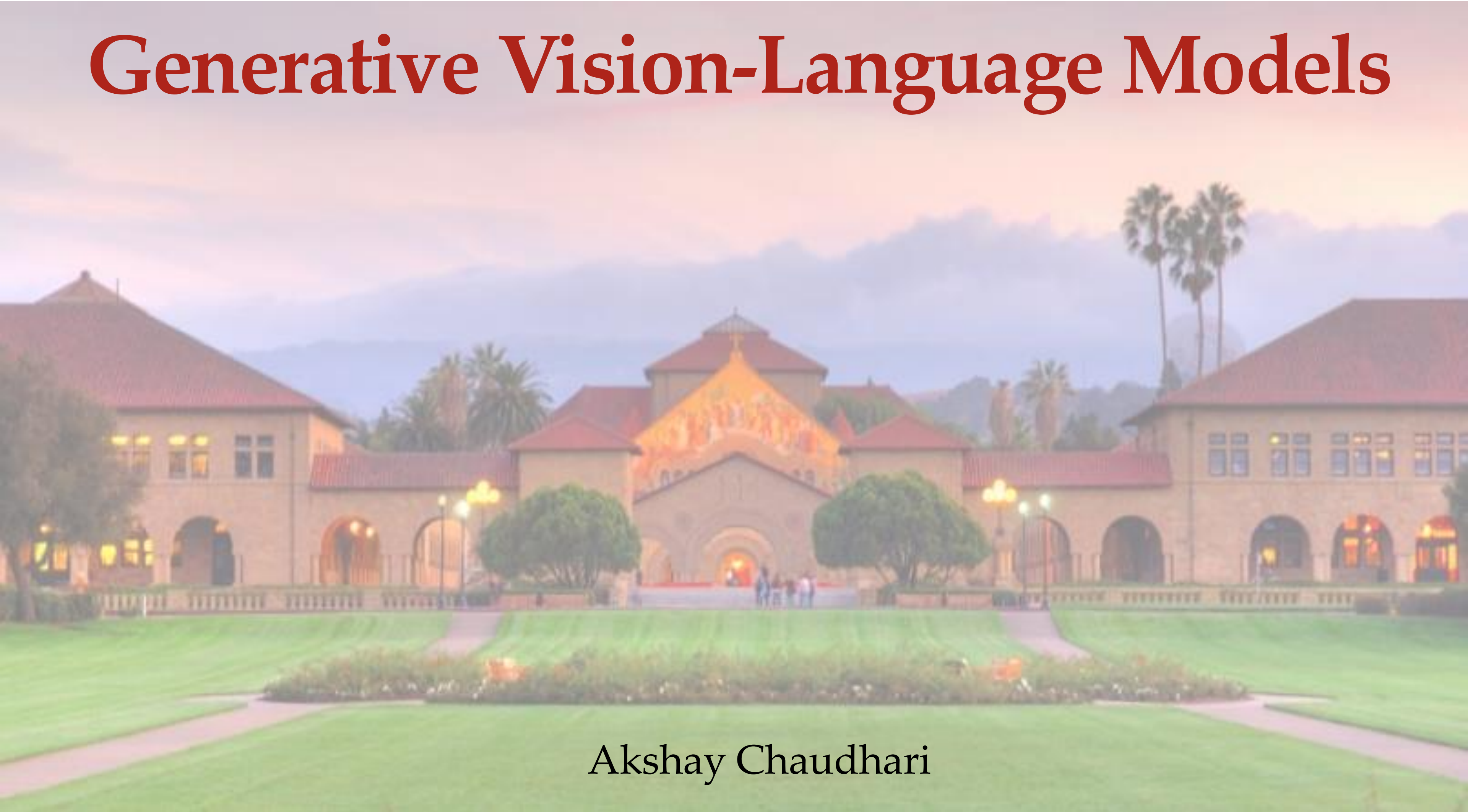
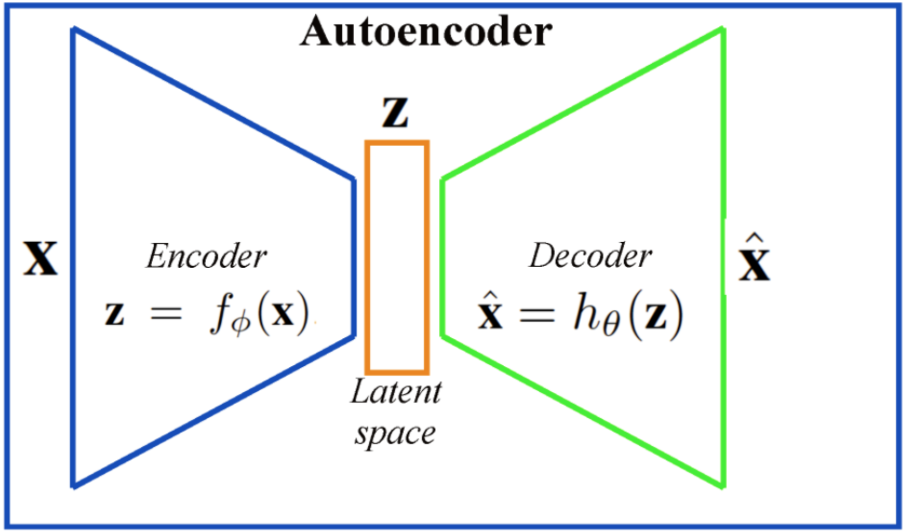
Generative Vision-Language Models



Akshay Chaudhari

Generative Models – Autoencoders

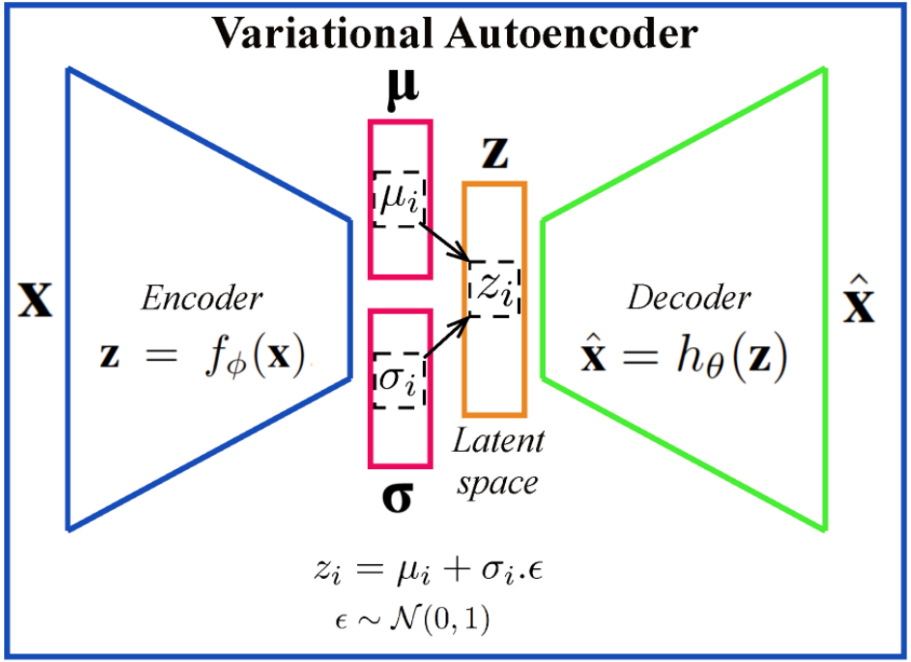
- Encoder – Reduce the dimensionality of some input
- Decoder – Use condensed dimensionality to recover signal



[1] Zemouri R. et al "Semi-supervised adversarial variational autoencoder.. Machine Learning and Knowledge Extraction (2020)

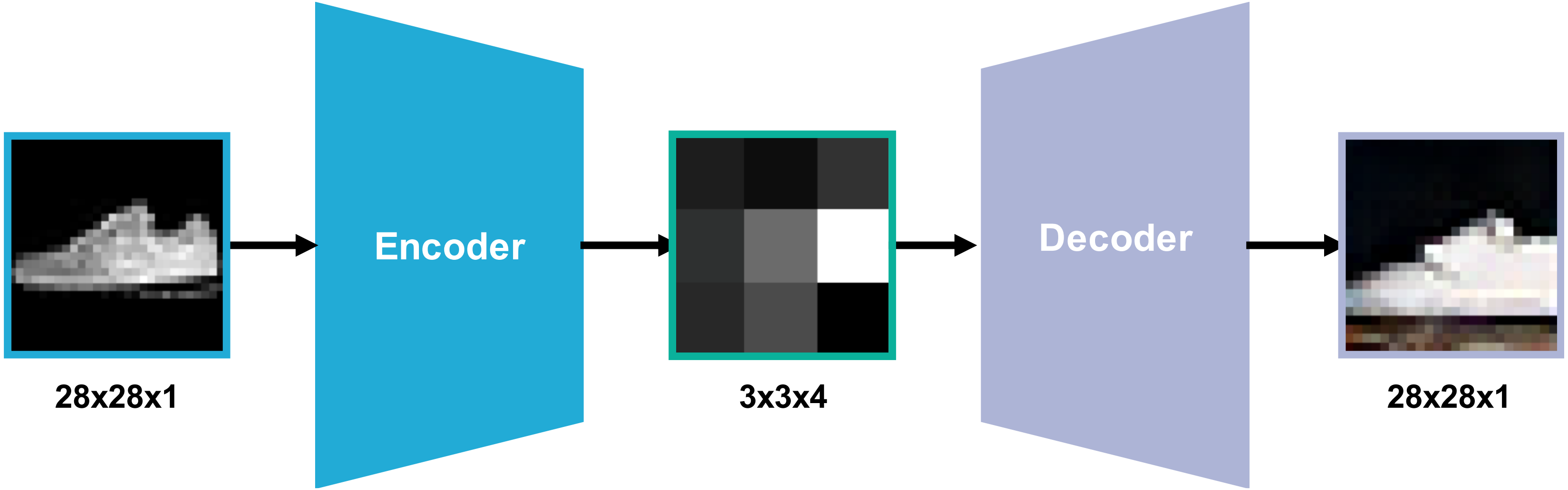
Variational Autoencoder

- Sample plausible distributions from latent space

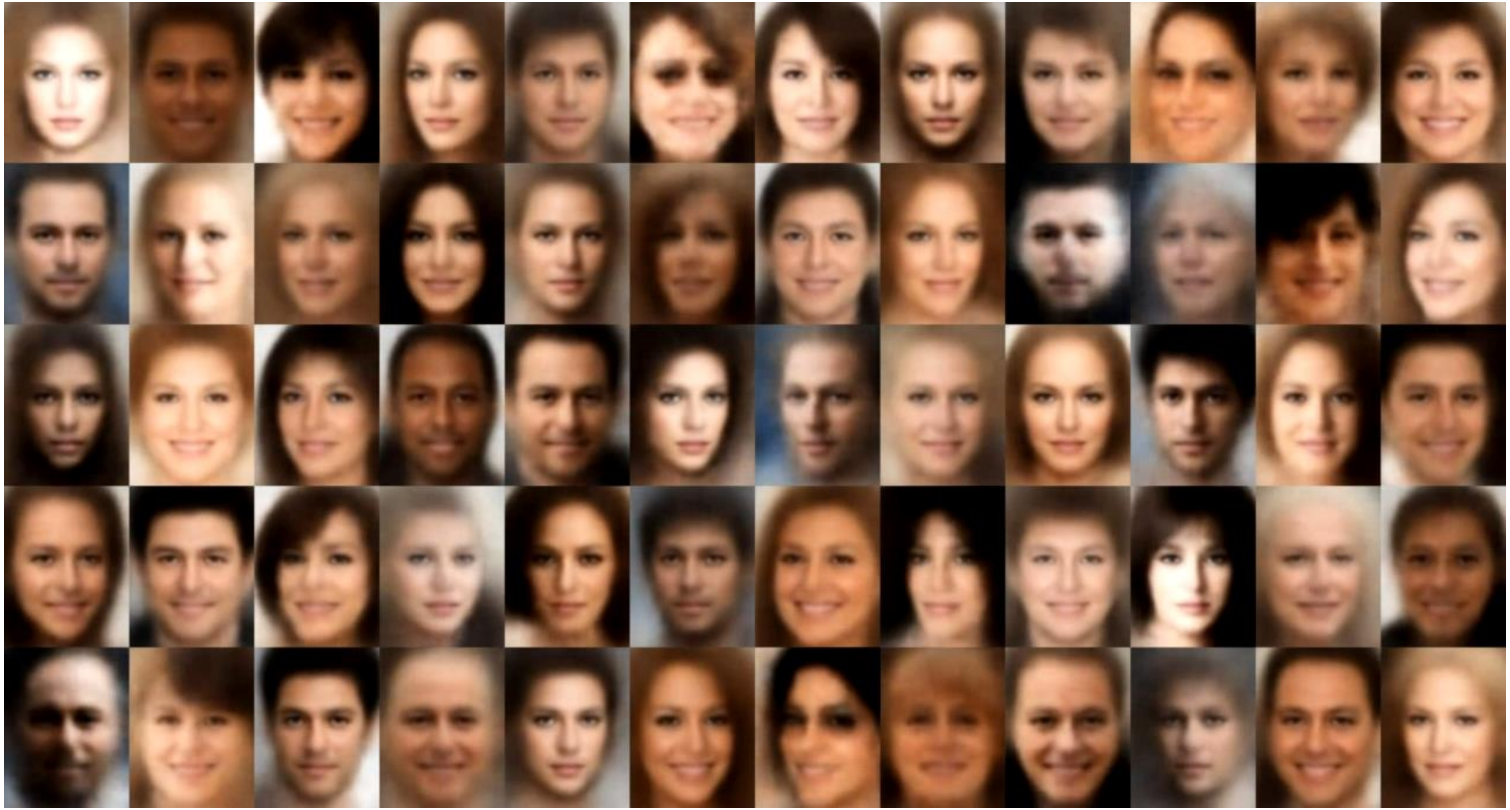

Variational Autoencoder
μ
μ_i
z
z_i
σ_i
σ
X
Encoder
$\mathbf{z} = f_\phi(\mathbf{x})$
Latent
space
Decoder
$\hat{\mathbf{x}} = h_\theta(\mathbf{z})$
$\hat{\mathbf{X}}$
$z_i = \mu_i + \sigma_i.\epsilon$
$\epsilon \sim \mathcal{N}(0,1)$

[1] Zemouri R. et al "Semi-supervised adversarial variational autoencoder.. Machine Learning and Knowledge Extraction (2020)

VAEs for Natural Images



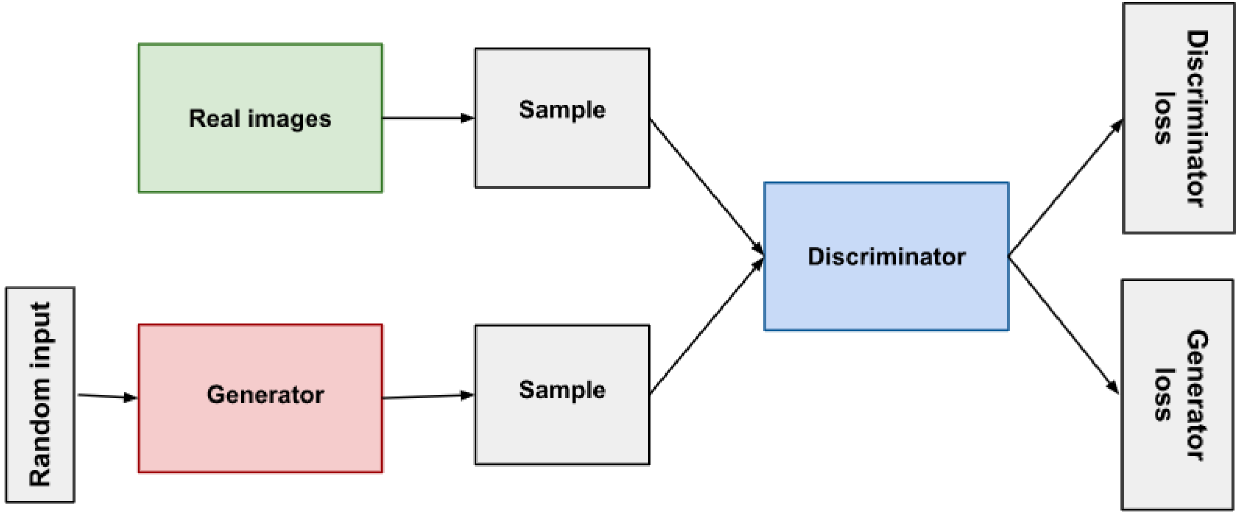
Variational Autoencoder



[1] Variational Autoencoder. https://github.com/wojciechmo/vae

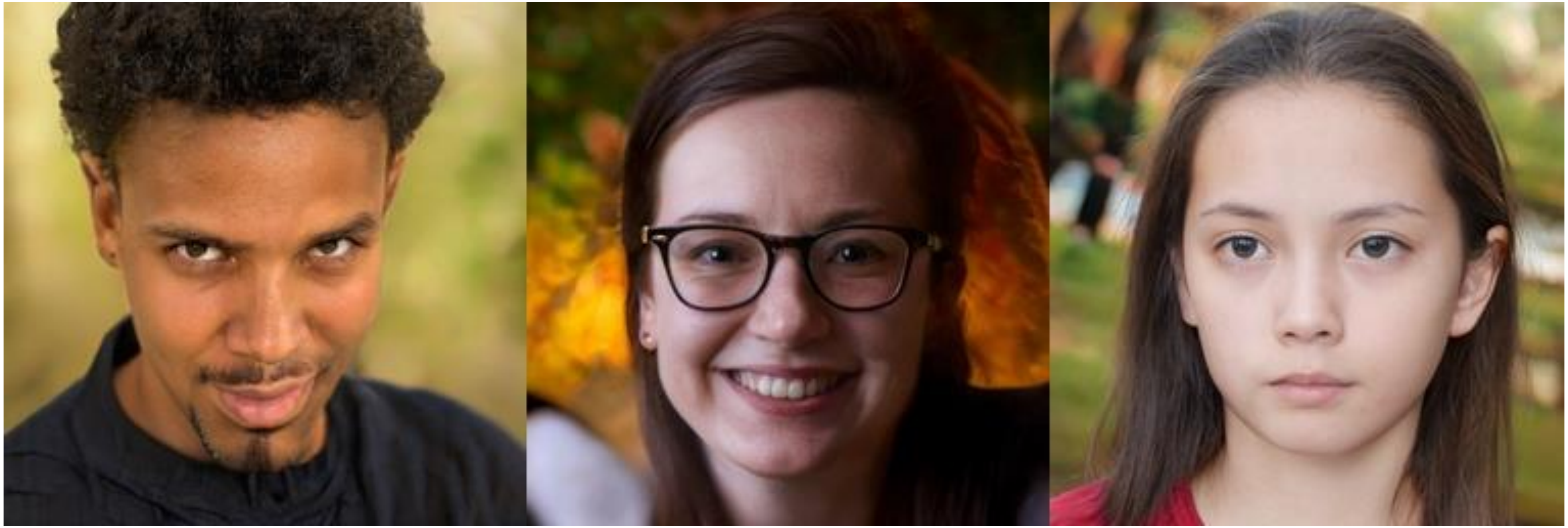
Generative Adversarial Network

- Use an additional *discriminator* network to enforce semantic consistency of *generator* images



[1] Goodfellow I. et al ”Generative Adversarial Nets”. NeuRIPS (2014)
[2] Overoview of GAN Structure (https://developers.google.com/machine-learning/gan/gan_structure)

StyleGAN2



[1] Kerras T et al. “Analyzing and Improving the Image Quality of StyleGAN”. CVPR (2020). https://github.com/NVlabs/stylegan2

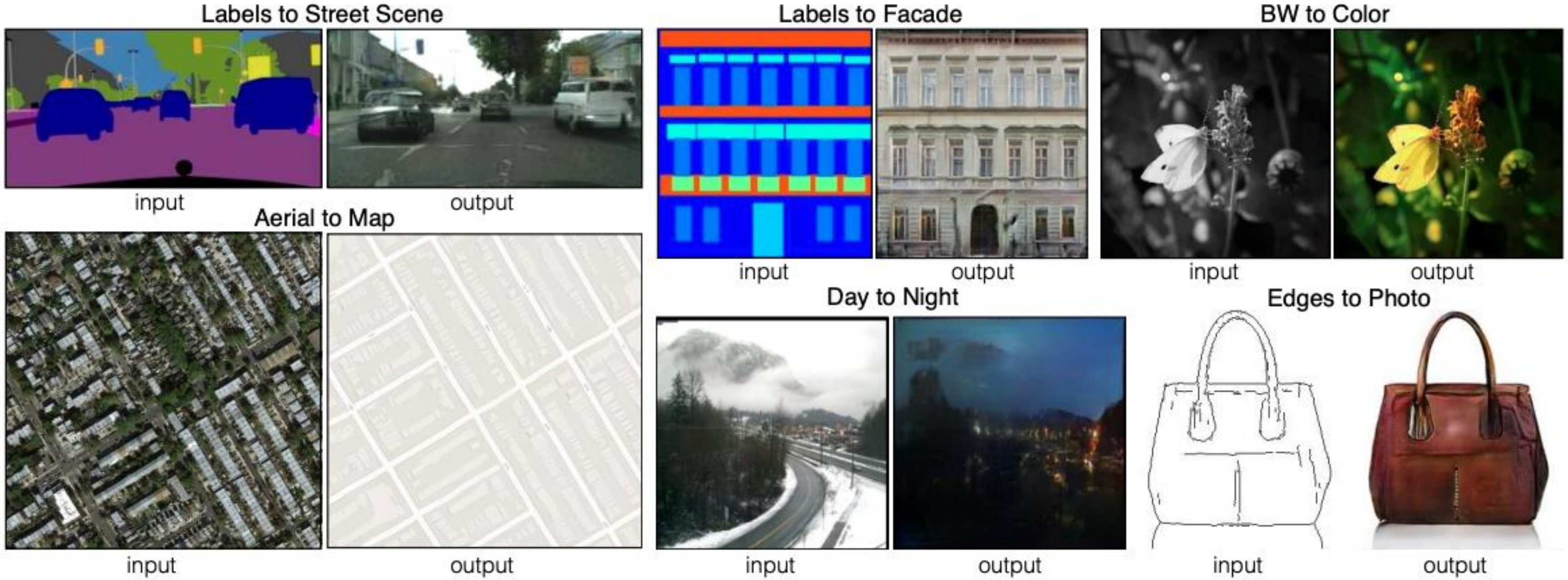
Conditional GANs

- Create generative models conditioned on input features
- Example: Generate an image with a specific class

[1] Brock A. et al "Large scale GAN training for high fidelity natural image synthesis." ICLR (2019)

Conditional GANs

- Create generative models conditioned on segmentation masks



[1] Isola P. et al "Image-to-Image Translation with Conditional Adversarial Networks." CVPR (2017)

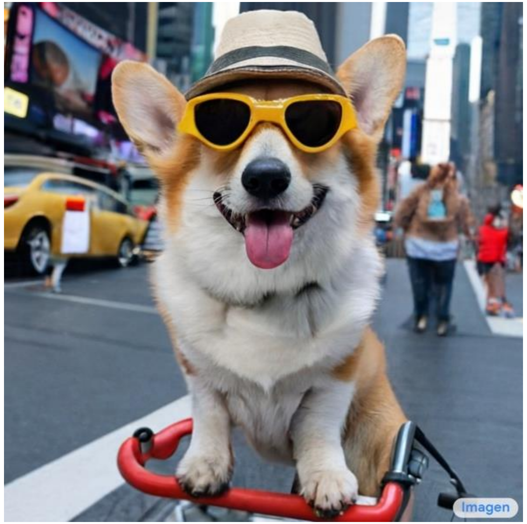
Denoising Diffusion Models

X^0

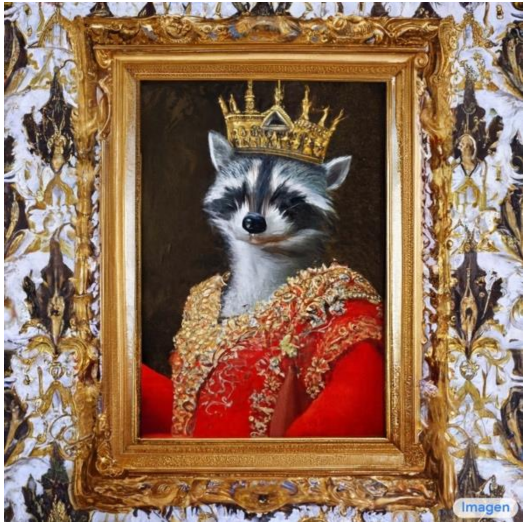
Song et al. Score-Based Generative Modeling through Stochastic Differential Equations. ICLR 2021

Text-to-Image Generative Models

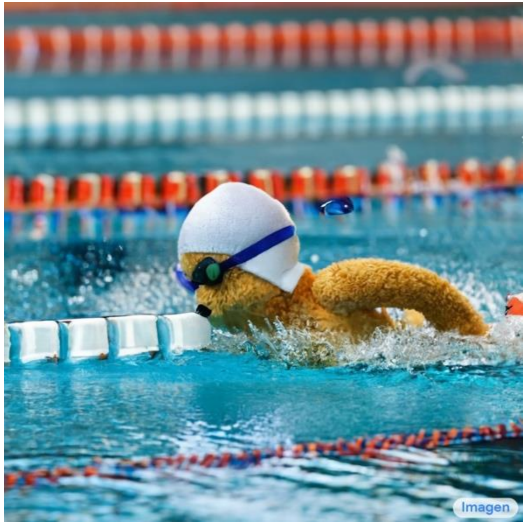
A photo of a corgi wearing a hat in Times Square. It is wearing sunglasses and a beach hat.



A majestic oil painting of a raccoon Queen wearing red French royal gown. The painting is hanging on an ornate wall decorated with wallpaper.



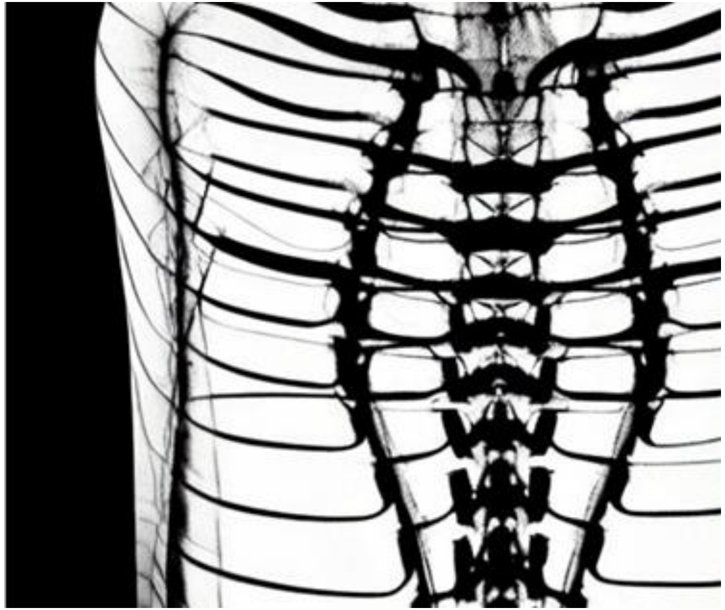
Teddy bears swimming at the Olympics 400m Butterfly event



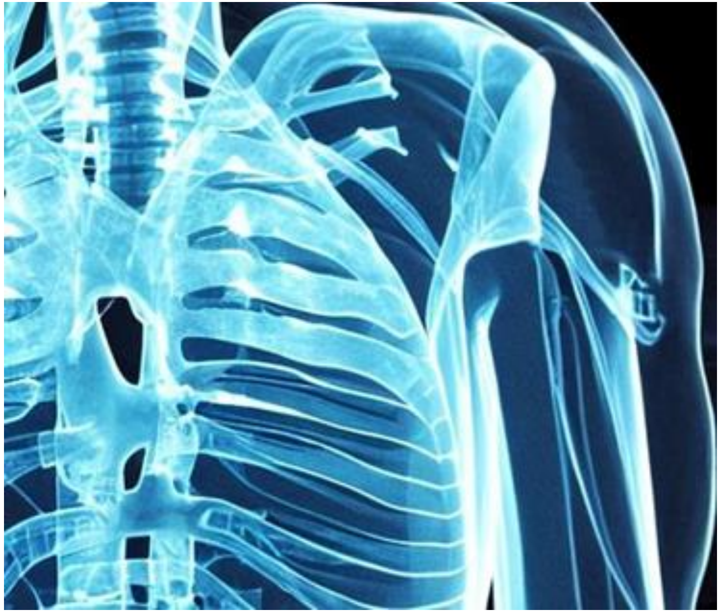
Saharia et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. NeurIPS 2022.

Text-to-Image Generative Models

A photo of a lung x-ray



A photo of a lung x-ray
with visible pleural effusion



High-resolution Image Synthesis with Latent Diffusion Models. Rombach et al. 2021

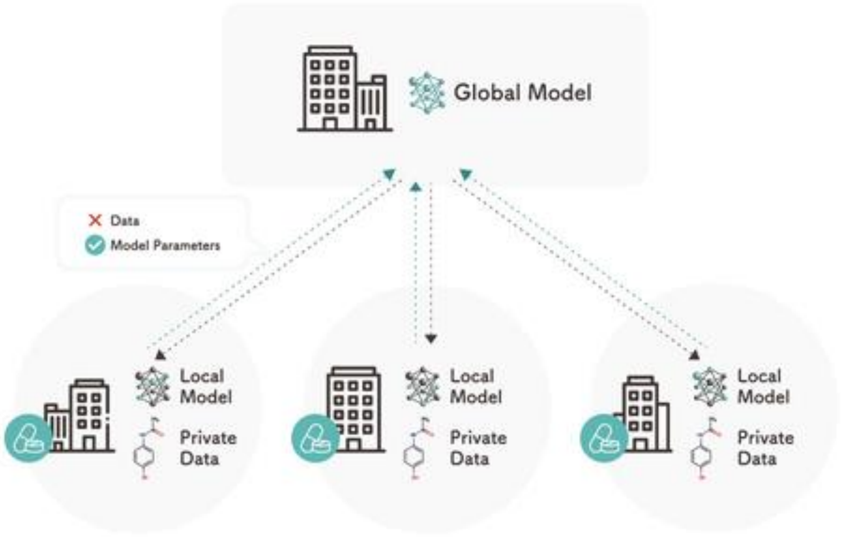
Why Use Synthetic Medical Images?

- Data augmentation
  - Scarcity of data
  - Rare diseases
  - Mitigating Bias

Why Use Synthetic Medical Images?

- Data augmentation
  - Scarcity of data
  - Rare diseases
  - Mitigating Bias

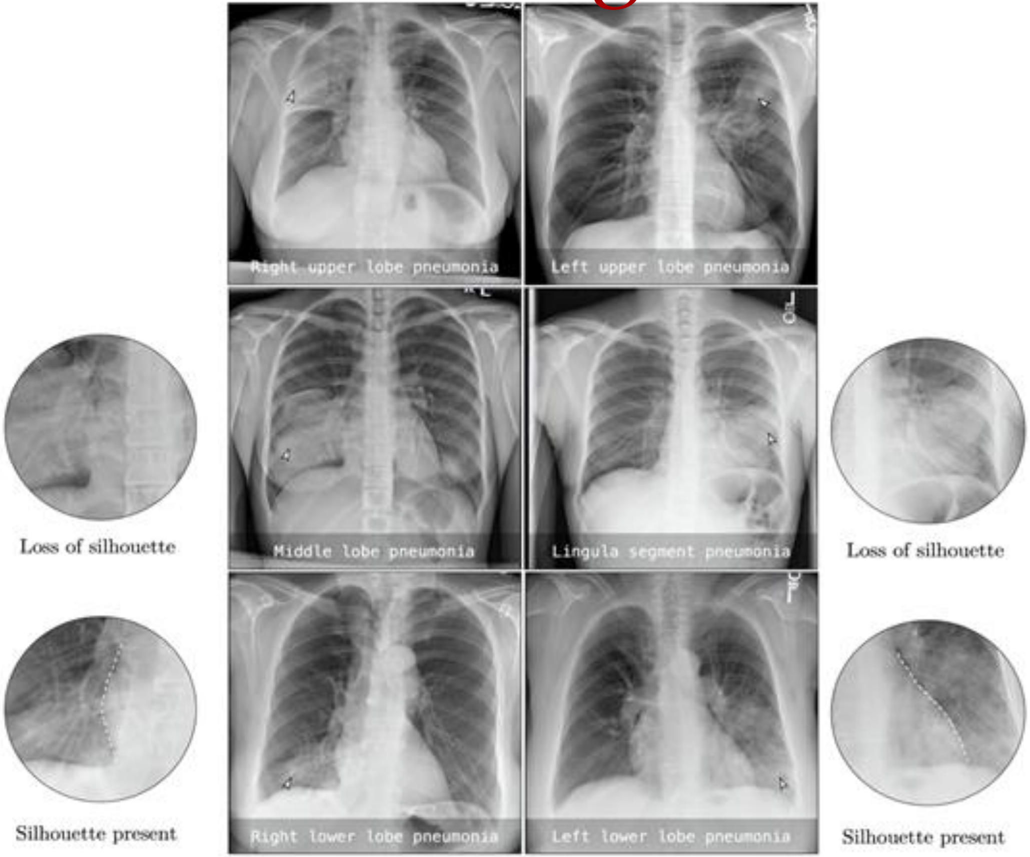
- Privacy preservation



BusinessWire. kMoL, a Machine Learning Library for AI Drug Discovery With Federated Learning Capabilities. 2022.

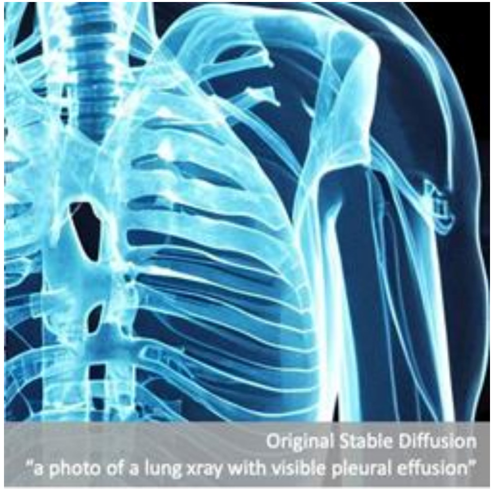
Why Use Synthetic Medical Images?

- Data augmentation
  - Scarcity of Data
  - Rare diseases
  - Mitigating Bias
- Privacy preservation
- Medical Education

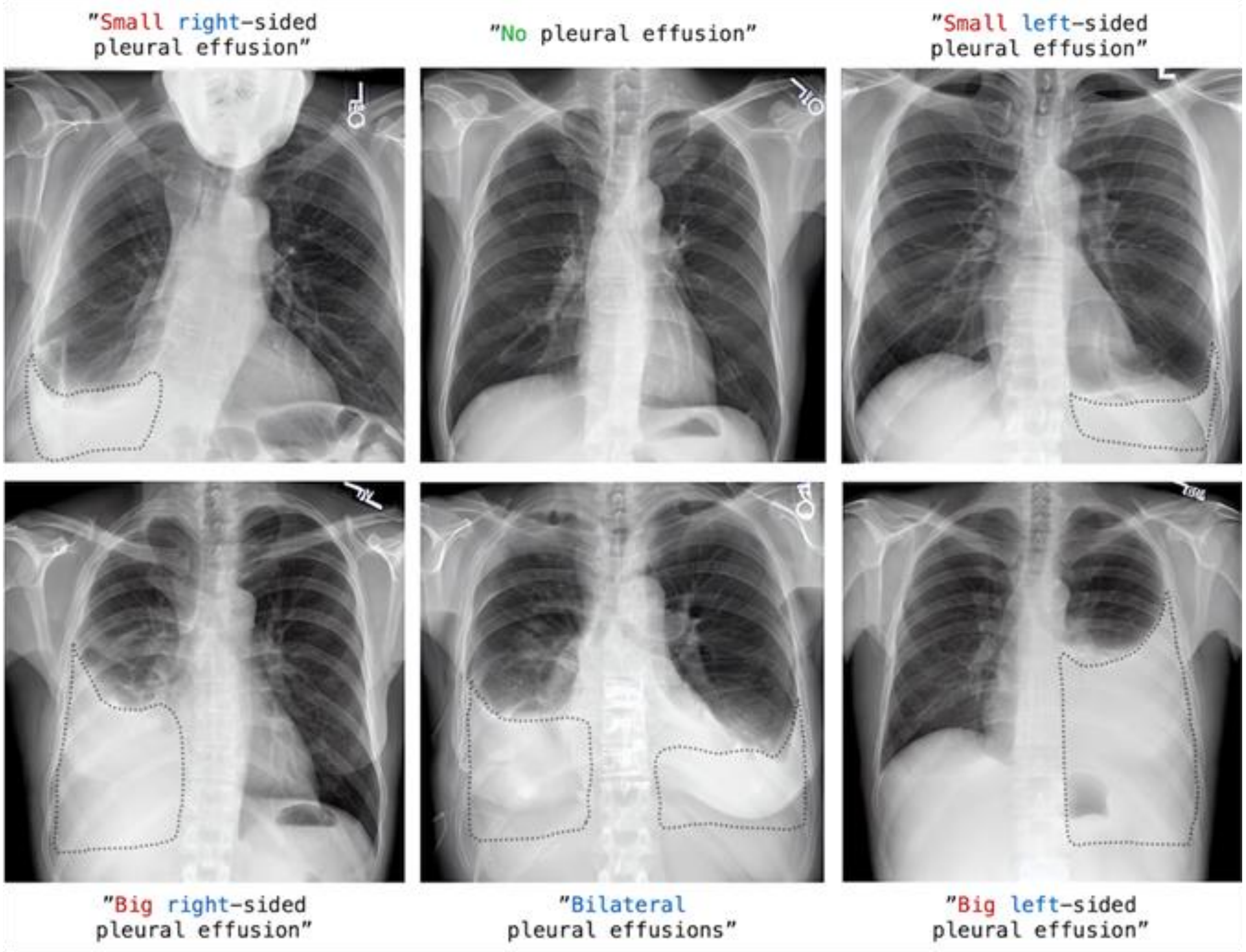


Right upper lobe pneumonia
Left upper lobe pneumonia
Loss of silhouette
Middle lobe pneumonia
Lingula segment pneumonia
Loss of silhouette
Silhouette present
Right lower lobe pneumonia
Left lower lobe pneumonia
Silhouette present

Chambon & Bluethgen et al. RoentGen: Vision-Language Foundation Model for Chest X-ray Generation. (arXiv preprint) 2022



Original Stable Diffusion
"a photo of a lung xray with visible pleural effusion"



"Small right-sided pleural effusion"

"No pleural effusion"

"Small left-sided pleural effusion"

"Big right-sided pleural effusion"

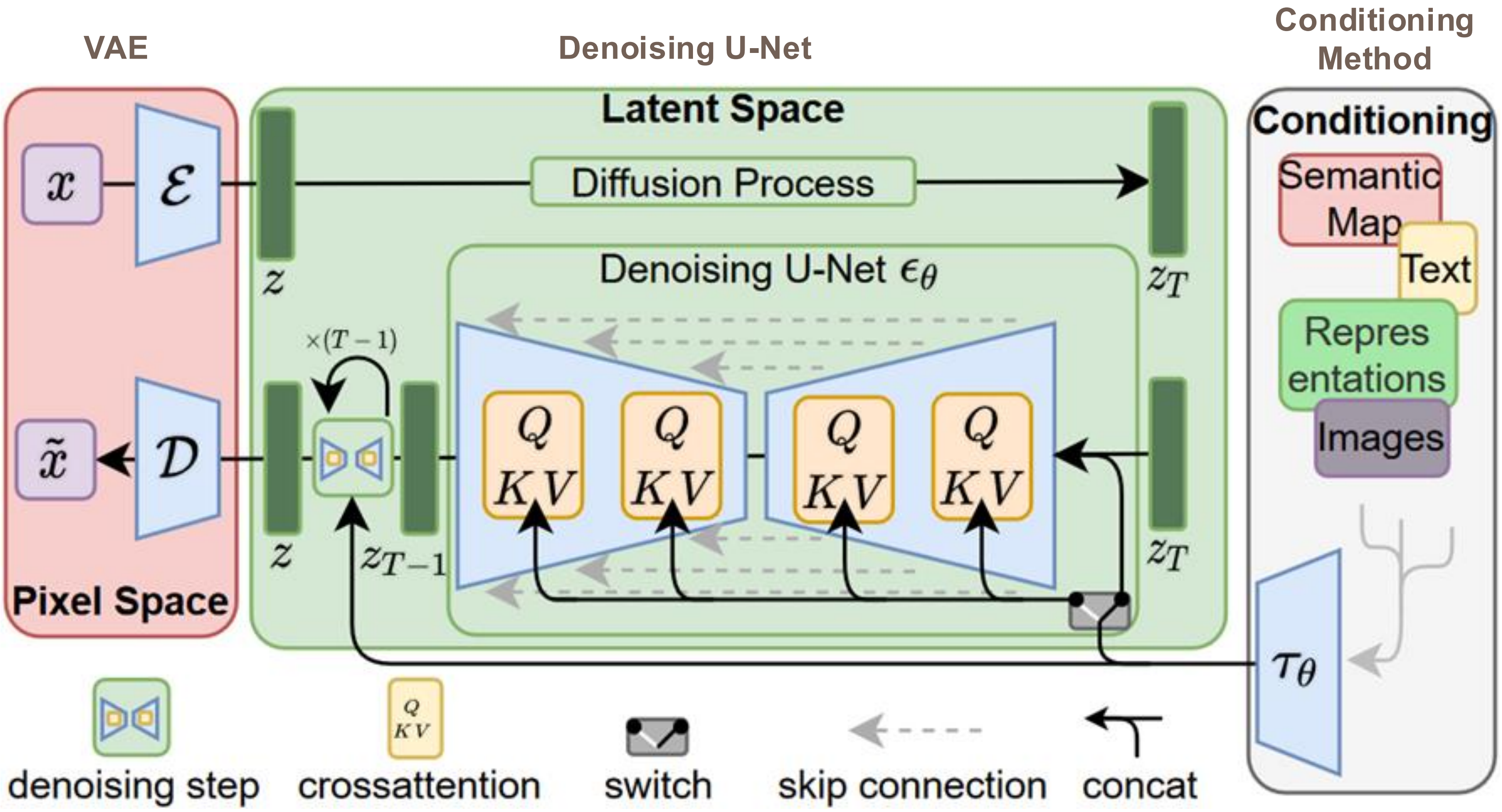
"Bilateral pleural effusions"

"Big left-sided pleural effusion"

Chris
Bluethgen

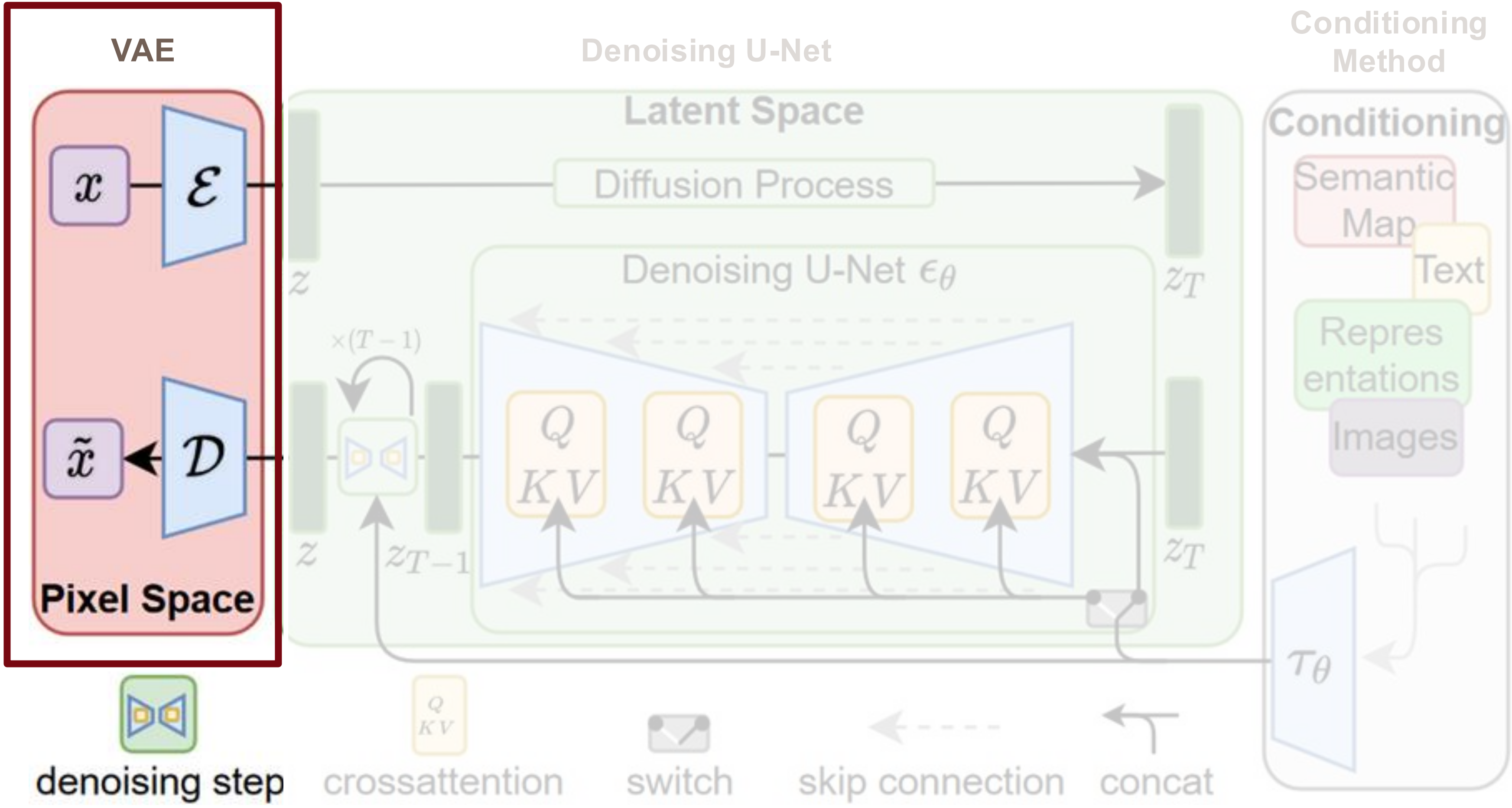
Pierre
Chambon

RoentGen: Vision-Language Foundation Model for Chest X-ray Generation. Chambon & Bluethgen et al. (arXiv preprint) 2022



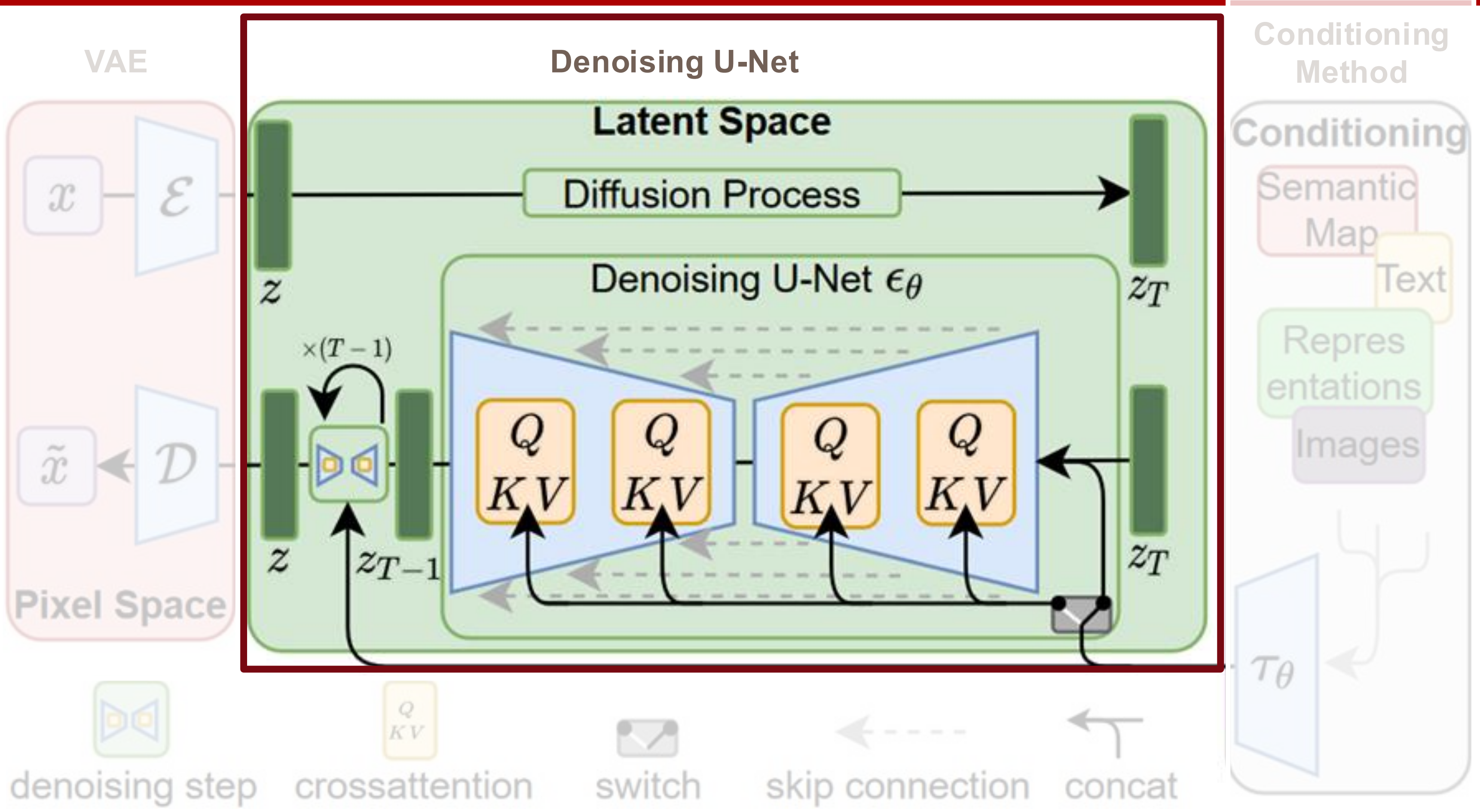
Stable Diffusion

High-resolution Image Synthesis with Latent Diffusion Models. Rombach et al. 2021



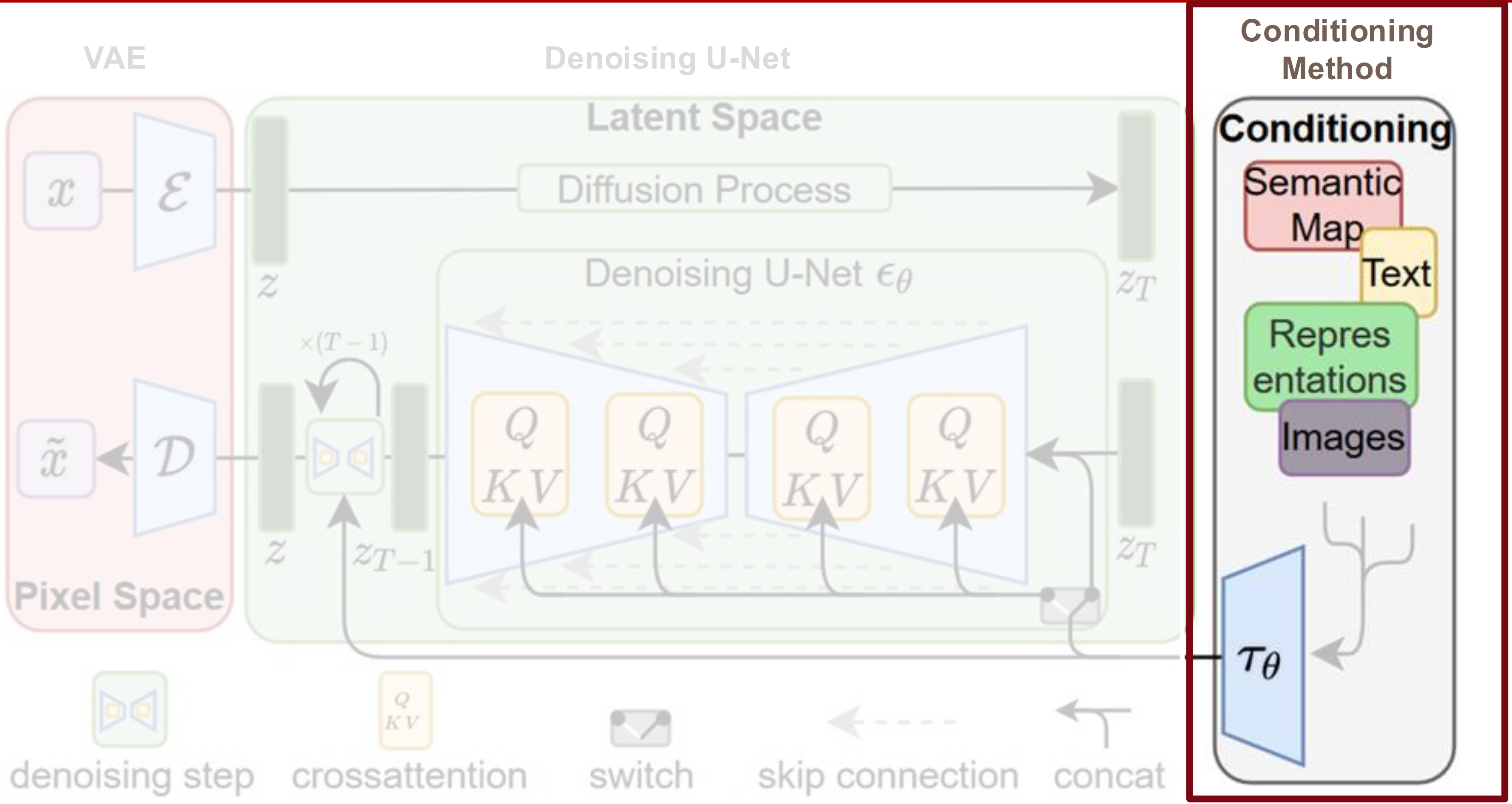
Stable Diffusion

High-resolution Image Synthesis with Latent Diffusion Models. Rombach et al. 2021



Stable Diffusion

High-resolution Image Synthesis with Latent Diffusion Models. Rombach et al. 2021



Stable Diffusion

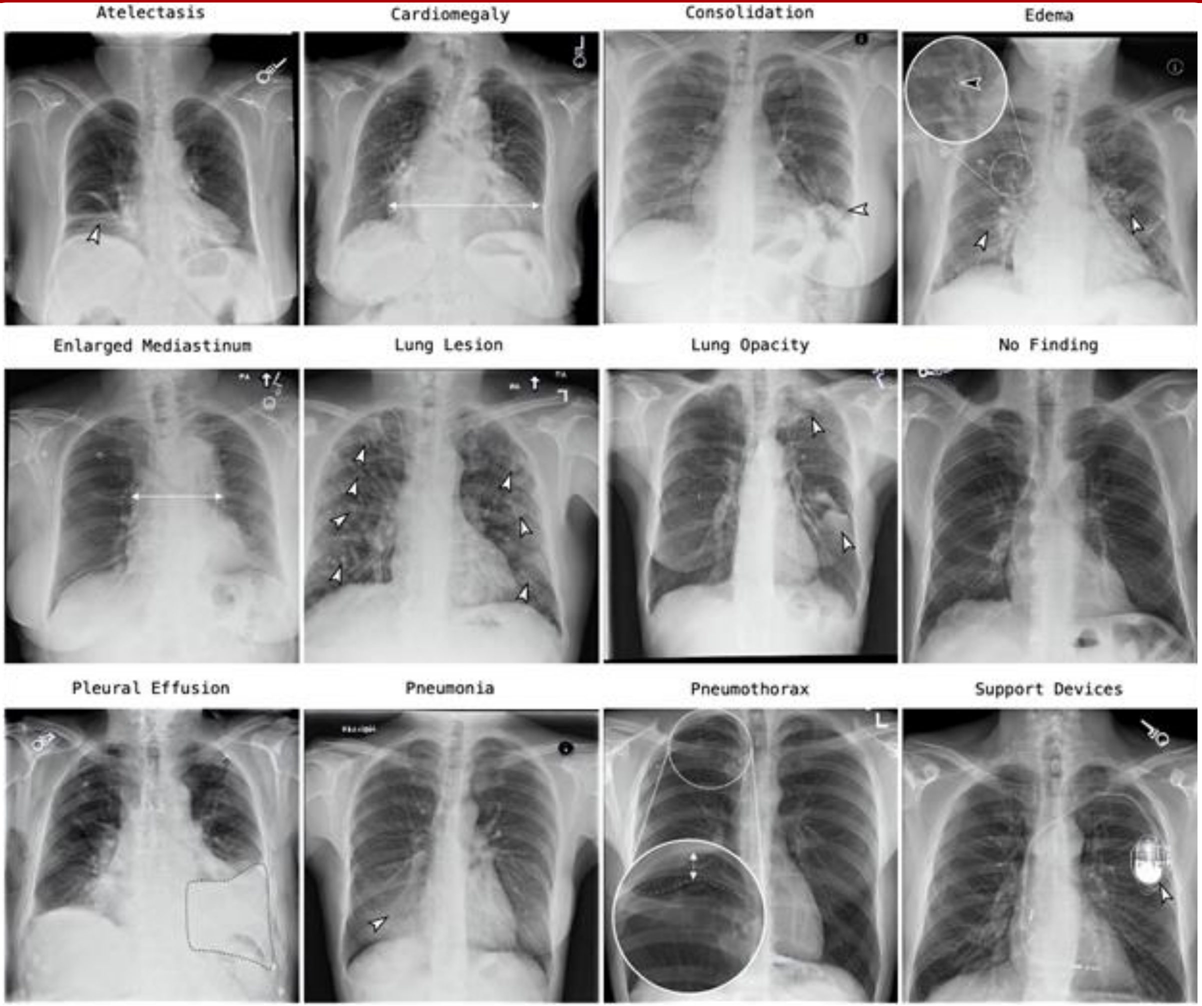
High-resolution Image Synthesis with Latent Diffusion Models. Rombach et al. 2021

Qualities of (Good) Generative Models

- Fidelity
  - Samples should resemble real data
- Diversity
  - Cover the variety of real data and produce diverse outputs
- Authenticity
  - Samples should not just be a copy of initial data

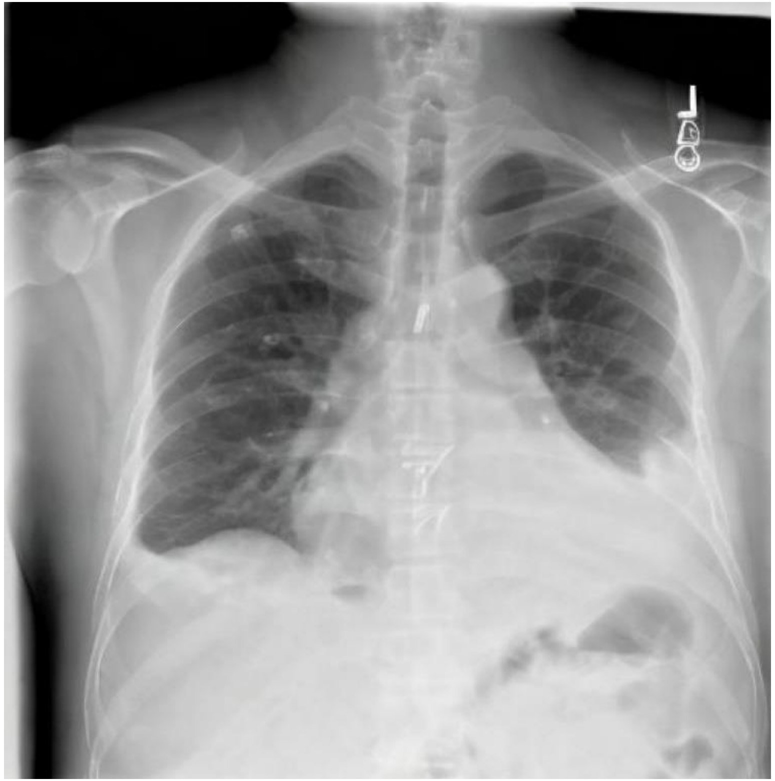
Alaa et al. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models. ICML 2022

Covering Range of Findings



Bluethgen*, Chambon*, et al "RoentGen: Vision-Language Foundation Model for Chest X-ray Generation". Nature Biomedical Engineering 2024

RoentGen-v2 Model



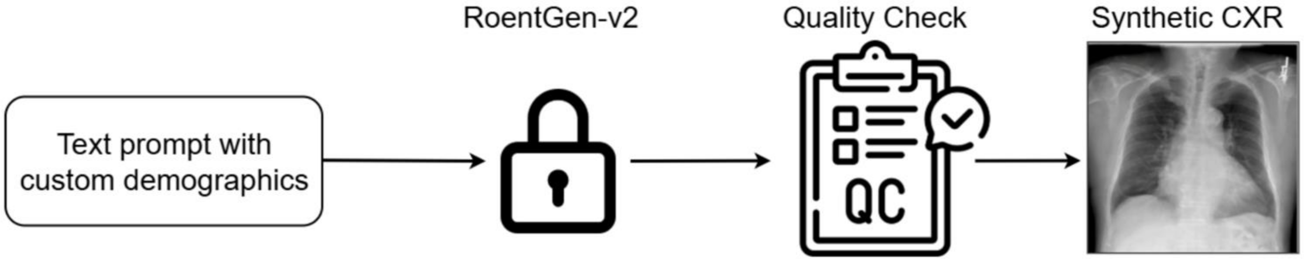
- Text to image model
- 500k+ images created

```
Bilateral , moderate right and small left,
pleural effusions with overlying atelectasis.
Medial right base opacity may be due to
combination of pleural effusion and
atelectasis , but consolidation due to
infection is not excluded in the appropriate
clinical setting.
```

48 year old ASIAN male.

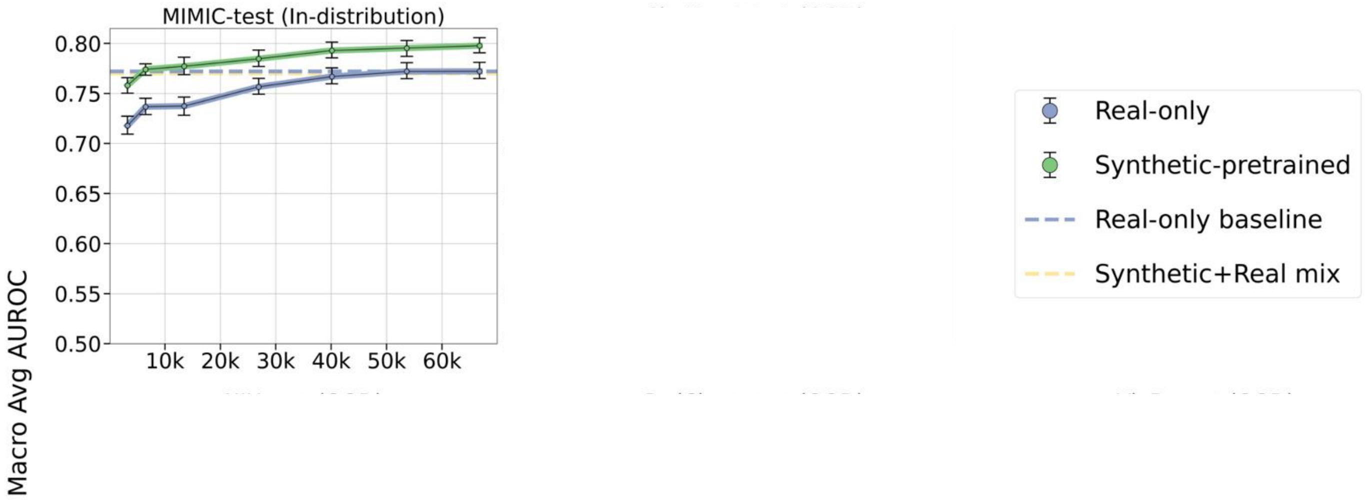
Moroianu et al. Improving Performance, Robustness, and Fairness of Radiographic AI Models with Finely-Controllable Synthetic Data. arXiv (2025).

Filtering Synthetic Data for Quality



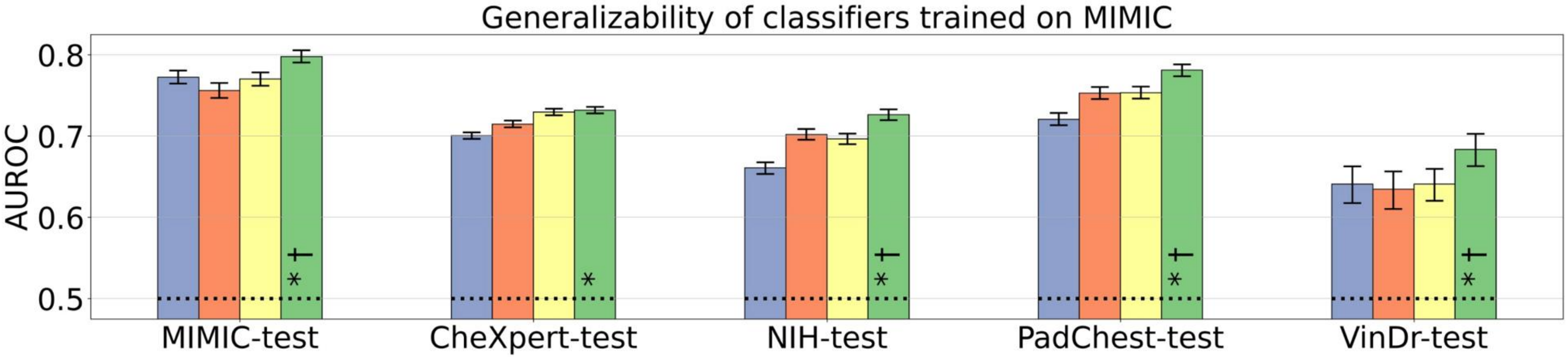
Moroianu et al. Improving Performance, Robustness, and Fairness of Radiographic AI Models with Finely-Controllable Synthetic Data. arXiv (2025).

Synthetic Data Improves Accuracy



Moroianu et al. Improving Performance, Robustness, and Fairness of Radiographic AI Models with Finely-Controllable Synthetic Data. arXiv (2025).

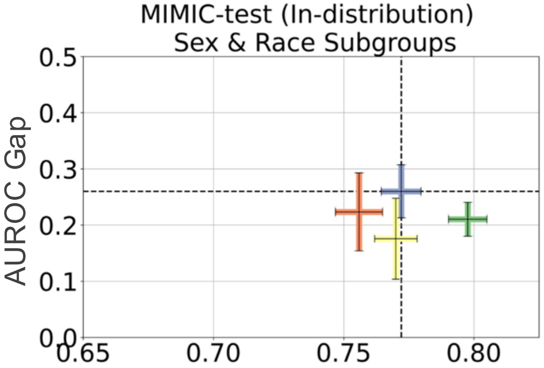
Synthetic Data Benefit



Generalizability of classifiers trained on MIMIC

Moroianu et al. Improving Performance, Robustness, and Fairness of Radiographic AI Models with Finely-Controllable Synthetic Data. arXiv (2025).

Synthetic Data Improves Fairness



Moroianu et al. Improving Performance, Robustness, and Fairness of Radiographic AI Models with Finely-Controllable Synthetic Data. arXiv (2025).

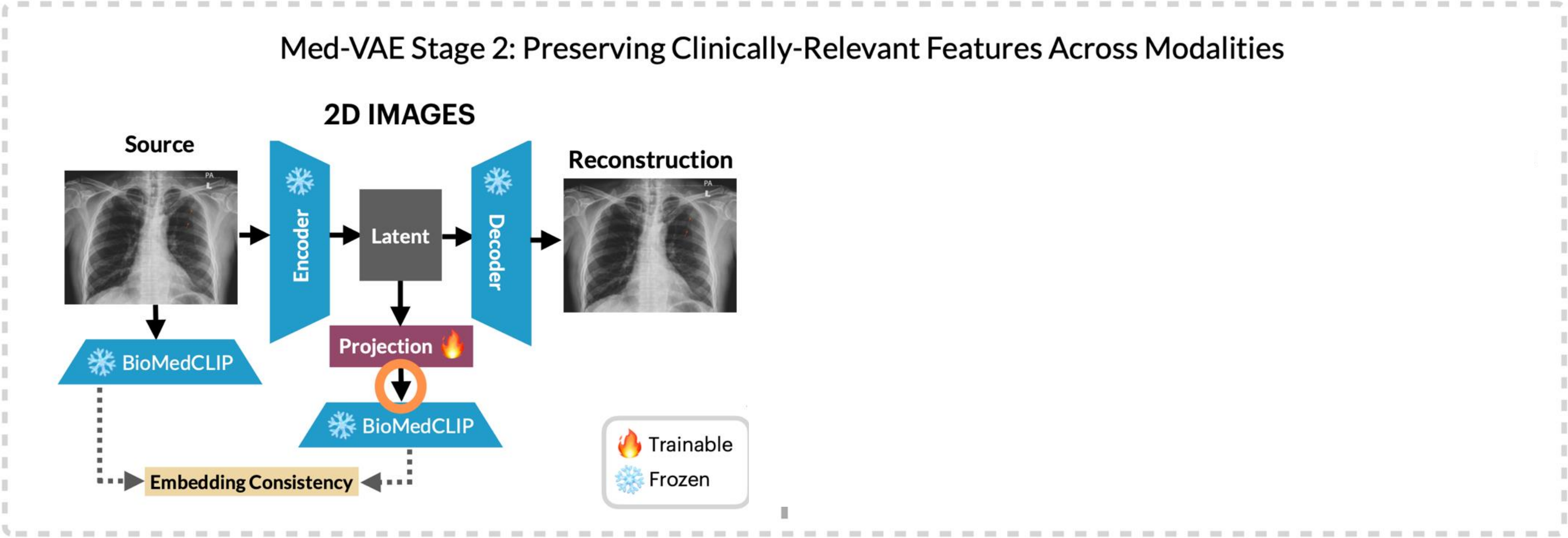
🟦 Real-only baseline
🟧 Synthetic-only
🟨 Synthetic+Real mix
🟩 Synthetic-pretrained
------- Real-only baseline

Medical Image Autoencoders

- Goal to improve efficiency of training large-scale models

- Develop tools for generative & discriminative models

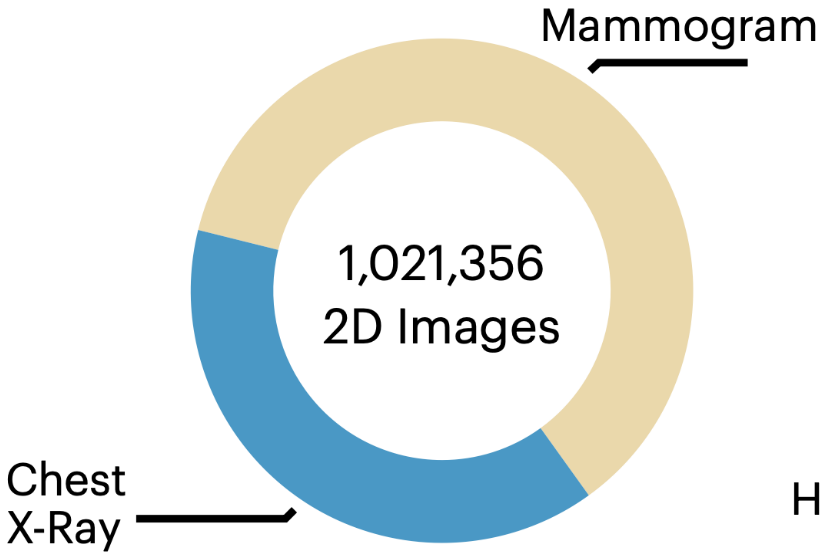
Medical Image Autoencoders

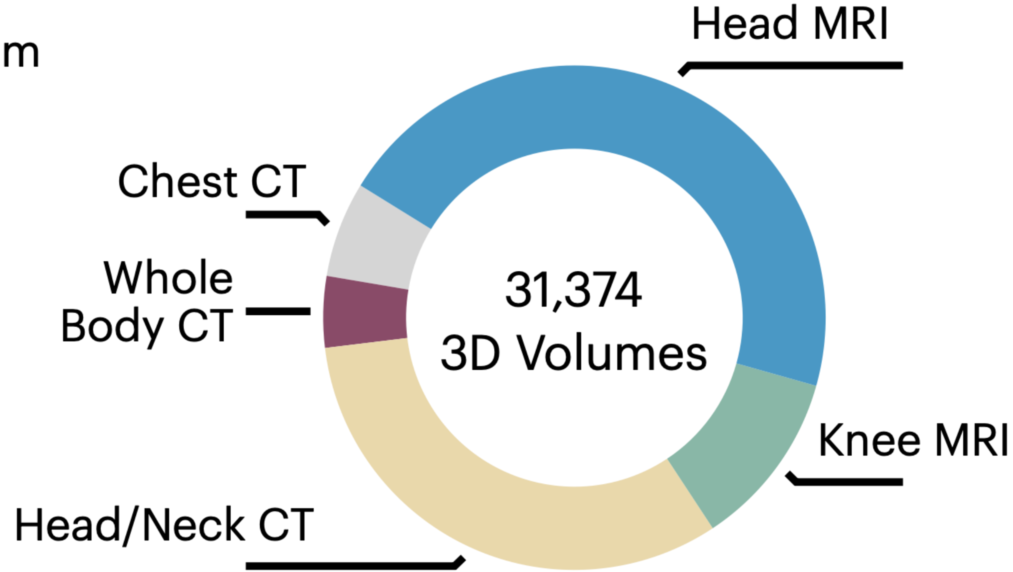


Med-VAE Stage 2: Preserving Clinically-Relevant Features Across Modalities

Varma, Kumar et al. MedVAE: Efficient Automated Interpretation of Medical Images with Large-Scale Generalizable Autoencoders. MIDL 2025.

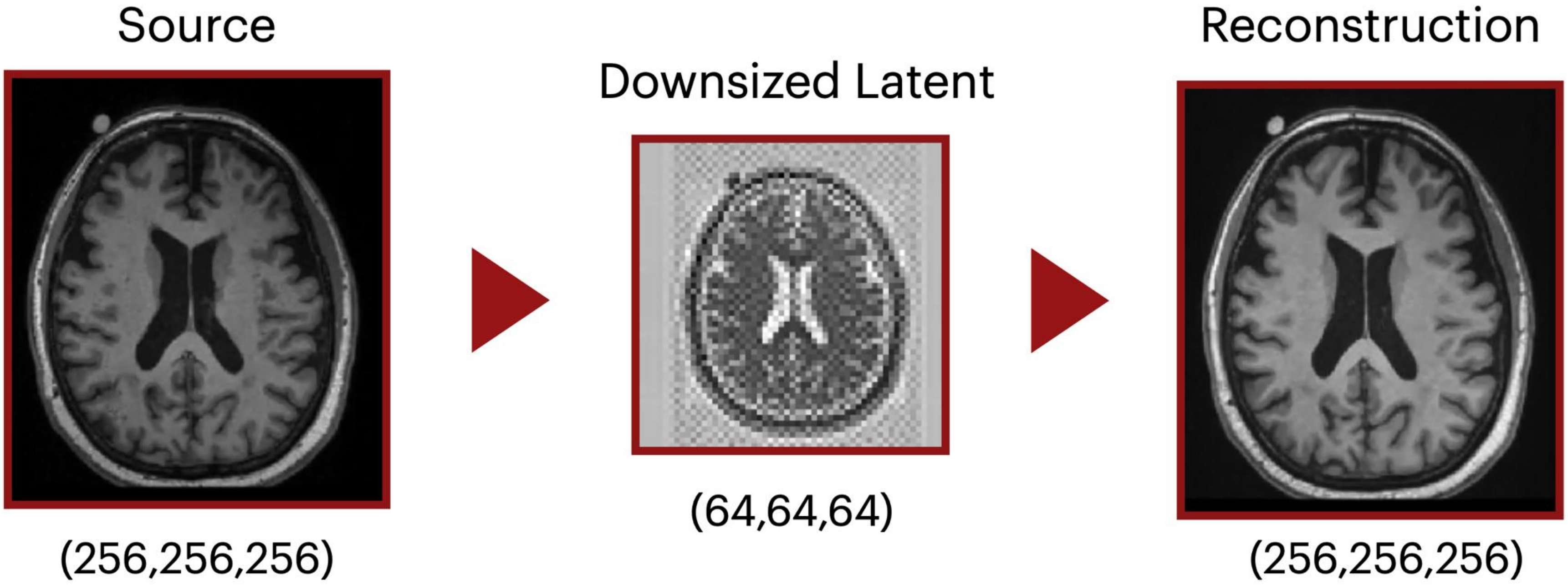
MedVAE Training Data





Varma, Kumar et al. MedVAE: Efficient Automated Interpretation of Medical Images with Large-Scale Generalizable Autoencoders. MIDL 2025.

MedVAE Latents



Source

(256,256,256)

Downsized Latent

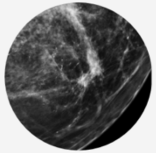
(64,64,64)

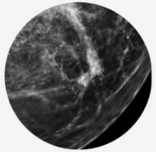
Reconstruction

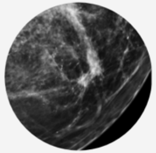
(256,256,256)

Varma, Kumar et al. MedVAE: Efficient Automated Interpretation of Medical Images with Large-Scale Generalizable Autoencoders. MIDL 2025.

MedVAE 2D Performance


Malignancy Detection
Mammogram


Calcification Detection
Mammogram


BI-RADS Prediction
Mammogram

Bone Age Prediction
X-Ray

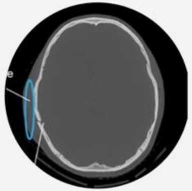
Fracture Detection
X-Ray

Method	Downsize Factor (f)	Mean AUROC
High-Res	1	69.2
Interpolation	16	67.9
General-Domain Autoencoder	16	64.6
2D MedVAE	16	**69.5**

Varma, Kumar et al. MedVAE: Efficient Automated Interpretation of Medical Images with Large-Scale Generalizable Autoencoders. MIDL 2025.

MedVAE 3D Performance

Spine Fracture Classification
CT



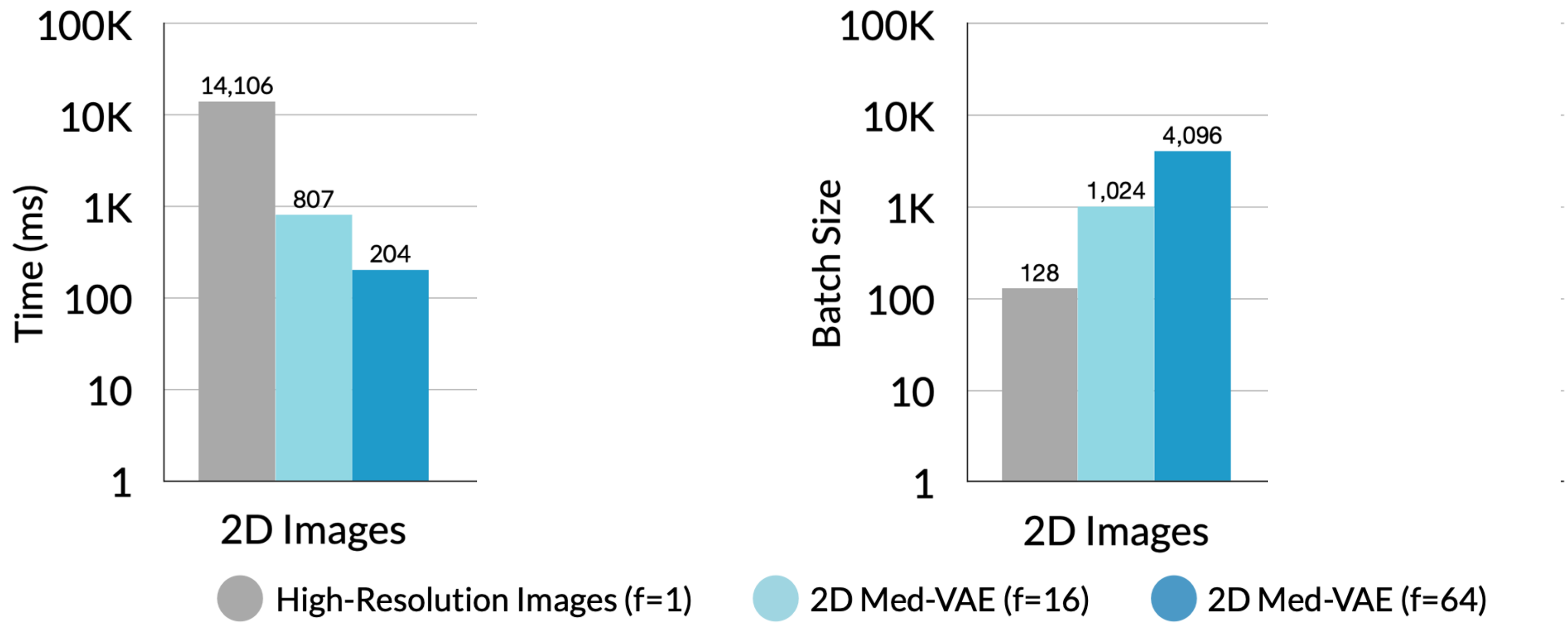
Head Fracture Detection
CT

ACL/Meniscus Tear Detection
MRI

Method	Downsize Factor (f)	Mean AUROC
High-Res	1	72.2
Interpolation	64	69.5
General-Domain Autoencoder	64	70.8
3D MedVAE	64	**79.7**

Varma, Kumar et al. MedVAE: Efficient Automated Interpretation of Medical Images with LargeScale Generalizable Autoencoders. MIDL 2025.

MedVAE Compute Efficiency



Time (in ms) to Perform Forward Pass

Maximum Batch Size

Varma, Kumar et al. MedVAE: Efficient Automated Interpretation of Medical Images with Large-Scale Generalizable Autoencoders. MIDL 2025.

Questions?

akshaysc@stanford.edu