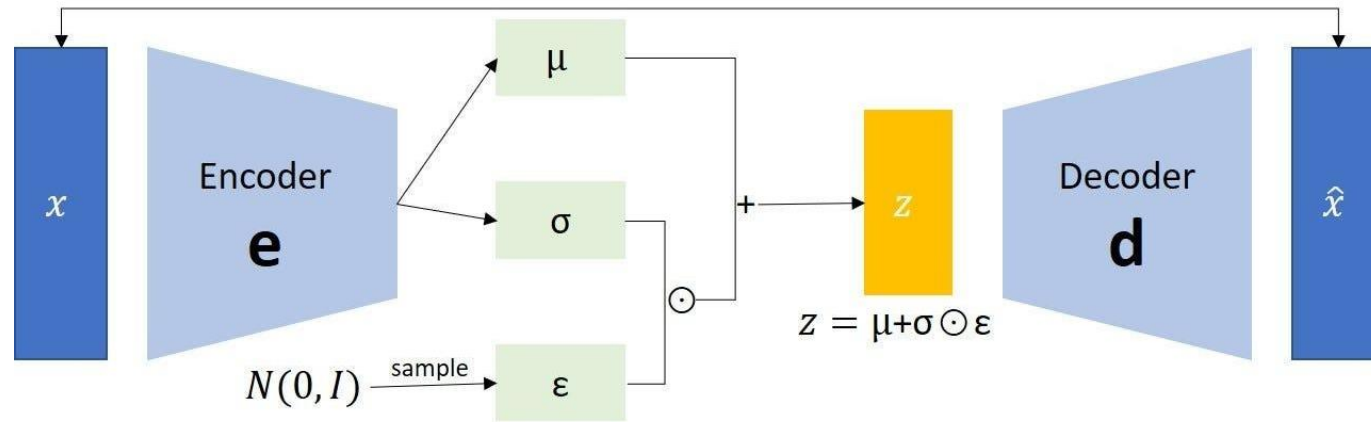


1. Take a datapoint x_i .
2. Map it to μ, σ using $q_\phi(z|x_i)$. **encoder**
3. Sample $\epsilon \sim N(0, I)$ and compute $\hat{z} = \mu + \sigma\epsilon$. **re-parameterize**
4. Reconstruct \hat{x} by sampling from $p(x|\hat{z}; \theta)$. **decoder**

$$\mathcal{L}(\mathbf{x}; \theta, \phi) = \underbrace{E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z}; \theta)]}_{\text{reconstruction } \mathbf{x} \rightarrow \mathbf{z} \rightarrow \mathbf{x}} - \underbrace{D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))}_{\text{prior over } \mathbf{z}, \text{ standard Gaussian}}$$

Prior on \mathbf{z} : $\mathbf{z} \sim \mathcal{N}(0, I)$

Learning Parameters of VAEs

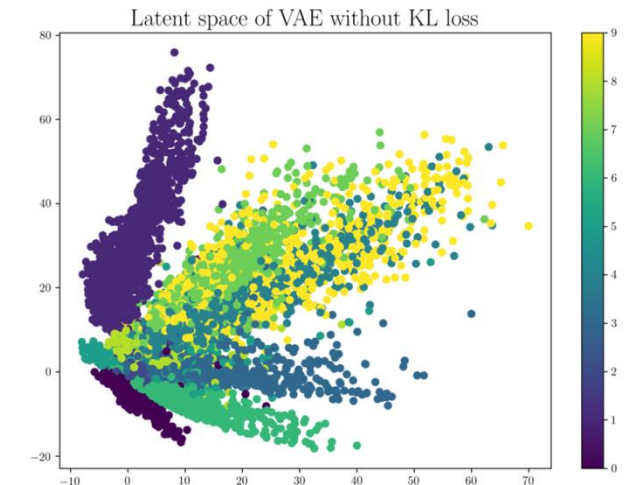
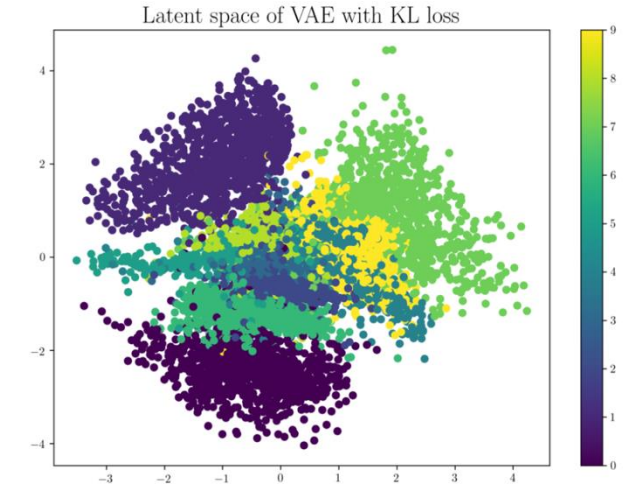


1. Take a datapoint x_i .
2. Map it to μ, σ using $q_\phi(z|x_i)$. **encoder**
3. Sample $\epsilon \sim N(0, I)$ and compute $\hat{z} = \mu + \sigma\epsilon$. **re-parameterize**
4. Reconstruct \hat{x} by sampling from $p(x|\hat{z}; \theta)$. **decoder**

$$\mathcal{L}(\mathbf{x}; \theta, \phi) = \underbrace{E_{q_\phi(z|x)}[\log p(\mathbf{x}|z; \theta)]}_{\text{reconstruction}} - \underbrace{D_{KL}(q_\phi(z|x) \| p(z))}_{\text{prior over } z, \text{ standard Gaussian}}$$

reconstruction
 $x \rightarrow z \rightarrow x$

prior over z ,
standard Gaussian

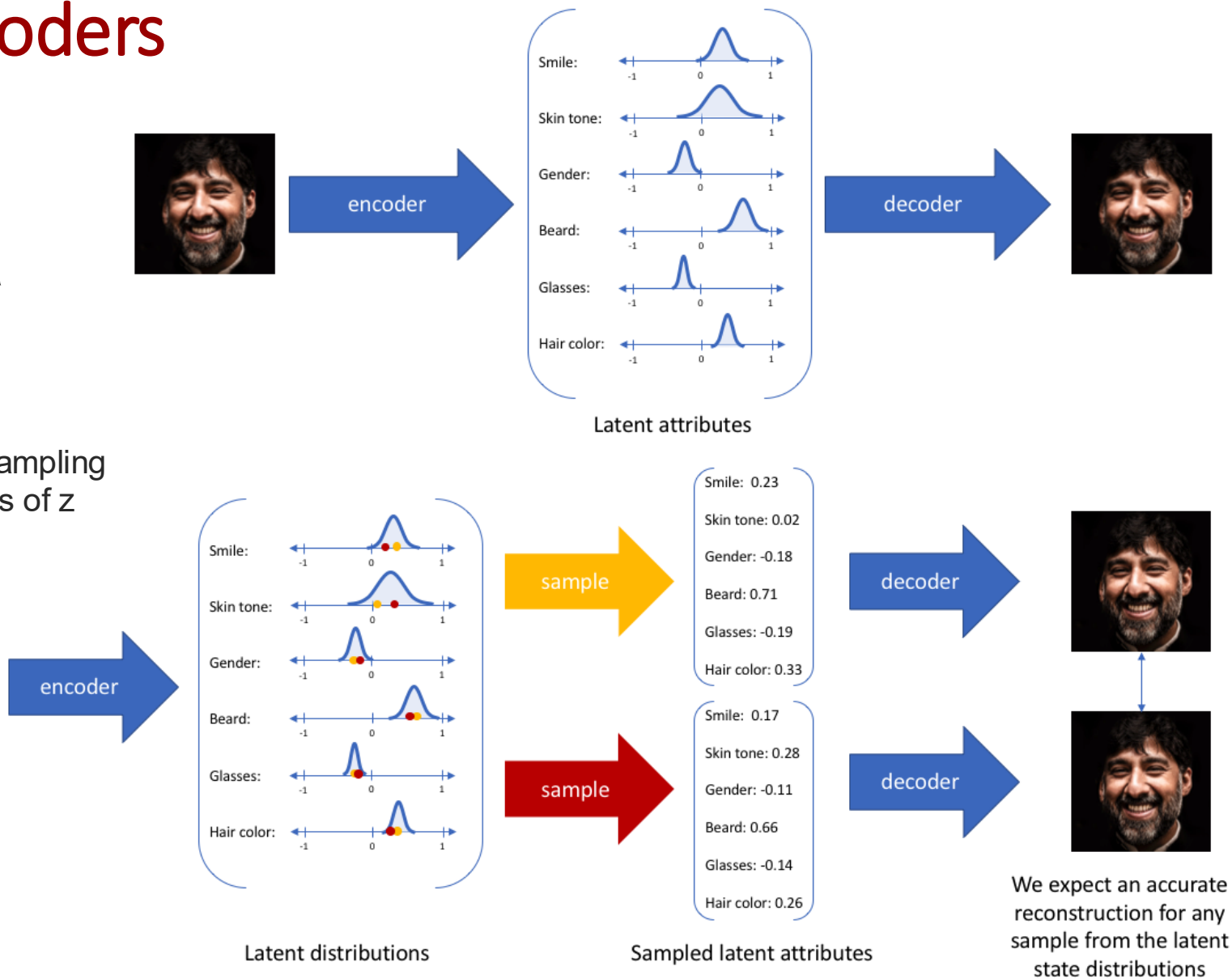


Prior on z : $\mathbf{z} \sim \mathcal{N}(0, I)$

Variational Autoencoders

Key ideas:

1. Encoder decoder structure.
2. Simple latent variable z .
3. Complex $p(x|z)$ decoder via neural networks.
4. Reconstruction objective.
5. Prior over latent variable z :
 - smoother latent space permits sampling
 - disentangles different dimensions of z



Lecture Summary

- 1 Text-to-image generation
- 2 Introduction to generative models
- 3 GMMs and VAEs

Assignments for This Coming Week

HW3 released, due 3/17

Midterm exam next Thursday 3/19. Midterm review next Tuesday 3/17.

Project mentors released. Try to meet as often as you can. Meet with me today.