Lecture 8: Vision Generative Models in Biomedicine

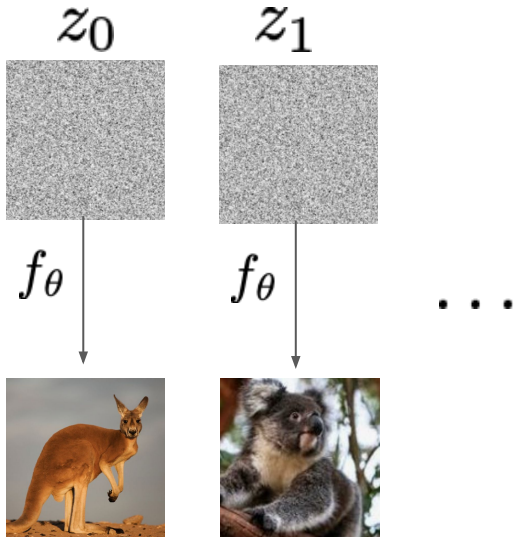
Announcements

- A1 is due today at 11:59pm
- Project proposal is due Wed Oct 23
- Submit discussion presentation preferences by tonight. See Ed for instructions.

Finishing up from last lecture: Vision Diffusion and Generative Models

Very quick recap

Generative modeling: the goal

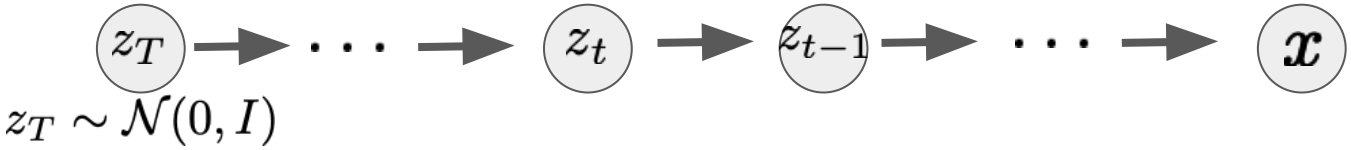


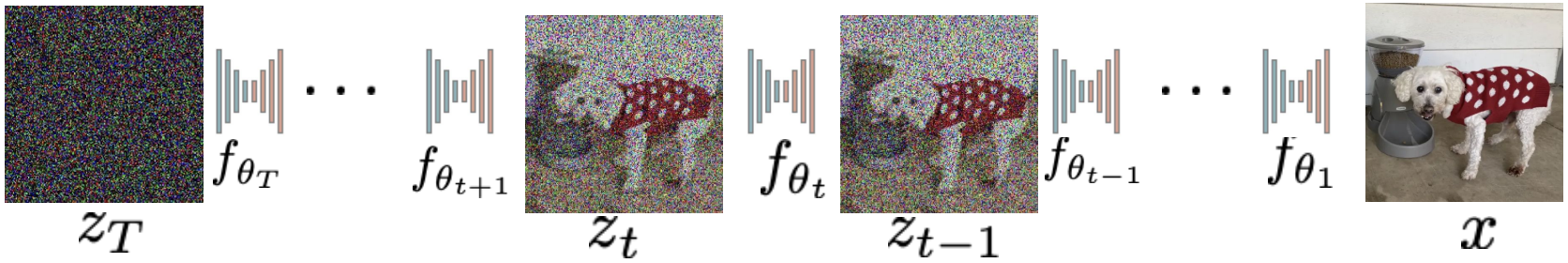
Training

Generation

Sampling a diffusion model

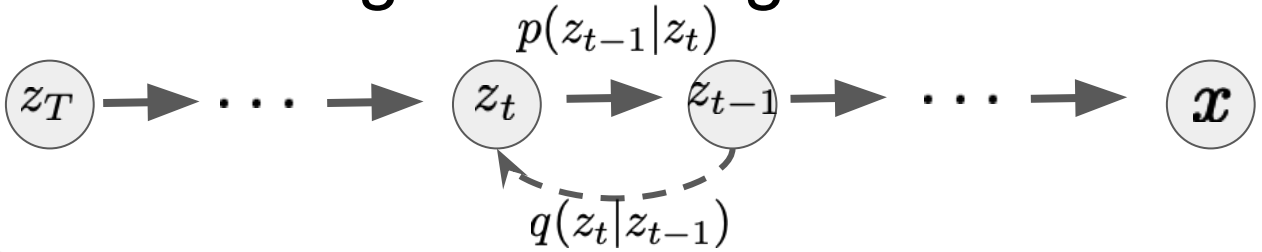
Noise to data with multiple steps



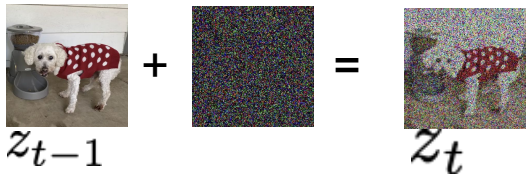


One neural net for each timestep, f_{θ_t}

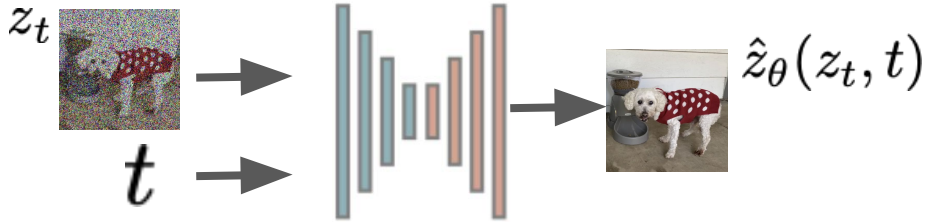
Diffusion training as learning to reverse noising



Sample t,
Apply noise: $q(z_t|z_{t-1})$



Use Unet to reverse the noise



Reconstruction loss

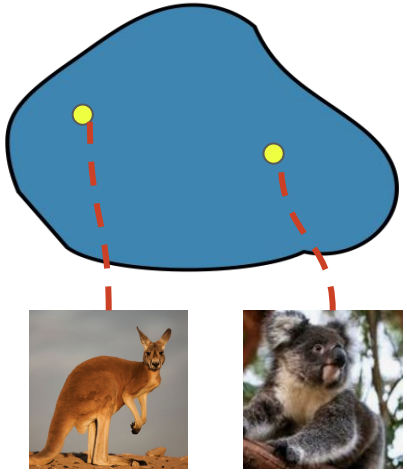
$$\mathbb{E}_{t\sim U(2,T)}\left[c(\alpha_t)\|\hat{z}_\theta(z_t,t) - z_{t-1}\|\right]$$

More advanced sampling

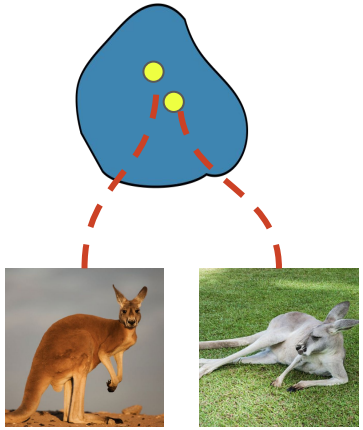
Conditional sampling

Recall: we want to be able to add conditioning information

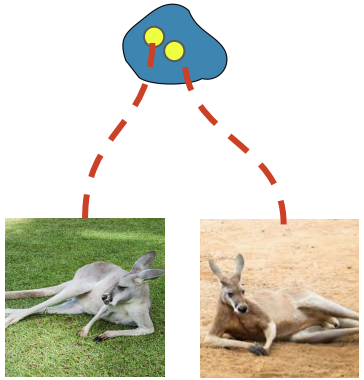
<no condition>



“kangaroo”



“Kangaroo lying down”’

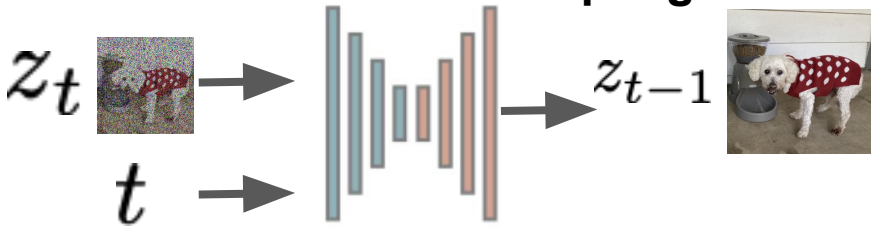


Conditional sampling

Solution: pass the condition info to the Unet

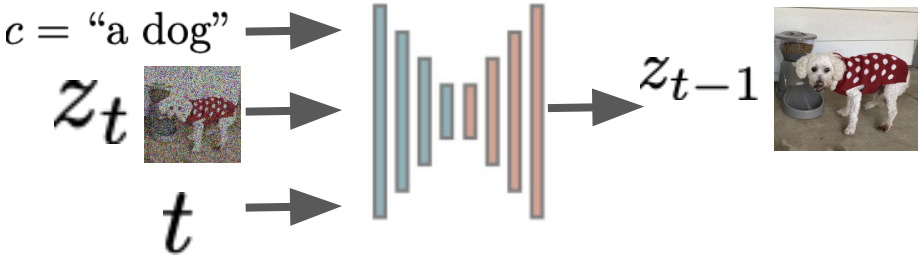
Te loss is almost the same

Before: unconditional sampling



$$\mathbb{E}_{t\sim(2,T)}\left[C(\alpha_t)\cdot\|\hat{z}_\theta(z_t,t)-z_t\|\right]$$

After: with a condition



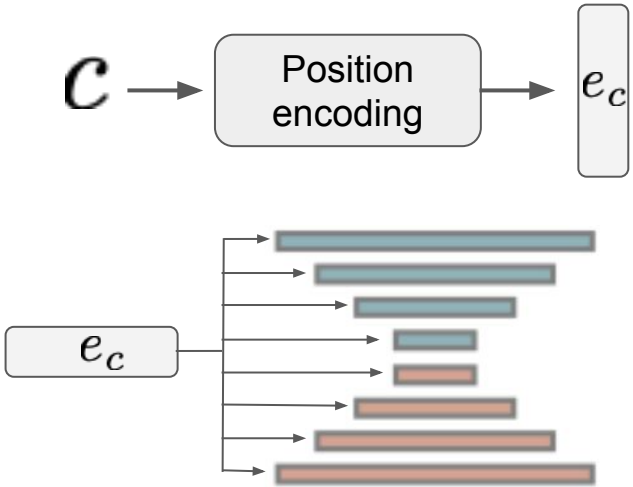
After

$$\mathbb{E}_{t\sim(2,T)}\left[C(\alpha_t)\cdot\|\hat{z}_\theta(z_t,t,c)-z_t\|\right.$$

Implementing conditional sampling

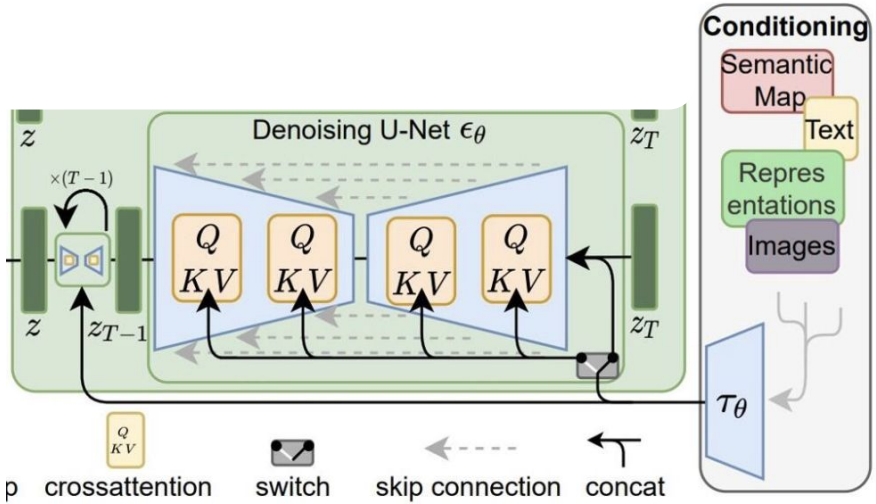
Class-based conditioning

Encode the same way `t` is encoded



General conditioning via cross-attention

Much more common



Conditional sampling in practice with “guidance”

Sampling (using ‘score matching’ perspective)
We move in the gradient direction

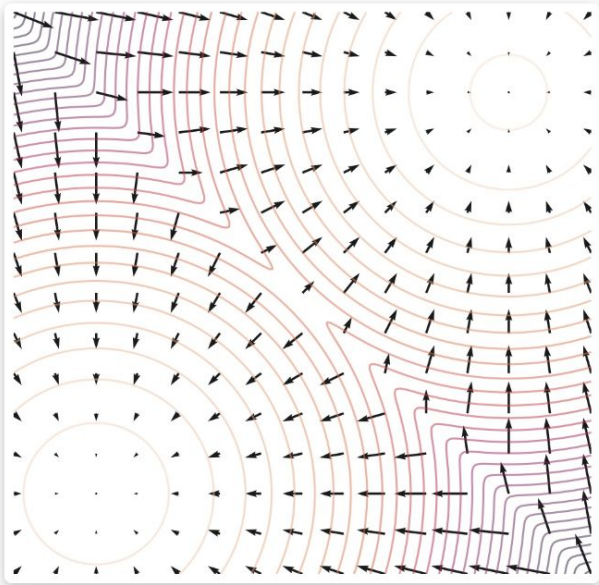
$$\nabla_x \log p(x)$$

Now we do conditioning:

$$\nabla_x \log p(x|c)$$

Guidance says:
“let’s balance the unconditional and conditional directions”

$$\nabla_x \log p(x) + \gamma \cdot \nabla_x \log p(x|c)$$



Good blog post: [Guidance: a cheat code for diffusion models]

Conditional sampling in practice with “guidance”

The γ parameter lets us trade off diversity and fidelity

c=“*A stain glass window of a panda eating bamboo*”

Better diversity - smaller γ

Better fidelity - bigger γ

Good blog post: [Guidance: a cheat code for diffusion models]

Conditional sampling in practice with “guidance”

Guidance has an unconditional and conditional term

$$\nabla_x \log p(x) + \gamma \cdot \nabla_x \log p(x|c)$$

Problem: we train $\nabla_x \log p(x|c)$ so how to get $\nabla_x \log p(x)$

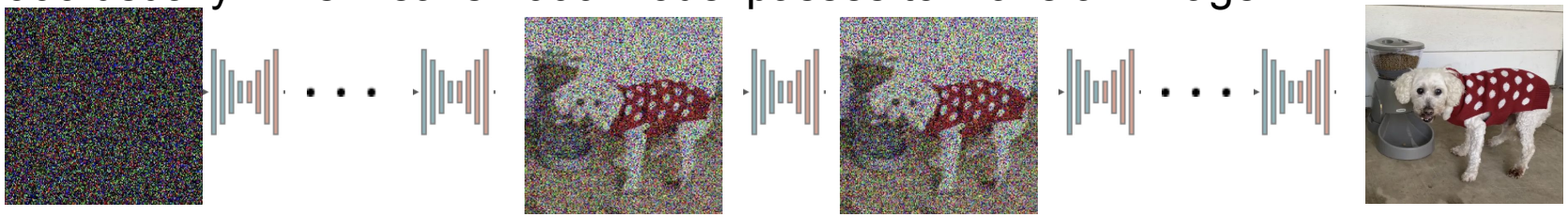
Answer: Use the same model, but let ‘unconditional’ be a class
$p(x) = p(x|c = \emptyset)$

In training, for 10% of samples, set the condition to this.

Good blog post: [Guidance: a cheat code for diffusion models]

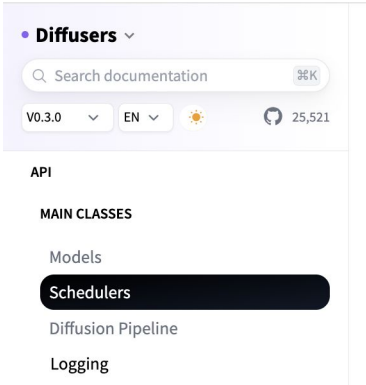
Faster sampling - schedulers

T=1000 usually. This means 1000 model passes to make an image



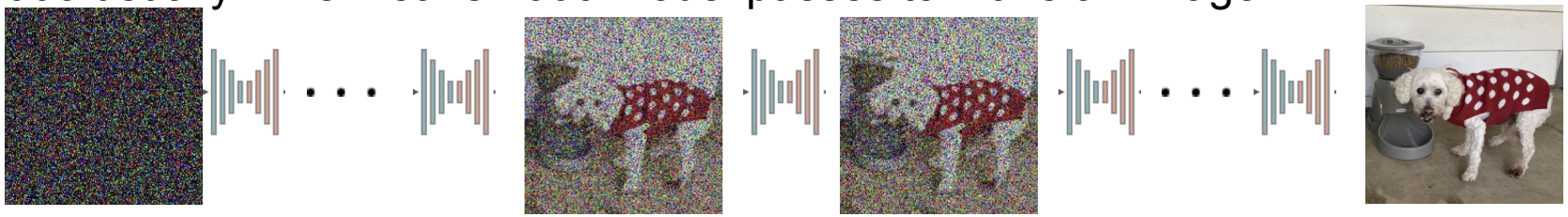
Different *samplers* (like DPMSolver) can reduce to 20 steps without changing the model

In diffusers library, look up “schedulers”

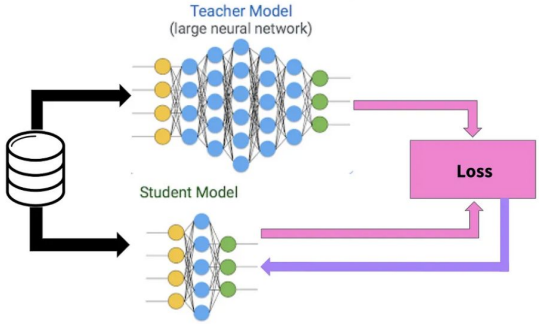


Faster sampling - distillation

T=1000 usually. This means 1000 model passes to make an image



Classic distillation: big model teaches smaller model



Diffusion distillation: High ‘T’ model teaches low ‘T’ model

See blog post [The paradox of diffusion distillation]

Important diffusion model designs

DDPM

Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

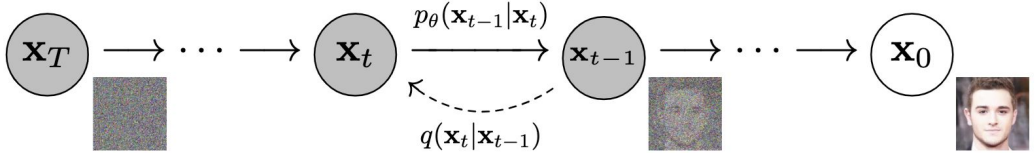
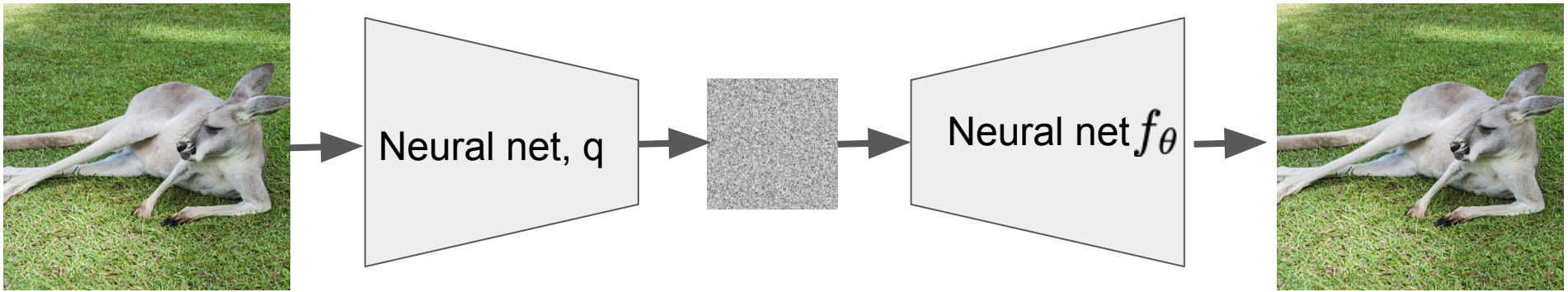


Figure 2: The directed graphical model considered in this work.

DDPM is *mostly* the model we have described up till now

Stable Diffusion or Latent Variable Models

Recall: VAEs can do data compression and reconstruction

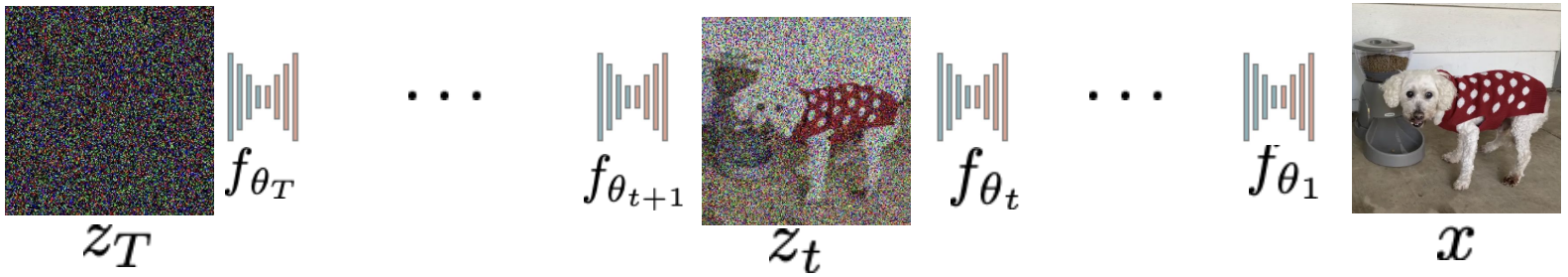


"High-Resolution Image Synthesis with Latent Diffusion Models"
"Score-based Generative Modeling in Latent Space"

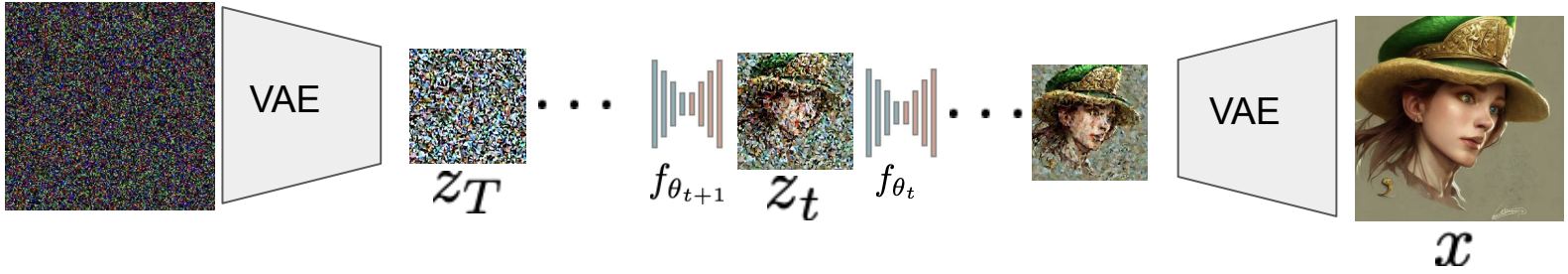
Stable Diffusion or Latent Variable Models

Idea: instead of diffusion models on *pixels*, do diffusion in a *compressed latent space*

Diffusion (DDPM)



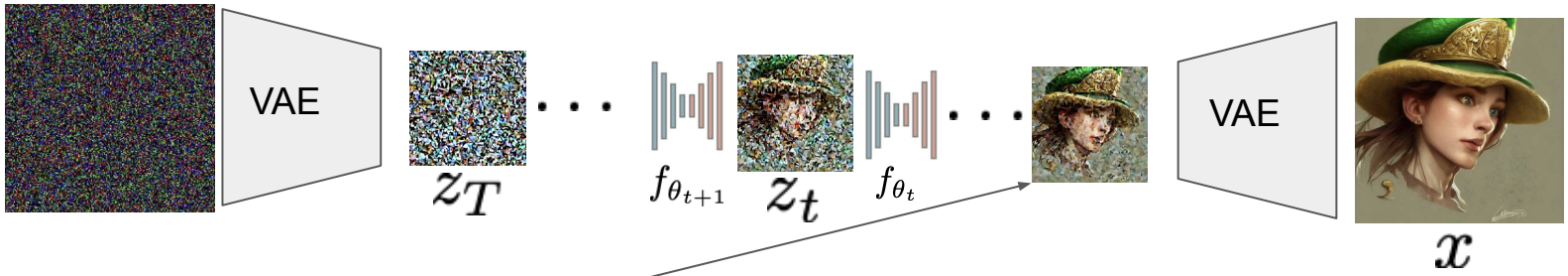
Latent diffusion (Stable Diffusion)



"High-Resolution Image Synthesis with Latent Diffusion Models"
"Score-based Generative Modeling in Latent Space"

Stable Diffusion or Latent Variable Models

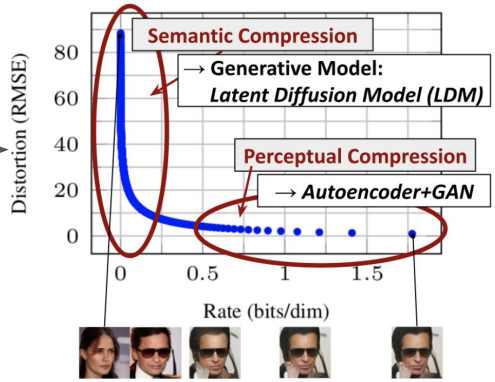
Idea: instead of diffusion models on *pixels*, do diffusion in a *compressed latent space*



Choose a VAE with low distortion

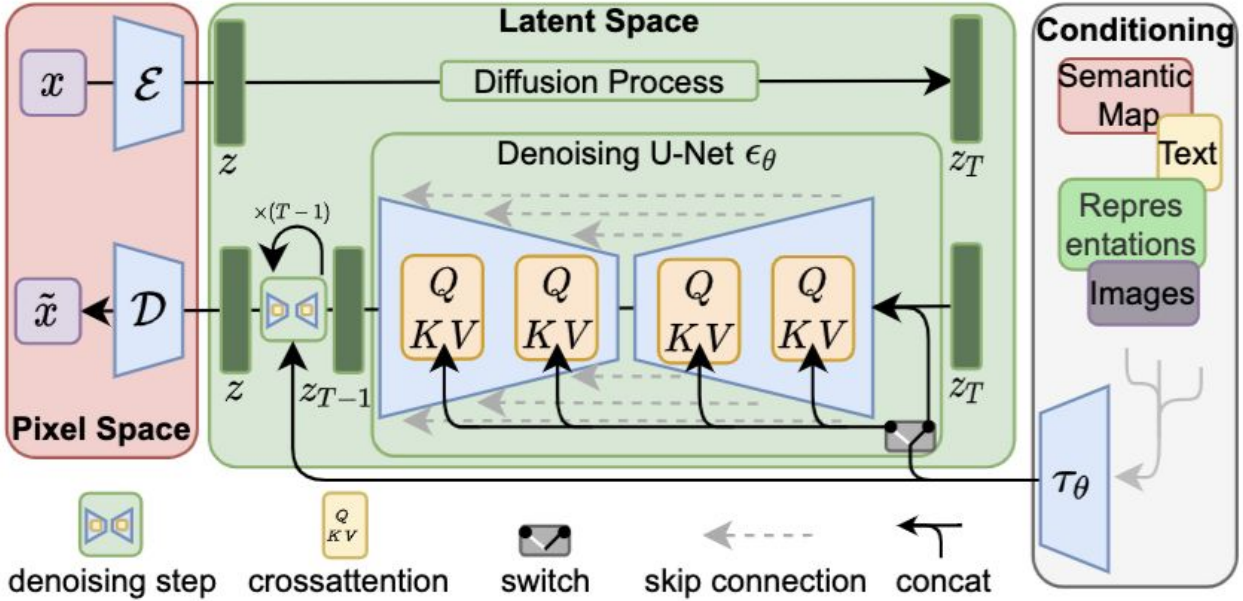
Key insight: VAE can downscale images a lot while maintaining semantic information.

Diffusion in latent is much more efficient



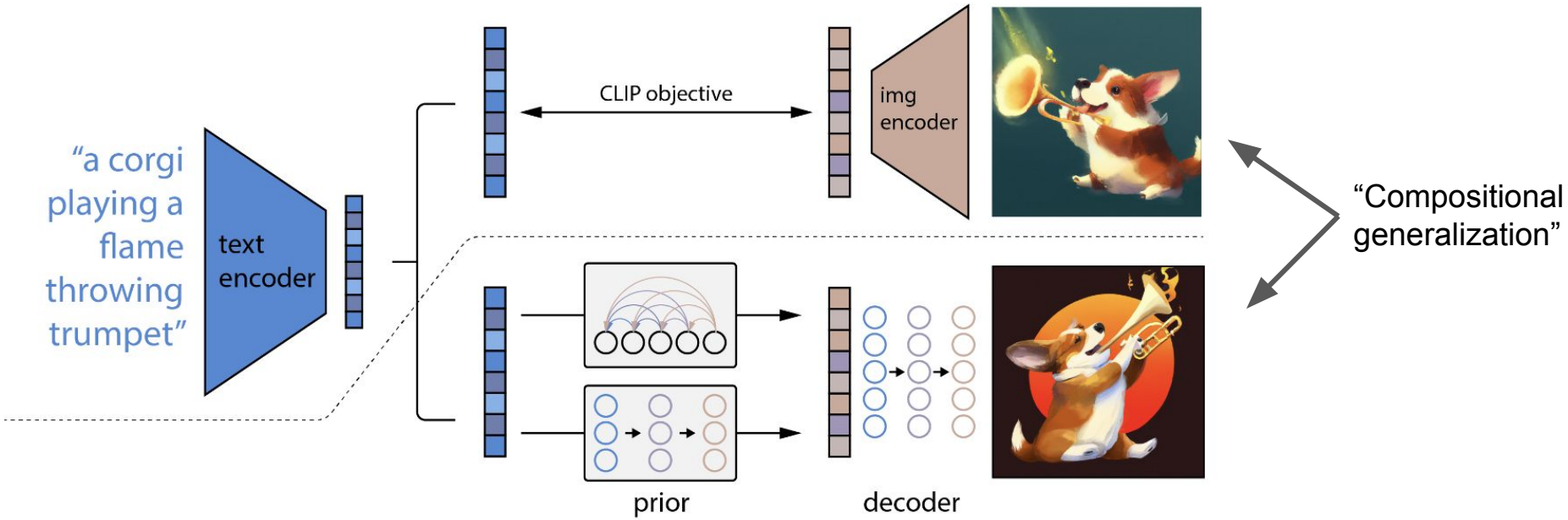
"High-Resolution Image Synthesis with Latent Diffusion Models"
"Score-based Generative Modeling in Latent Space"

Stable Diffusion or Latent Variable Models



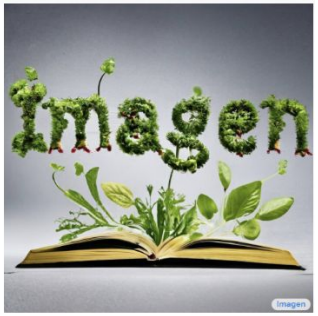
"High-Resolution Image Synthesis with Latent Diffusion Models"
"Score-based Generative Modeling in Latent Space"

DALLE-2



"Hierarchical Text-Conditional Image Generation with CLIP Latents"

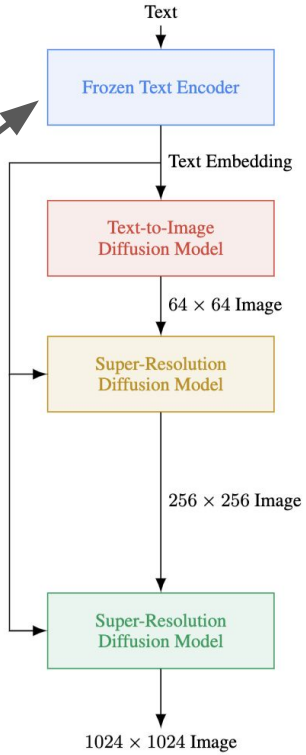
Imagen

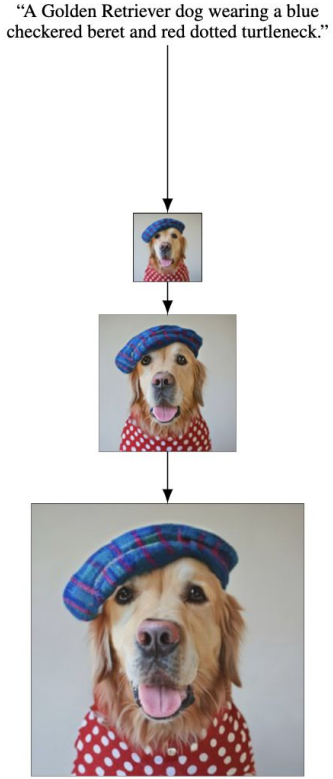


Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

Teddy bears swimming at the Olympics 400m Butterfly event.

Pretrained language-only model (others use CLIP language encoder)



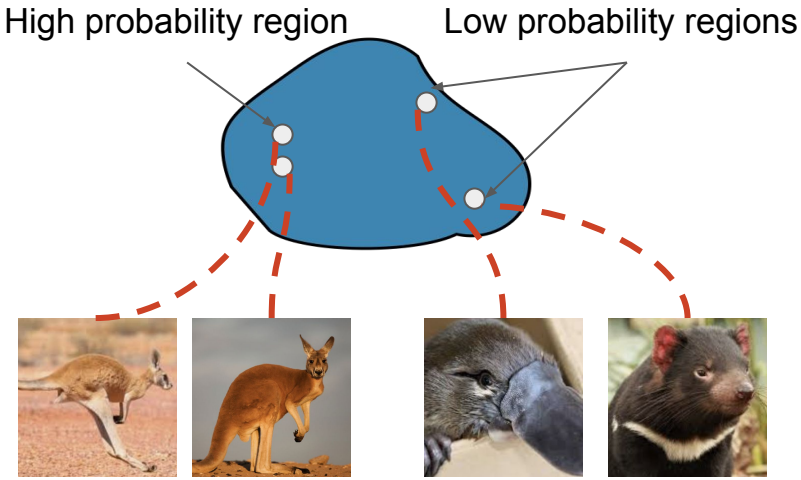


"Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding"

Evaluation

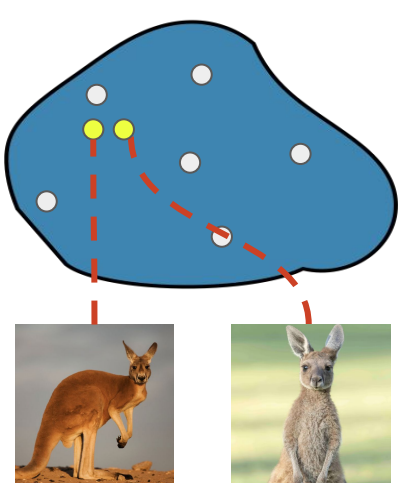
Evaluation

Desiderata for diffusion models

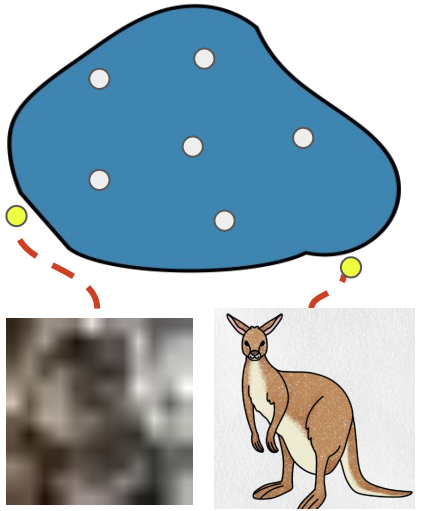


Sampling likelihood should match the data likelihood

Common issues



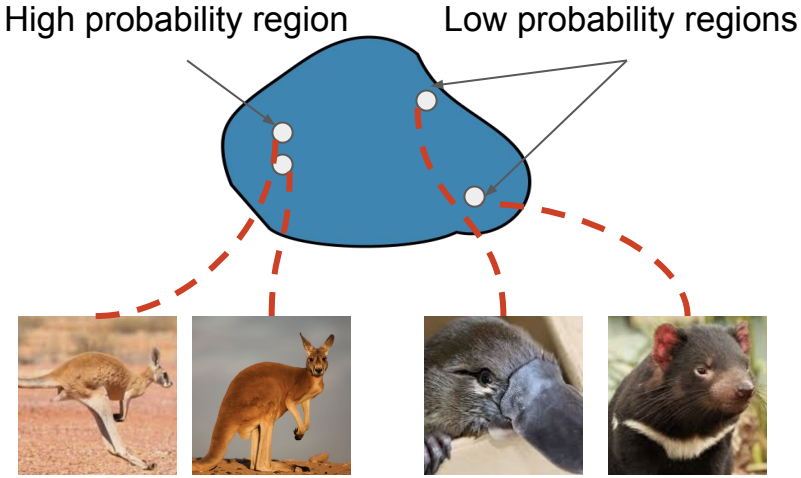
Diversity: not sampling the whole distribution



Fidelity: samples not from the right distribution

Evaluation

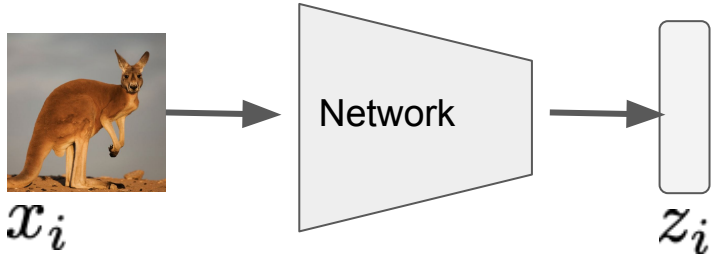
Desiderata for diffusion models



Sampling likelihood should match the data likelihood

Challenge: how to evaluate distribution in this “semantic latent space”

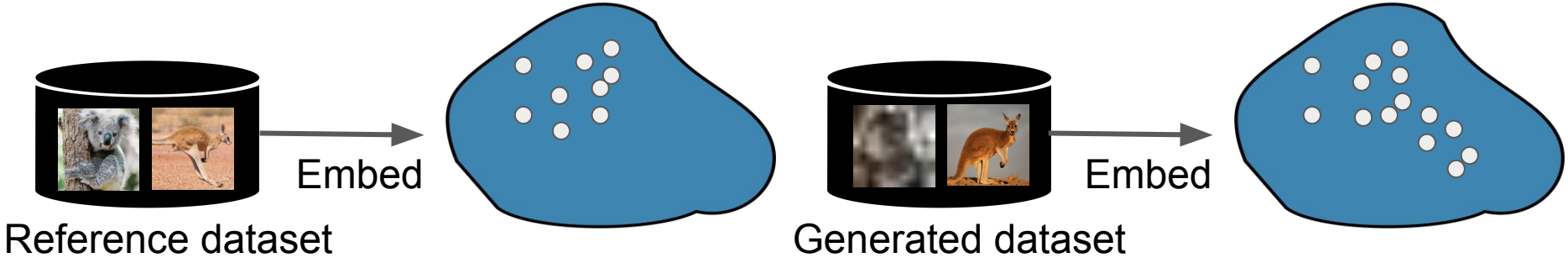
Approach: embed with a pretrained network, e.g. ‘inception network’ trained on ImageNet



*choice of network will depend on data
E.g. MRI images should not use networks trained on ImageNet

Evaluation: Frechet Inception Distance (FID)

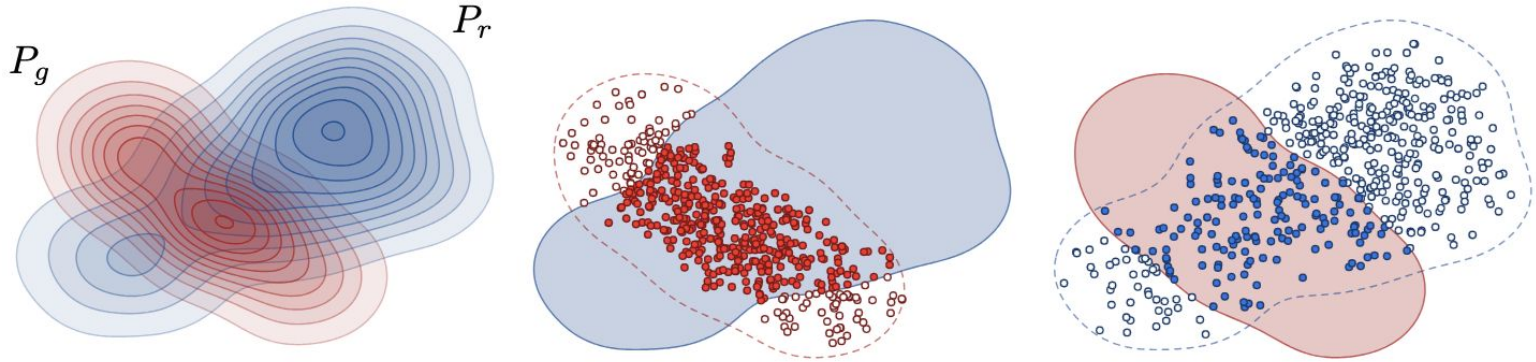
Idea: generated samples should match a target distribution, e.g. the ImageNet test set



1) Map dataset and embeddings to embeddings
2) fit a Gaussian to each distribution of embeddings
3) Measure a distance between those distributions

$$d_F(\mathcal{N}(\mu,\Sigma),\mathcal{N}(\mu',\Sigma'))^2 = \|\mu-\mu'\|_2^2 + \mathrm{tr}\left(\Sigma+\Sigma'-2(\Sigma\Sigma')^{\frac{1}{2}}\right)$$

Evaluation: precision and recall



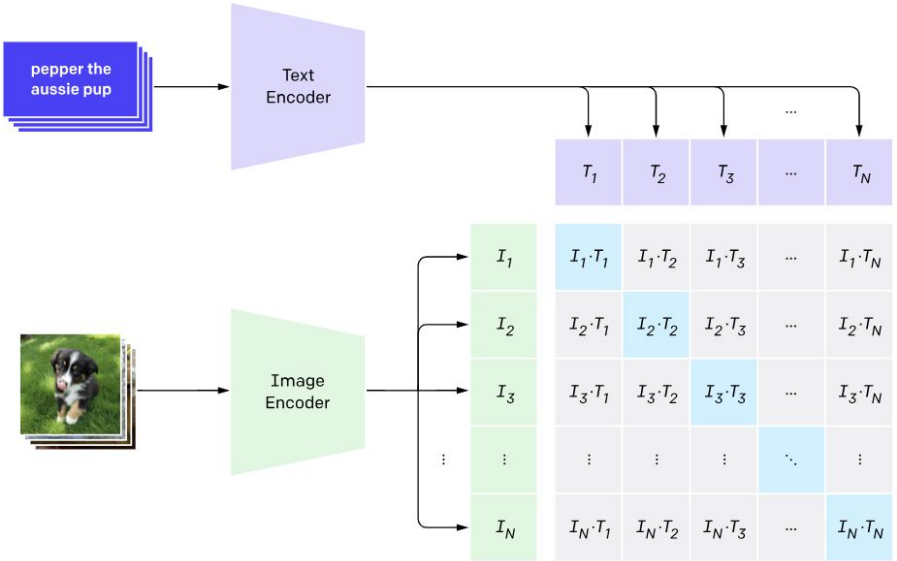
(a) Example distributions (b) Precision (c) Recall

Figure 1: Definition of precision and recall for distributions [25]. (a) Denote the distribution of real images with P_r (blue) and the distribution of generated images with P_g (red). (b) Precision is the probability that a random image from P_g falls within the support of P_r. (c) Recall is the probability that a random image from P_r falls within the support of P_g.

"Improved Precision and Recall Metric for Assessing Generative Models"

Evaluation: CLIP score

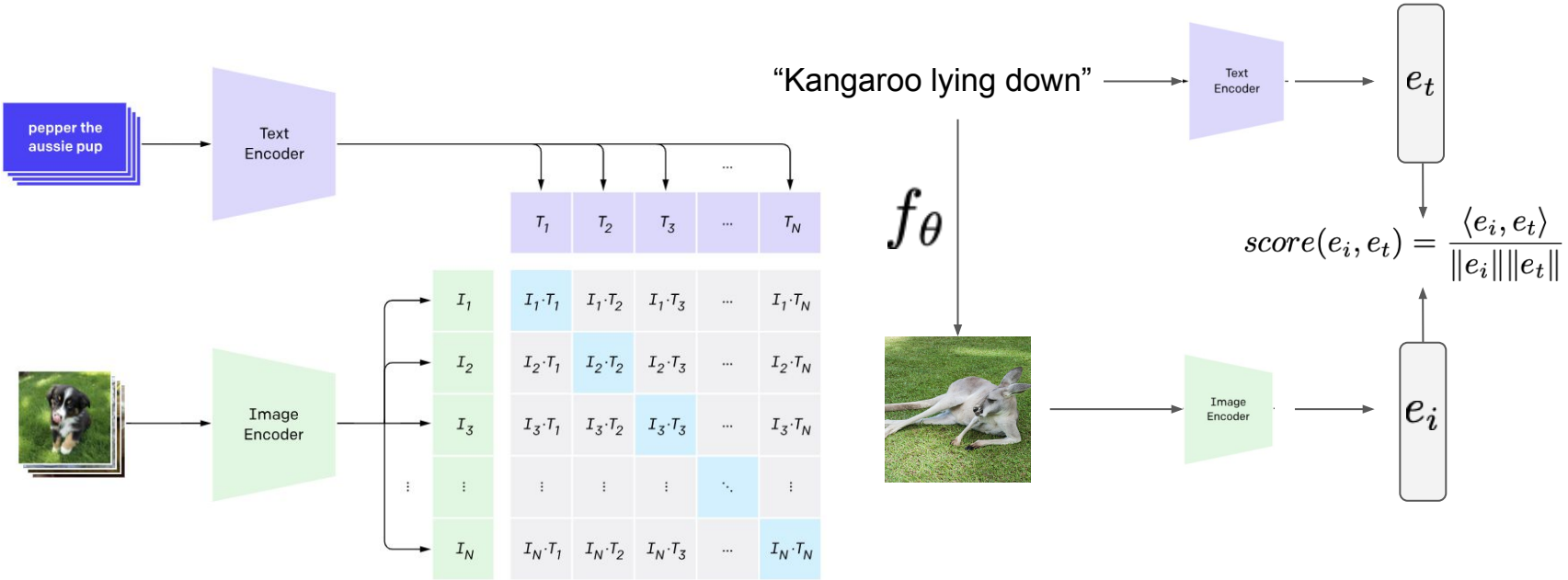
Idea: image should be close to its text prompt in CLIP space



"CLIPScore: A Reference-free Evaluation Metric for Image Captioning"

Evaluation: CLIP score

Idea: image should be close to its text prompt in CLIP space



"CLIPScore: A Reference-free Evaluation Metric for Image Captioning"

Evaluation: compositional generalization

People are interested in composing concepts in novel ways in diffusion models

Concept categories: object | texture | color | shape | style | number | spatial | size

A small pink duck stands on a hill with a metallic texture. The image is photorealism.

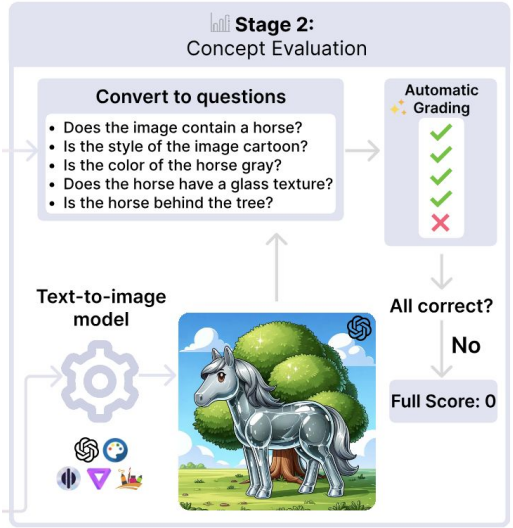
k=5

Four tiny, rectangular sushi pieces with a glass-like texture are positioned behind a tray in an expressionist style.

k=6

In a cubism style, four tiny, gray, glass-textured, rectangular laptops are positioned on top of a table.

k=7



This benchmark uses automatic eval using an image recognition model

"ConceptMix: A Compositional Image Generation Benchmark with Controllable Difficulty"

Diffusion beyond text-to-image

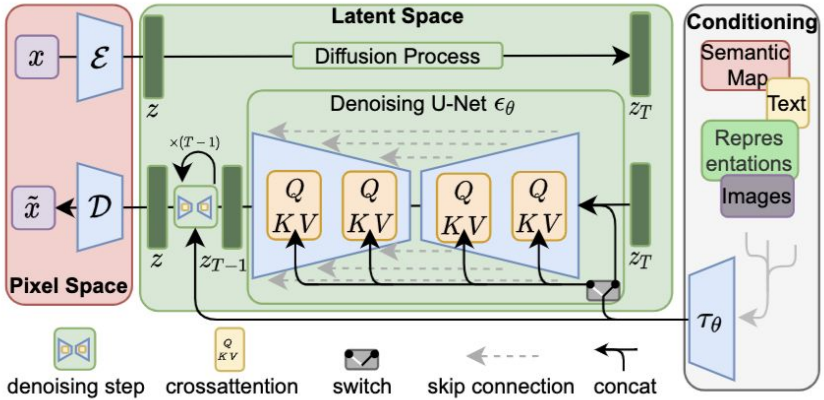
A few examples of generative model tasks in images

Text-to-image

"A kangaroo lying down"'

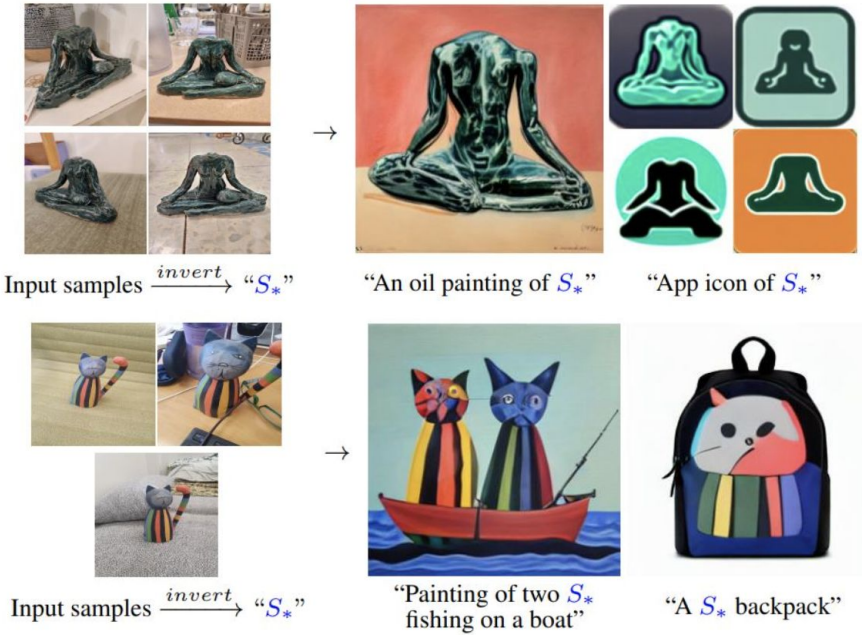
Super-resolution

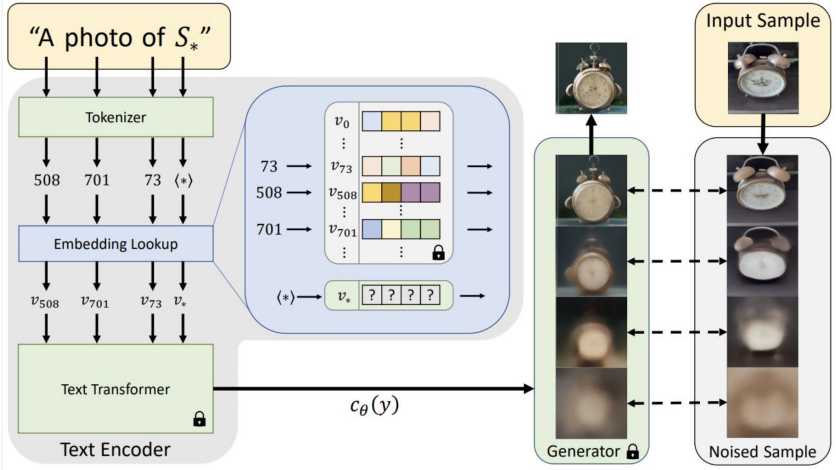
Image inpainting



Textual inversion: learning to prompt new concepts

Idea: learn an input “word” to capture a new concept from a few images, and generate it

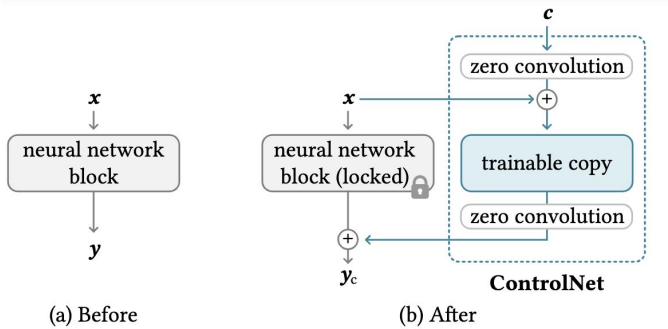




“An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion”

Controlnet: adding new conditions to text-to-image

If we start with a t2i model like StableDiffusion, then it *should* be easier to learn to map other conditioning information



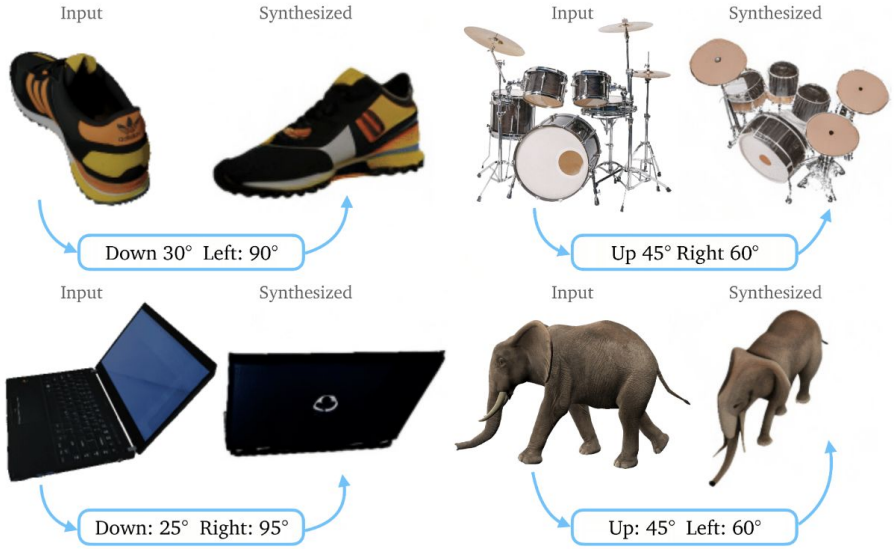
"Adding Conditional Control to Text-to-Image Diffusion Models"

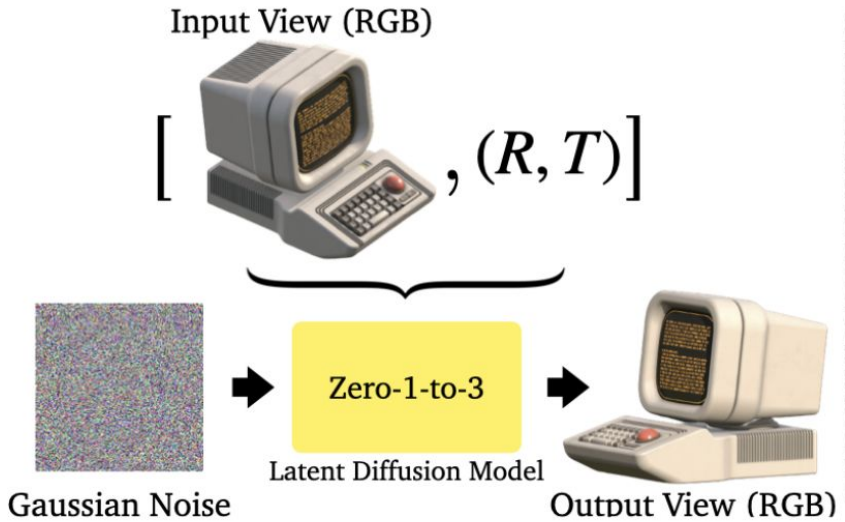
General theme: better control of image generation

"MoA : Mixture-of-Attention for Subject-Context Disentanglement in Personalized Image Generation"

Zero-1-to-3: rotating objects in 3D

Key idea: rotate objects by conditioning diffusion on starting image = angles



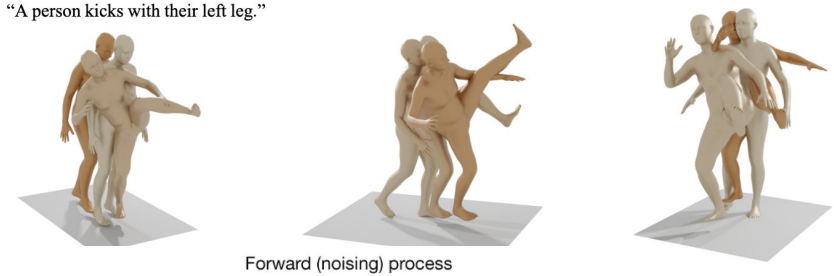


"Zero-1-to-3: Zero-shot One Image to 3D Object"

Diffusion on other data types

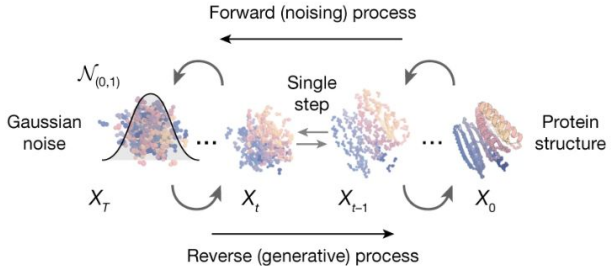
Human motion

"MDM: Human Motion Diffusion Model"



Proteins

"De novo design of protein structure and function with RFdiffusion"



Language

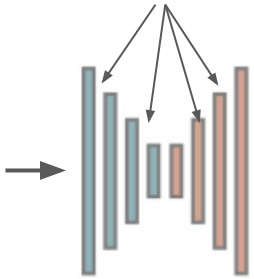
"Language Modeling by Estimating the Ratios of the Data Distribution"

study ants bear burrito skyline song

Diffusion representations for dense perception tasks

Idea: Unet activations capture meaningful content at the pixel level

Features from Unet

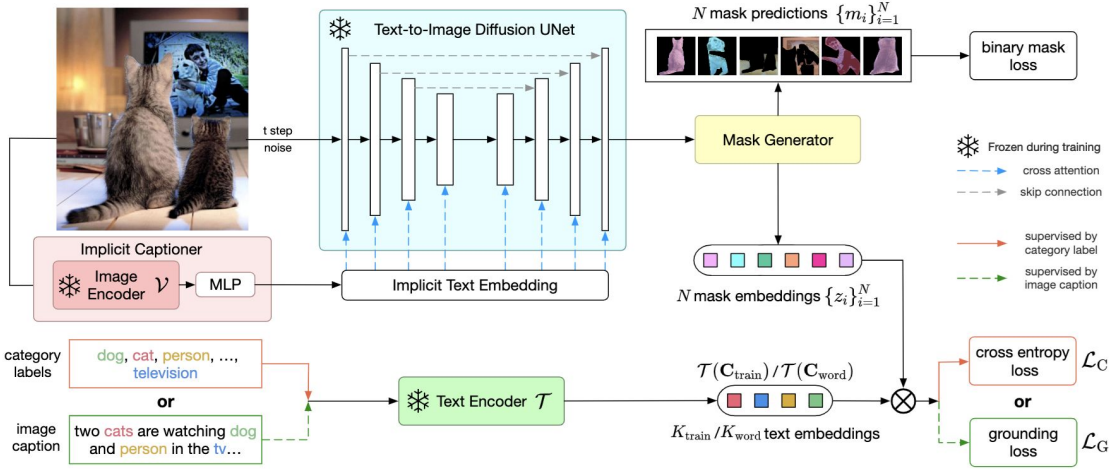


K-Means Clustering of
Frozen Diffusion Features

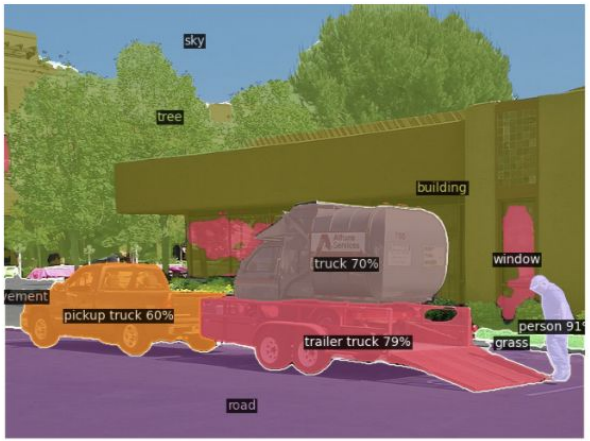
“Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models”

Diffusion representations for dense perception tasks

Idea: Unet activations capture meaningful content at the pixel level
So: use it as a backbone in segmentation



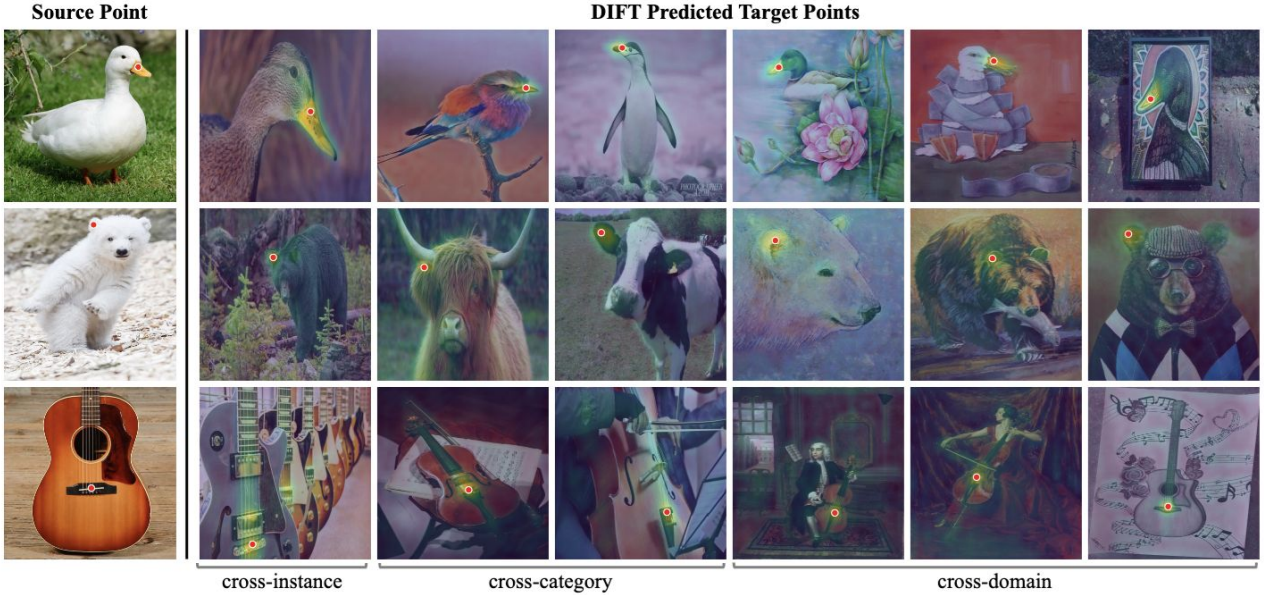
Open-Vocabulary Panoptic Segmentation
Prediction from ODISE



"Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models"

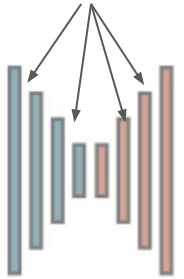
Diffusion representations for dense perception tasks

Idea: use Unet activations and find point-point correspondence by feature similarity



"Emergent Correspondence from Image Diffusion"

Copy features from Unet

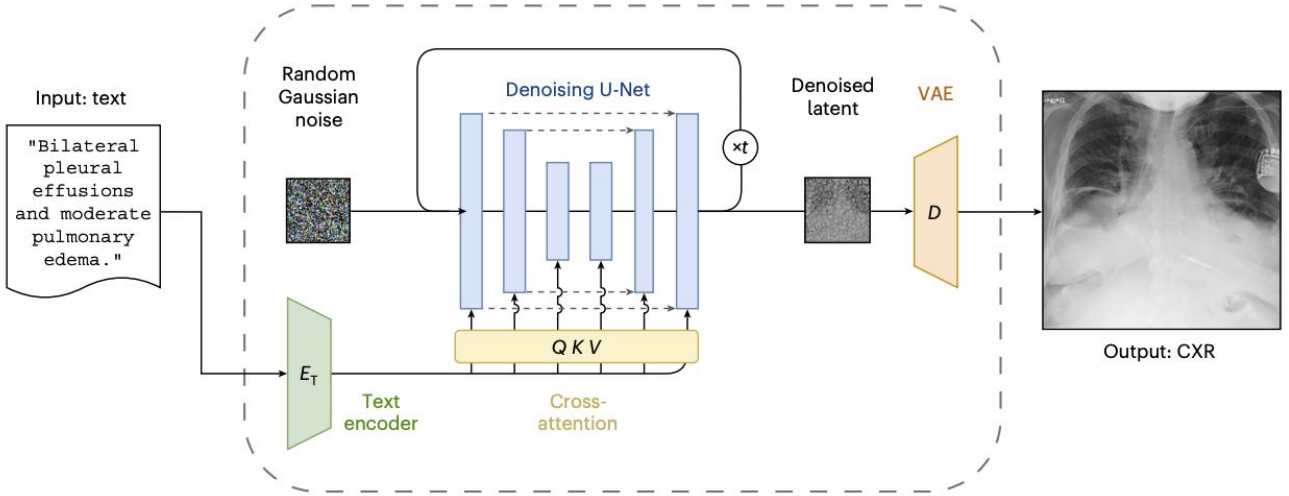


Use vector dot product to find most similar point in 2nd image

Next:
Vision Diffusion and Generative Models in Biomedicine

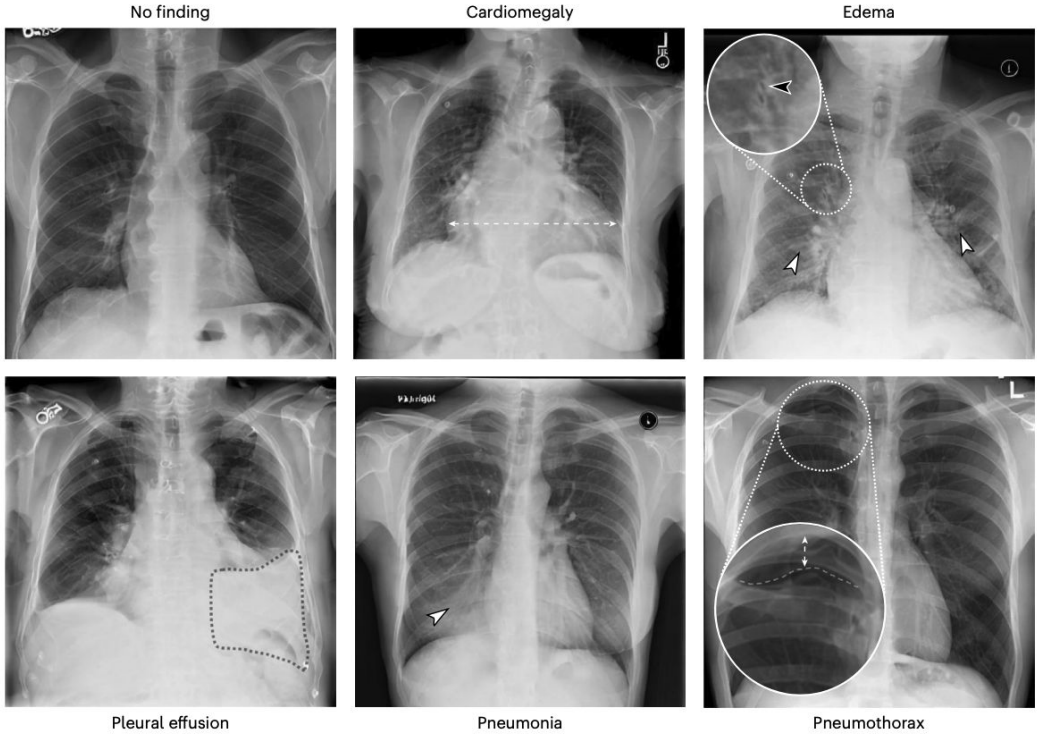
One widely considered application area of vision generative models: augmenting training data

- RoentGen fine-tunes the Stable Diffusion model on the MIMIC-CXR dataset of chest x-ray (CXR) images and corresponding text reports (about ~175k images for training)



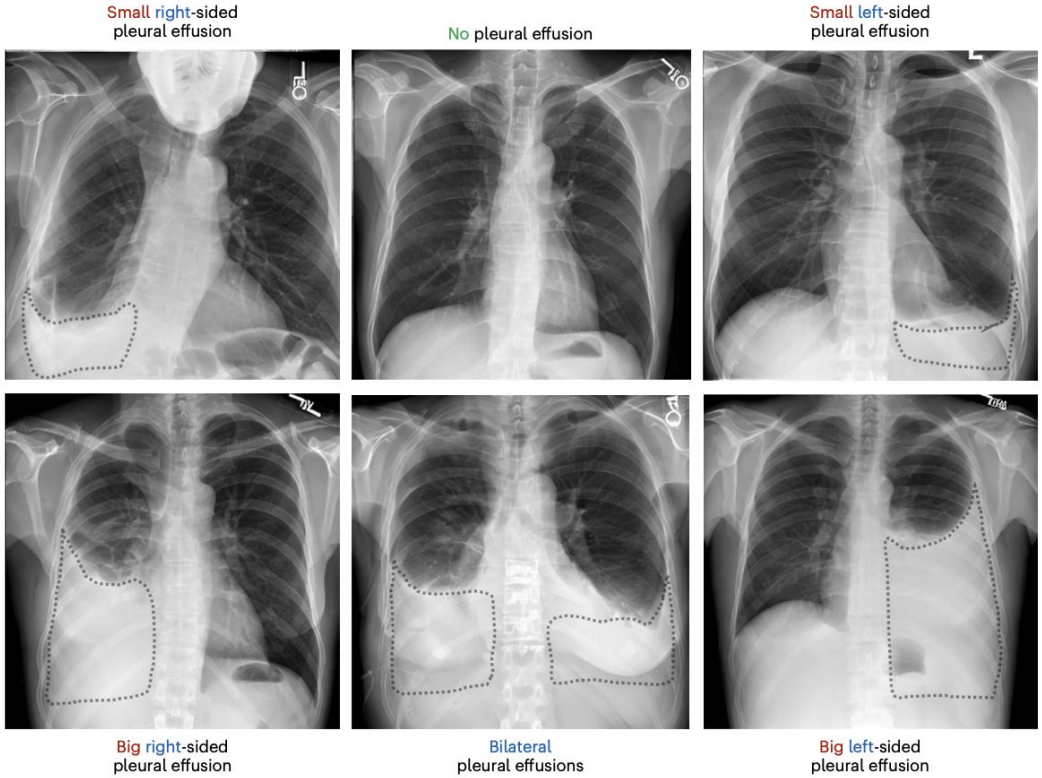
Bluethgen et al. A vision–language foundation model for the generation of realistic chest X-ray images. Nature Biomedical Engineering 2024.

Examples of synthetically generated chest x-rays for different conditions



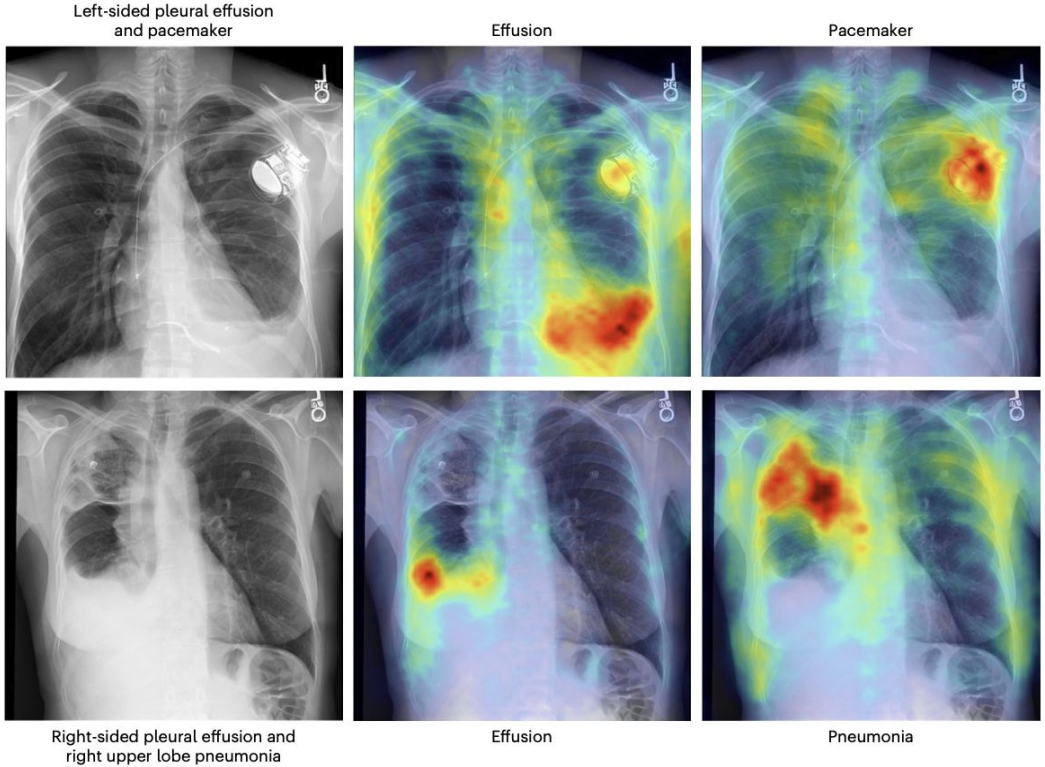
Bluethgen et al. A vision–language foundation model for the generation of realistic chest X-ray images. Nature Biomedical Engineering 2024.

Finer-grained text control



Bluethgen et al. A vision–language foundation model for the generation of realistic chest X-ray images. Nature Biomedical Engineering 2024.

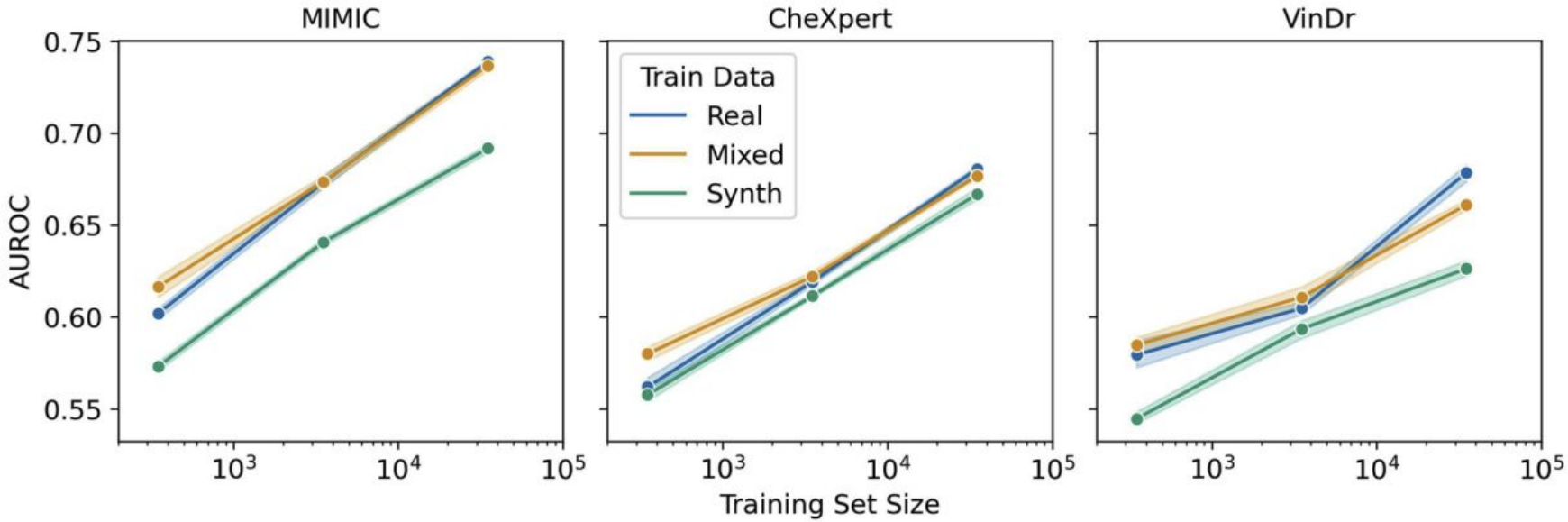
Combination of multiple abnormalities



Bluethgen et al. A vision–language foundation model for the generation of realistic chest X-ray images. Nature Biomedical Engineering 2024.

Benchmark classification performance using real vs mixed vs synthetic data

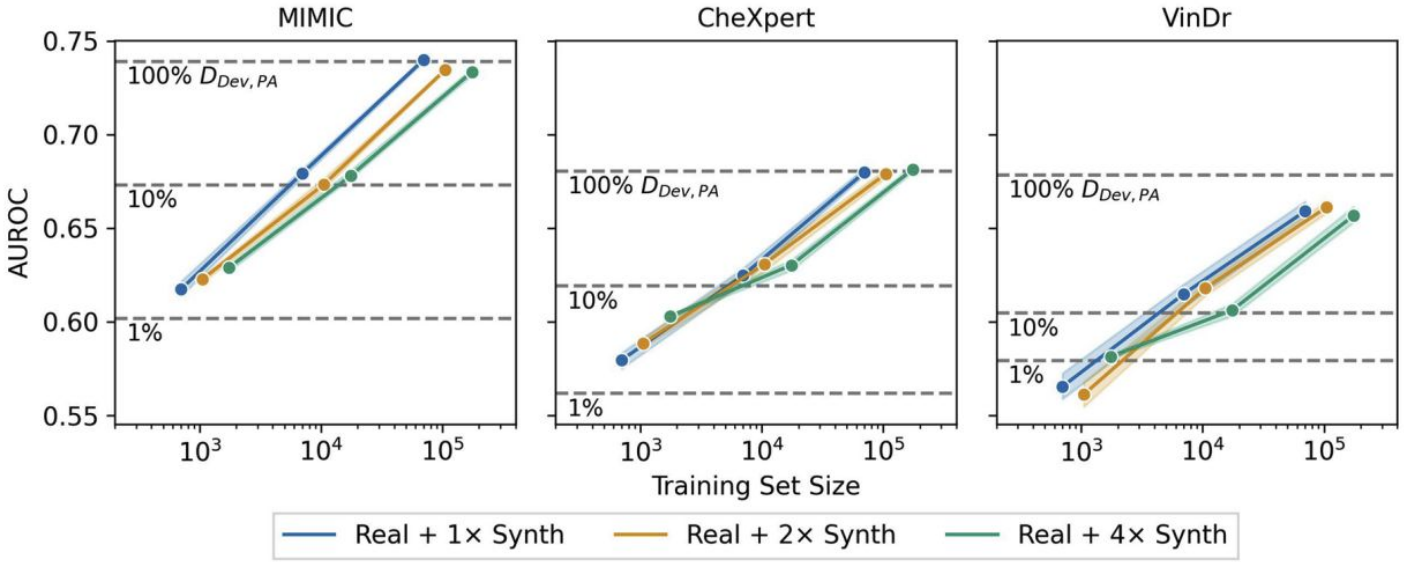
- Using DenseNet-121 as CNN backbone



Bluethgen et al. A vision–language foundation model for the generation of realistic chest X-ray images. Nature Biomedical Engineering 2024.

Benchmark classification performance by adding varying amounts of synthetic data to real data

- Using DenseNet-121 as CNN backbone



Bluethgen et al. A vision–language foundation model for the generation of realistic chest X-ray images. Nature Biomedical Engineering 2024.

Linear probing of models SSL-pretrained on real data

- Using DenseNet-121 as CNN backbone

Evaluation Data	Training Data (%)	Real CXR	Synthetic CXR	Mixed (1:1)
MIMIC	1	0.596 ± 0.006 (+0.3%)	0.557 ± 0.007 (-6.2%)	0.598 ± 0.006 (+0.8%)
	10	0.660 ± 0.001 (+2.8%)	0.614 ± 0.002 (-4.3%)	0.647 ± 0.001 (+0.8%)
	100	0.714 ± 0.001 (+1.1%)	0.650 ± 0.000 (-7.9%)	0.701 ± 0.000 (-0.7%)
CheXpert	1	0.534 ± 0.006 (-0.6%)	0.528 ± 0.008 (-1.7%)	0.527 ± 0.006 (-2.0%)
	10	0.590 ± 0.003 (-0.6%)	0.561 ± 0.002 (-5.4%)	0.579 ± 0.002 (-2.4%)
	100	0.647 ± 0.001 (+0.4%)	0.629 ± 0.001 (-2.3%)	0.652 ± 0.000 (+1.3%)
VinDr	1	0.567 ± 0.009 (+4.4%)	0.552 ± 0.009 (+1.7%)	0.550 ± 0.009 (+1.4%)
	10	0.564 ± 0.003 (+0.2%)	0.556 ± 0.003 (-1.2%)	0.557 ± 0.002 (-0.9%)
	100	0.603 ± 0.001 (-1.8%)	0.575 ± 0.001 (-6.5%)	0.601 ± 0.001 (-2.3%)

Bluethgen et al. A vision–language foundation model for the generation of realistic chest X-ray images. Nature Biomedical Engineering 2024.

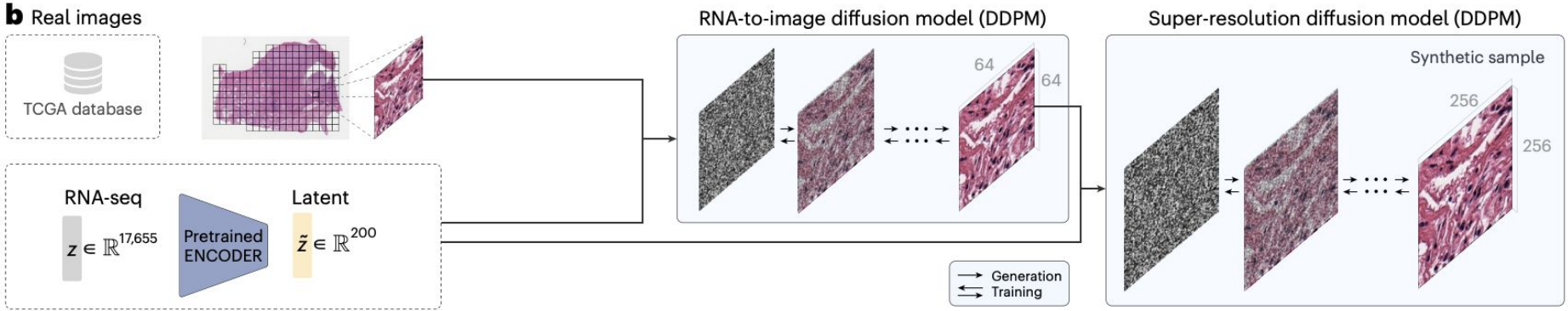
Linear probing of models SSL-pretrained on synthetic data

- Using DenseNet-121 as CNN backbone

Evaluation Data	Training Data (%)	Real CXR	Synthetic CXR	Mixed (1:1)
MIMIC	1	0.602 ± 0.006 (+1.0%)	0.550 ± 0.011 (-1.3%)	0.599 ± 0.005 (+0.2%)
	10	0.655 ± 0.002 (-0.8%)	0.609 ± 0.002 (-0.8%)	0.642 ± 0.001 (-0.8%)
	100	0.710 ± 0.001 (-0.5%)	0.650 ± 0.001 (±0.0%)	0.697 ± 0.000 (-0.6%)
CheXpert	1	0.534 ± 0.005 (+0.1%)	0.523 ± 0.008 (-0.9%)	0.533 ± 0.007 (+1.3%)
	10	0.581 ± 0.002 (-1.5%)	0.561 ± 0.002 (±0.0%)	0.578 ± 0.001 (-0.2%)
	100	0.642 ± 0.001 (-0.7%)	0.632 ± 0.001 (+0.5%)	0.640 ± 0.000 (-1.9%)
VinDr	1	0.546 ± 0.008 (-3.7%)	0.553 ± 0.013 (+0.2%)	0.551 ± 0.009 (+0.2%)
	10	0.558 ± 0.004 (-1.0%)	0.557 ± 0.003 (+0.2%)	0.558 ± 0.002 (+0.1%)
	100	0.617 ± 0.001 (+2.2%)	0.589 ± 0.001 (+2.5%)	0.614 ± 0.001 (+2.2%)

Bluethgen et al. A vision–language foundation model for the generation of realistic chest X-ray images. Nature Biomedical Engineering 2024.

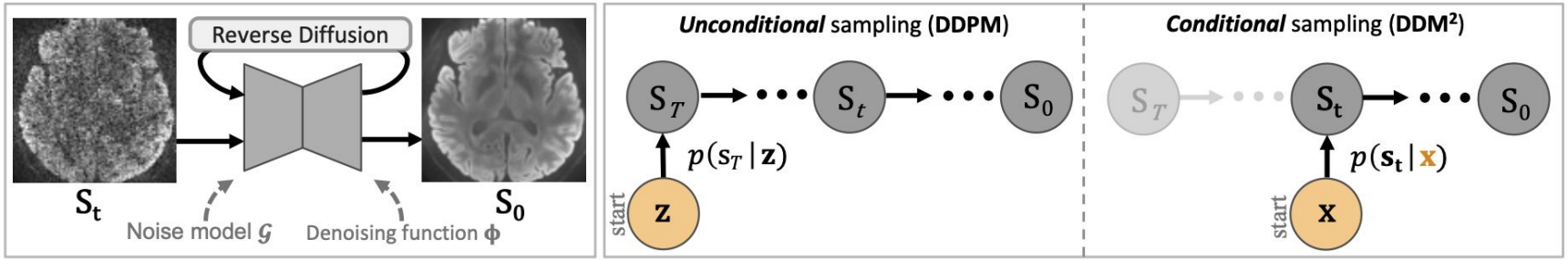
Preview to discussion paper presentations: generating pathology whole-slide image titles conditioned on RNA-sequencing



Carrillo-Perez et al. Generation of synthetic whole-slide image tiles of tumours from RNA-sequencing data via cascaded diffusion models. Nature Biomedical Engineering 2024.

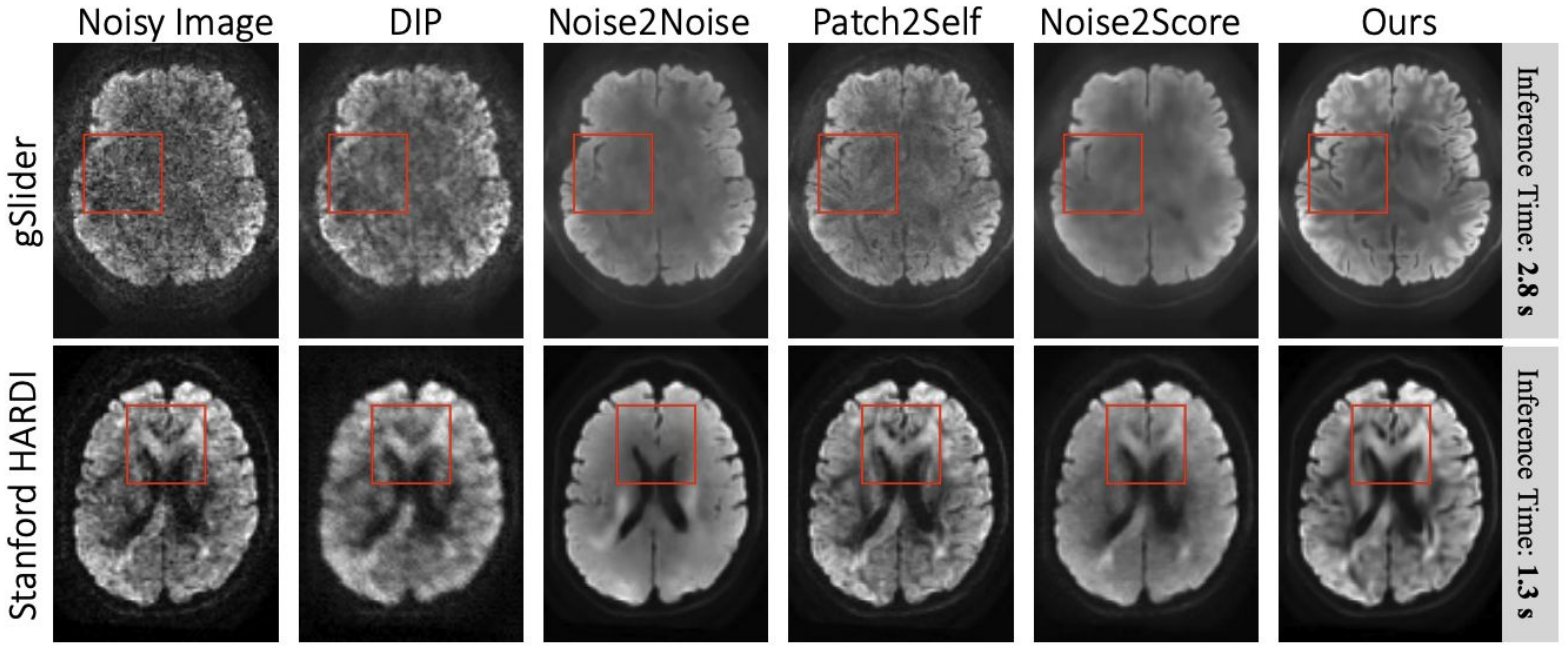
Diffusion models for MRI denoising: DDM^2

- MRI denoising is a key problem in modern MRI research due to the trade-off between achieving high signal-to-noise MRI scans and requiring long scan time (increased cost and discomfort, harder to accommodate overall patient demand)
- DDM^2 (Denoising Diffusion Models for Denoising Diffusion MRI) uses diffusion models to achieve this denoising. It conditionally samples an image generation based on a noisy image as condition, and matches to an intermediate timestep in the diffusion process.



Xiang et al. DDM2: Self-Supervised Diffusion MRI Denoising with Generative Diffusion Models. ICLR 2023.

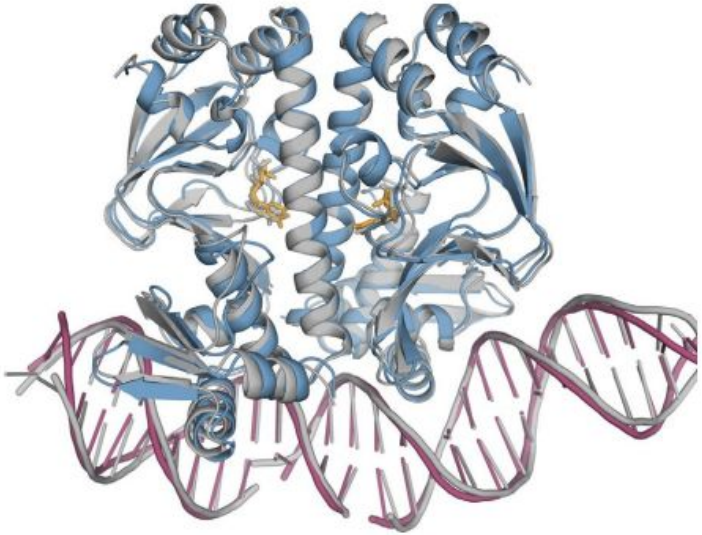
Diffusion models for MRI denoising: DDM^2

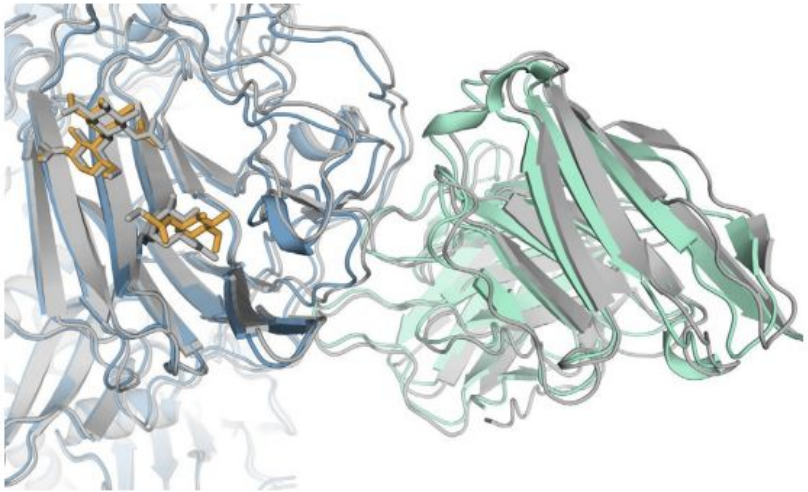


Xiang et al. DDM2: Self-Supervised Diffusion MRI Denoising with Generative Diffusion Models. ICLR 2023.

AlphaFold 3

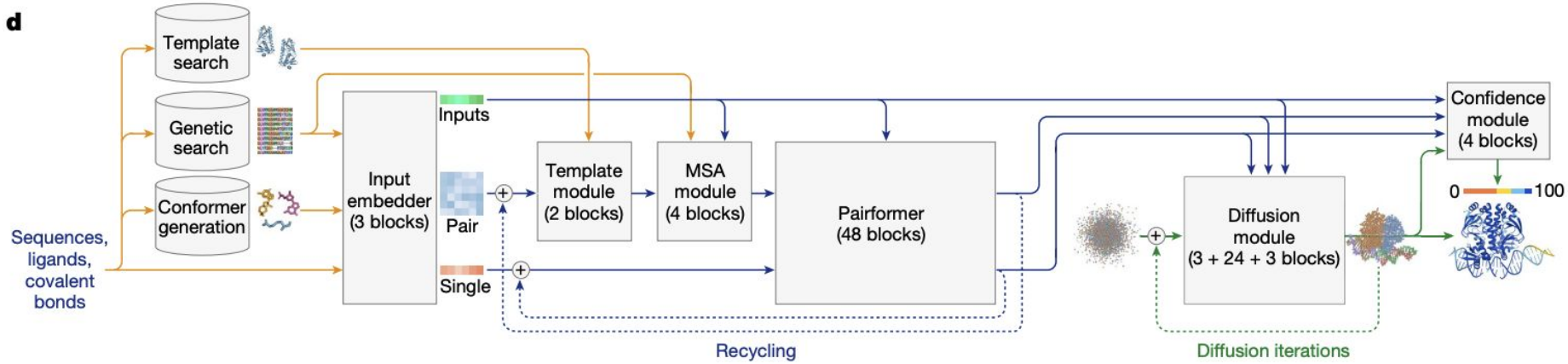
- Diffusion-based architecture that predicts the joint 3D structure of complexes including proteins, nucleic acids, small molecules, etc. from sequences





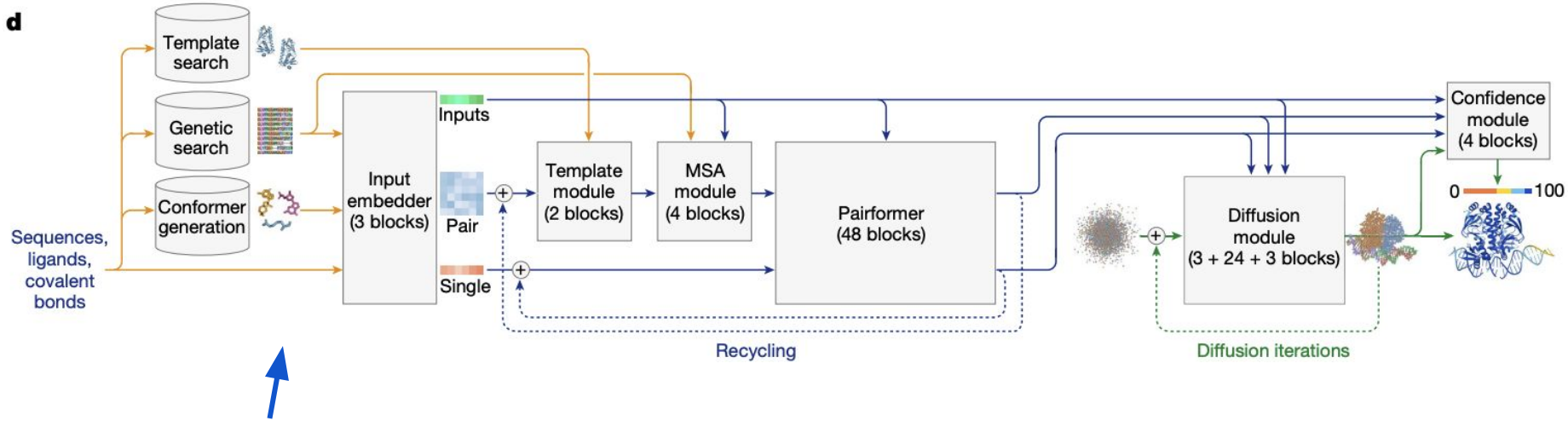
Abramson et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024.

AlphaFold 3: a bird's eye view



Abramson et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024.

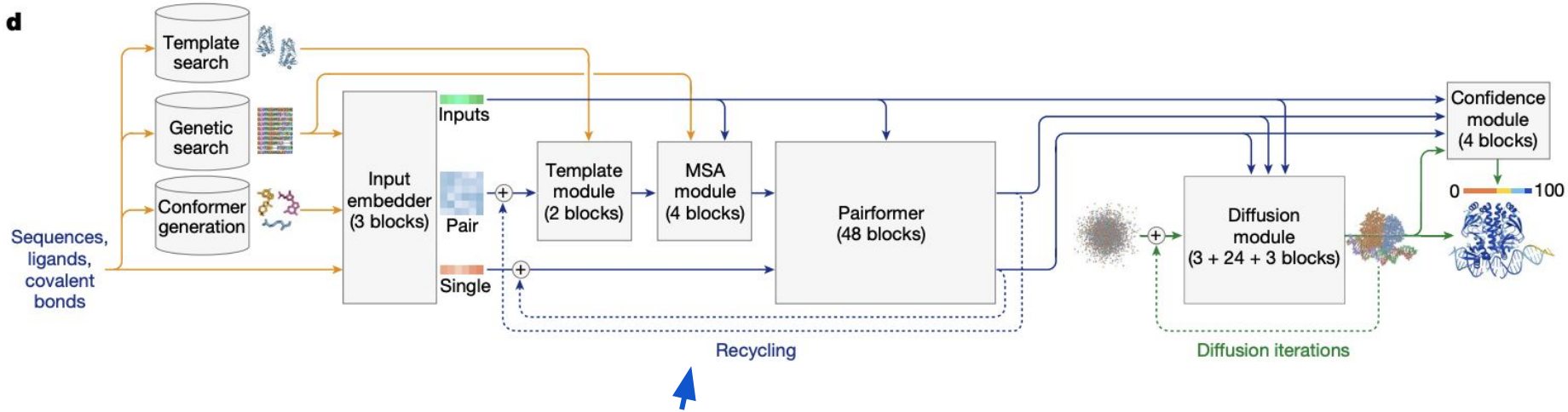
AlphaFold 3: a bird's eye view



Process and initially embed inputs

Abramson et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024.

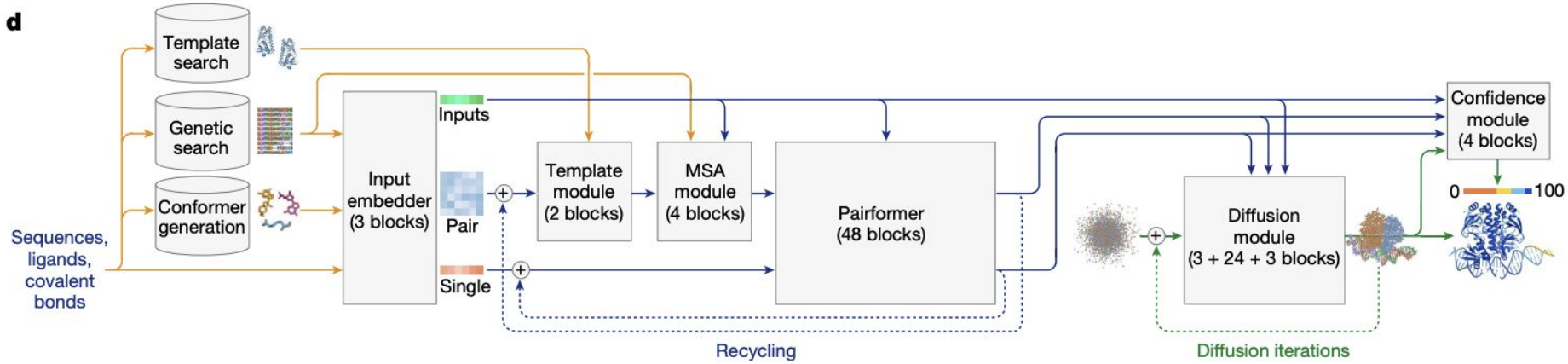
AlphaFold 3: a bird's eye view



Transform inputs into powerful representations to condition diffusion model

Abramson et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024.

AlphaFold 3: a bird's eye view

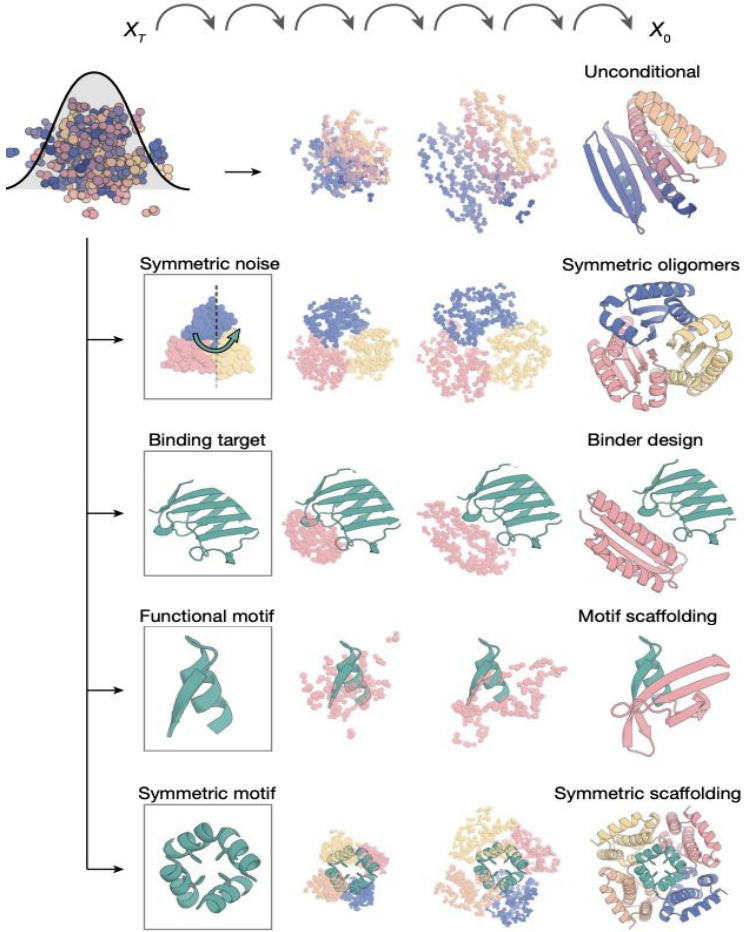


Starting from random atom coordinates (noise), use conditional diffusion model to generate refined coordinate values (i.e. 3D structure)

Abramson et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024.

RFdiffusion: designing protein structures based on functional constraints (conditions)

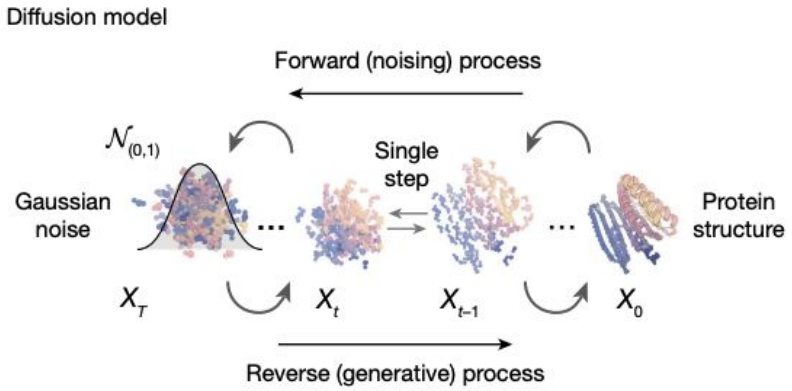
Generated output is still 3D protein structure (similar to AlphaFold 2, not the more diverse outputs of AlphaFold 3), but conditioning is now based on embedding functional conditions such as a desired binding target, functional motif, etc.



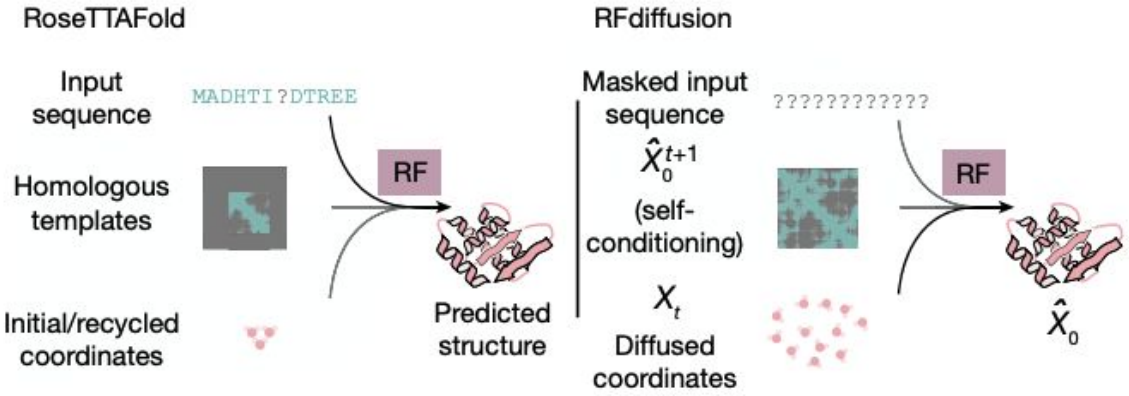
Watson et al. De novo design of protein structure and function with RFdiffusion. Nature 2023.

RFdiffusion: designing protein structures based on functional constraints (conditions)

Core diffusion model looks familiar: generating a valid 3D protein structure from a starting noise structure



Compared to RosaTTAFold (the authors' previous work performing sequence-based prediction similar to AlphaFold), RFdiffusion no longer uses an input sequence as input



Watson et al. De novo design of protein structure and function with RFdiffusion. Nature 2023.

Next time

- Vision-Language Generative Models