https://arxiv.org/abs/2401.14044

Joint work with Zhuang Liu, Saining Xie, Kaiming He

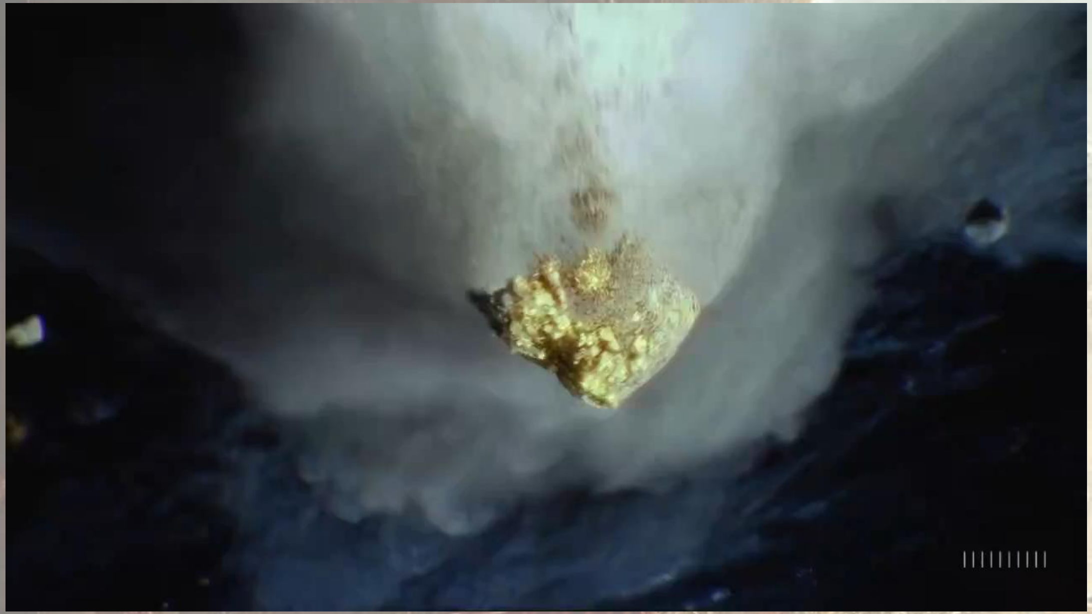
Diffusion Models for Self-Supervised Learning: A Deconstructive Journey

Xinlei Chen

ECCV 2024 Workshop on Self-Supervised Learning – What’s Next?

facebook
Artificial Intelligence Research

Diffusion Models for Generation



"An AI-Generated Picture Won an Art Prize. Artists Aren't Happy."
https://openai.com/sora

Impressive *Generation*, but does it *Understand*?

What I cannot create,
I do not understand

If your goal is to train a world model for recognition or planning, using pixel-level prediction is a terrible idea

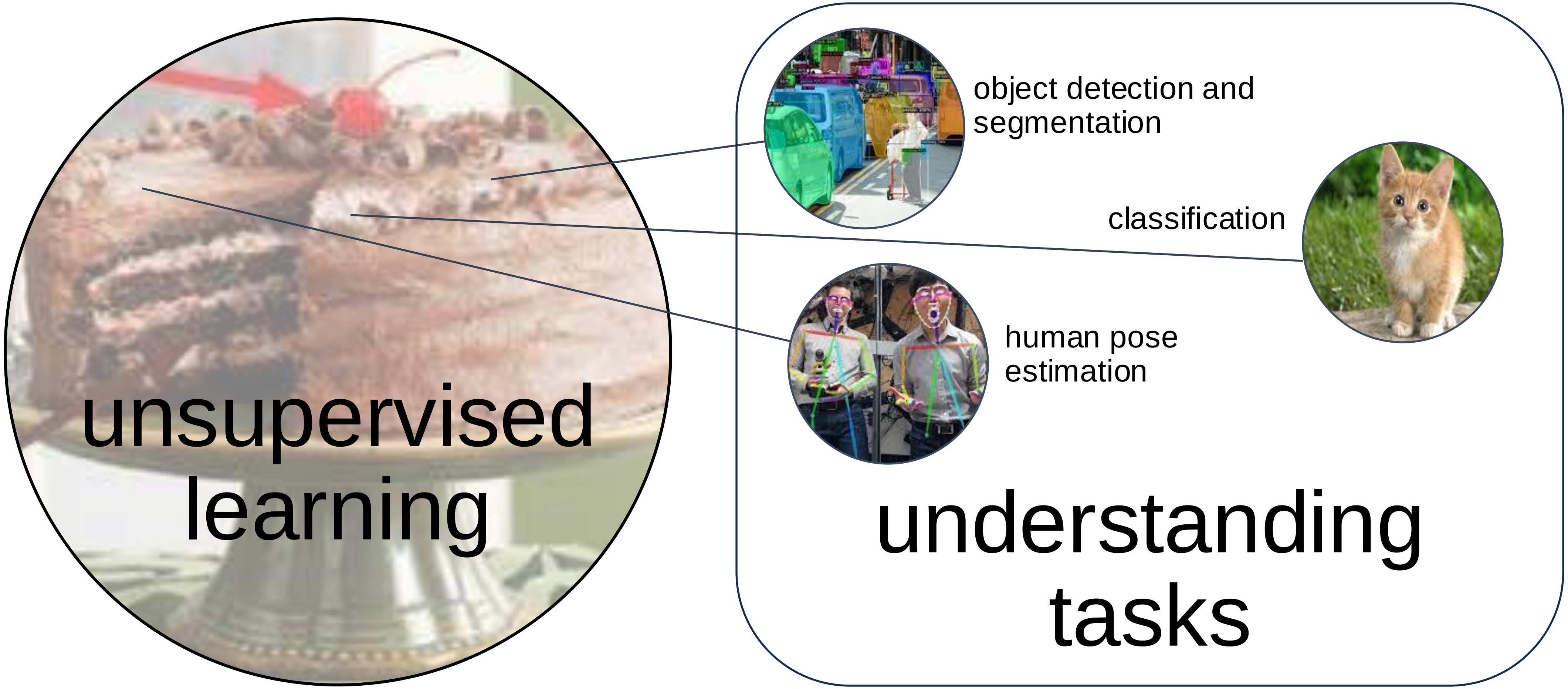
Richard Feynman

So, how much do diffusion models understand?

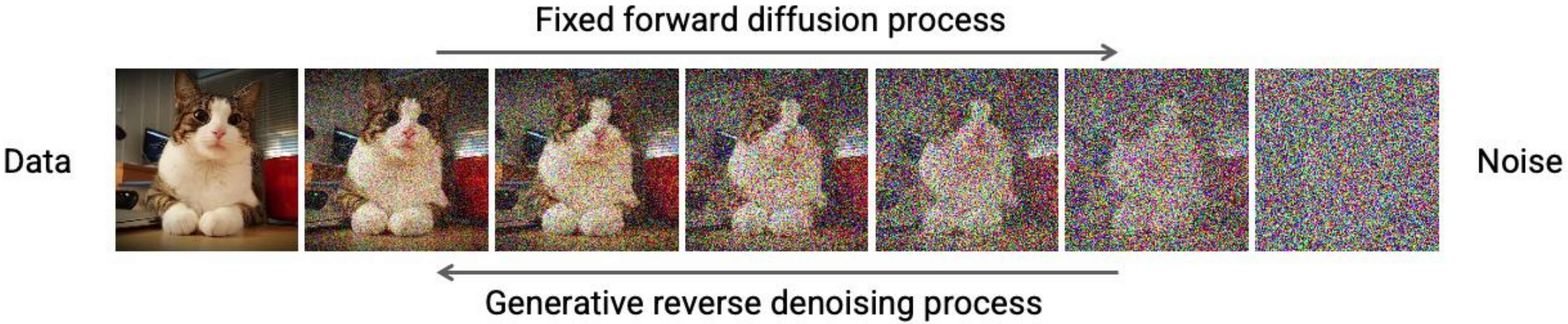
Yann LeCun

Self-Supervised Learning (SSL)

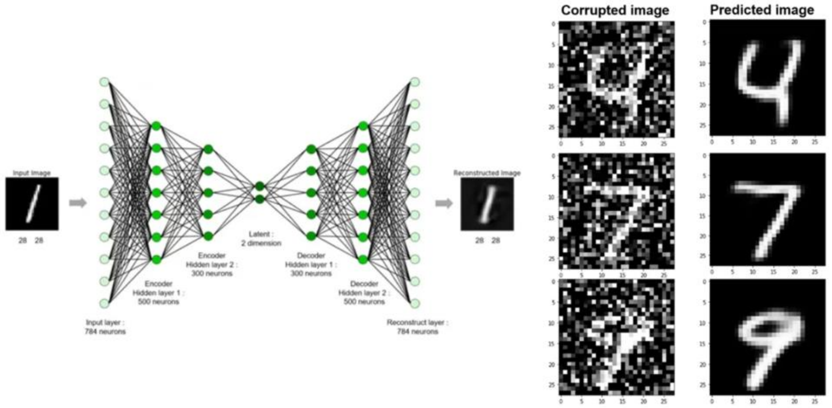
- Pre-train representations without human annotated labels



SSL from Diffusion Models?

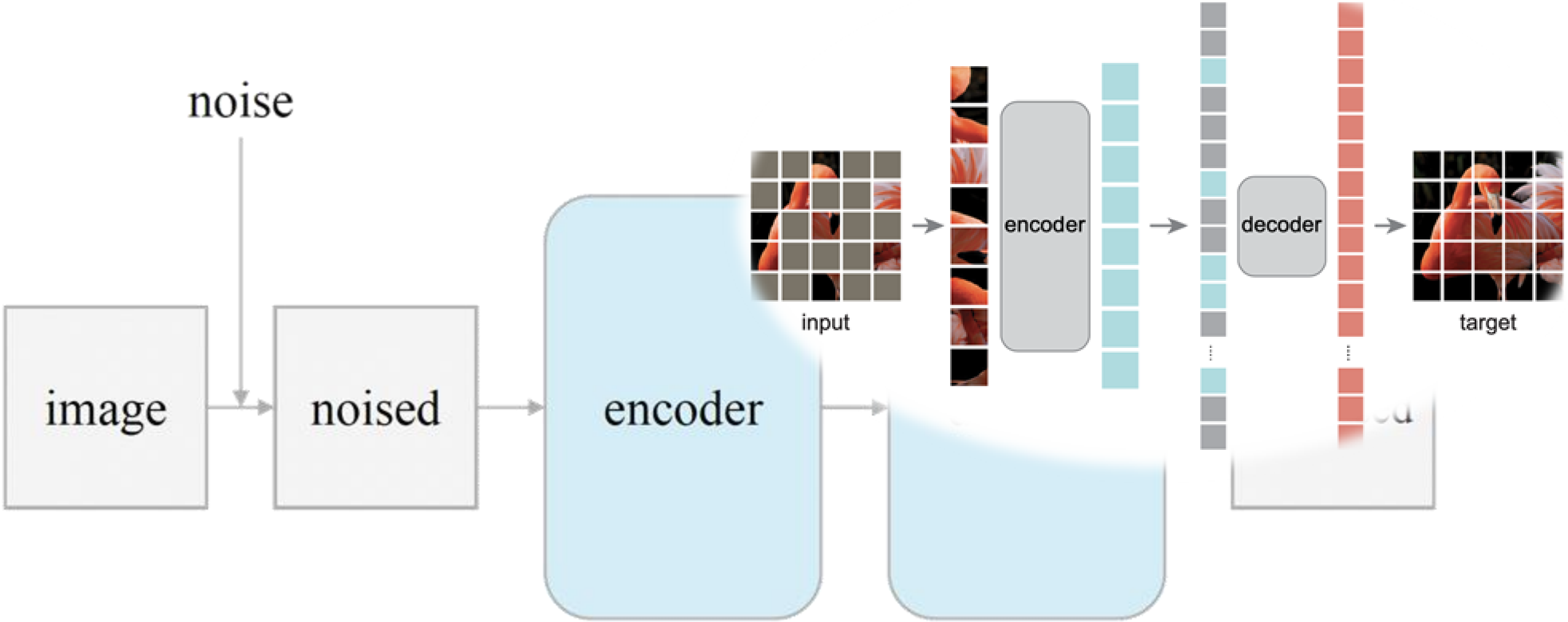


Every time step is essentially a
Denoising Auto-Encoder (DAE)
that does the underlying work



[Vincent et al, ICML 2008]

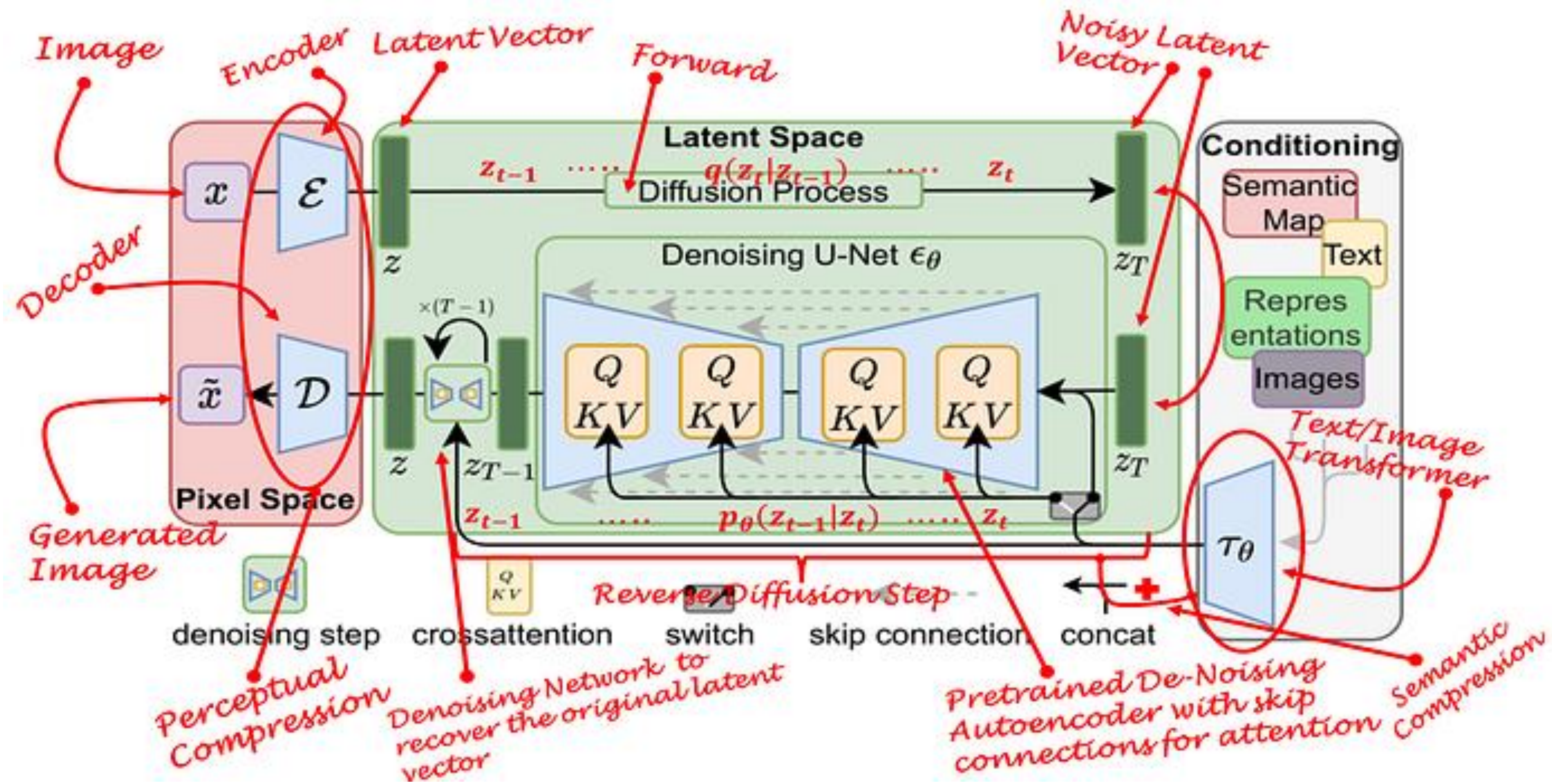
Classical Denoising Auto-Encoders (DAE)



[He et al, ECCV 2022]

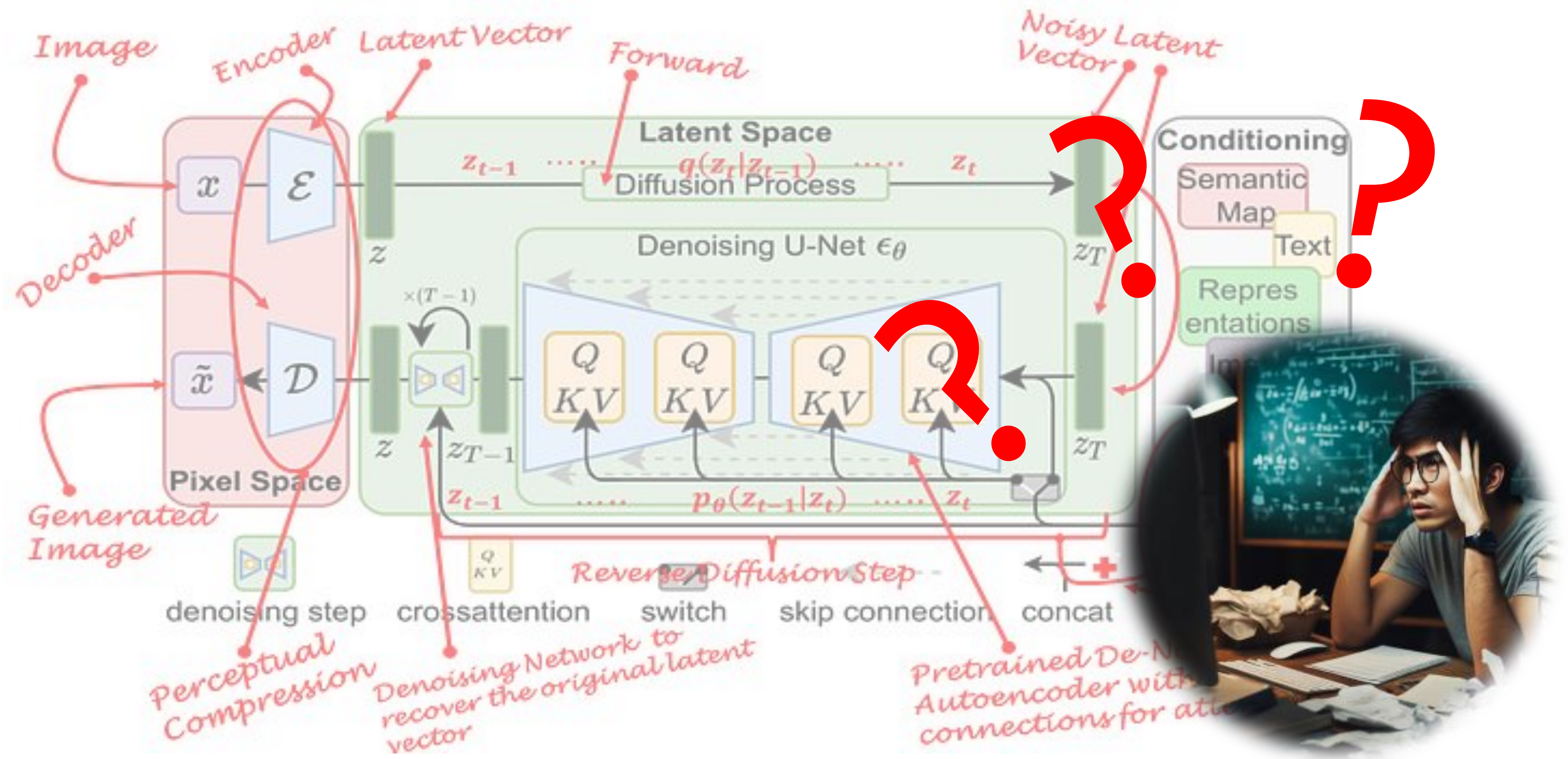
[Rombach et all, CVPR 2022]

Modern Denoising Diffusion Models (DDM)

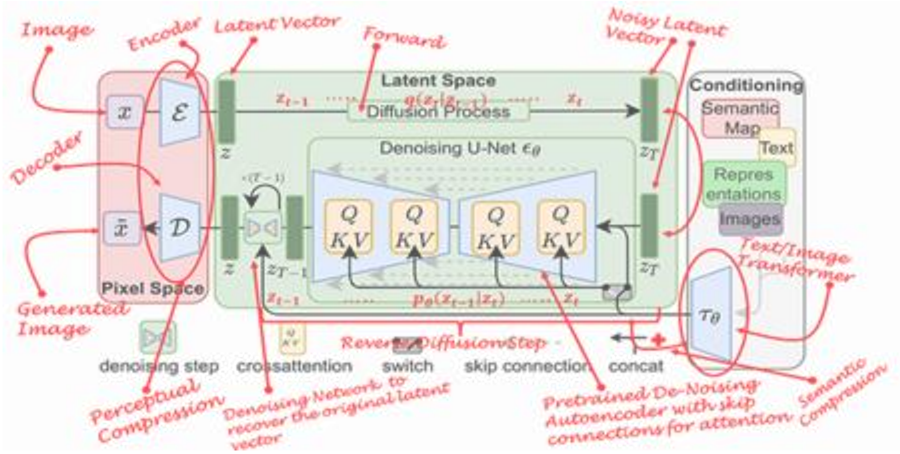


[Rombach et all, CVPR 2022]

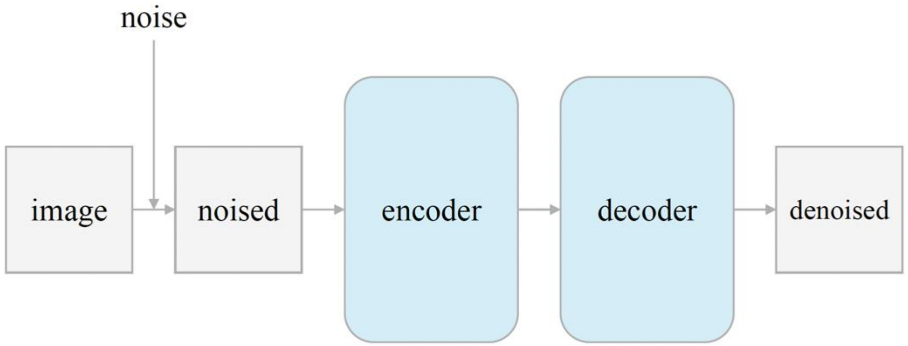
Modern Denoising Diffusion Models (DDM)



Goal: Deconstruct DDM toward DAE

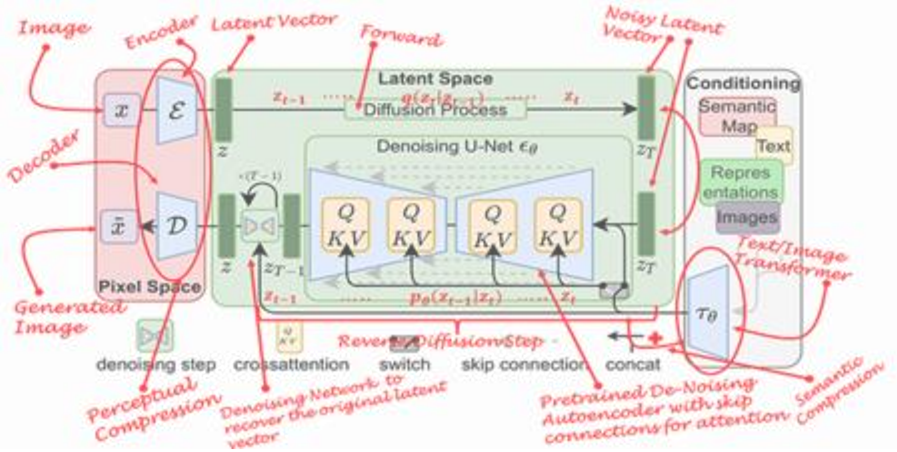


Modern DDM
for Image Generation

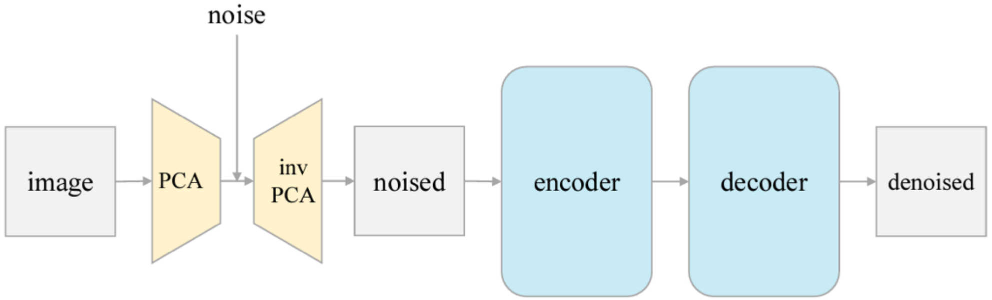


Classical DAE
for Image Understanding

L-DAE: Outcome after Deconstruction



Modern DDM
for Image Generation



latent-DAE
for Image Understanding

adding noise in the low-dimensional *latent* space is crucial
l-DAE: drastically closed the gap to existing working paradigms

Overview of the Deconstructive Journey

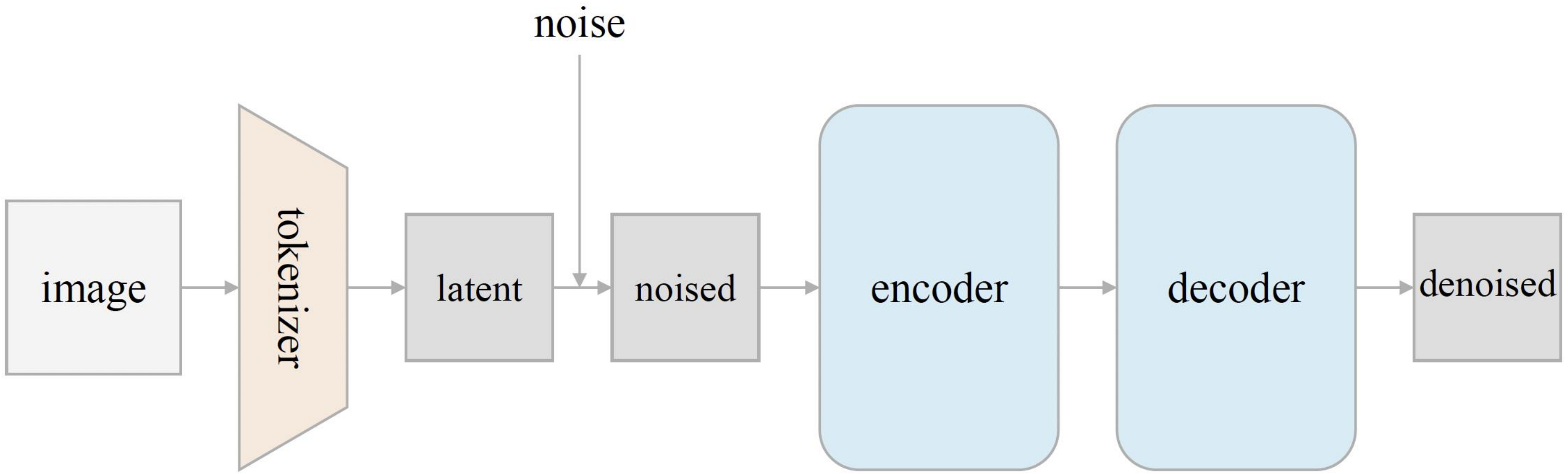
1. Initialization: DiT

2. Re-orienting DiT for SSL

3. Deconstructing the tokenizer

4. Toward classical DAE

1. Initialization: Diffusion Transformer (DiT)



ImageNet:

Acc ↑	57.5
FID ↓	11.6

significantly better than we expected!

- Transformer blocks for the autoencoder
- Another autoencoder (VQGAN) provides latent token space for denoising
- DiT-L overall, so DiT-$\frac{1}{2}$L as encoder for linear probing

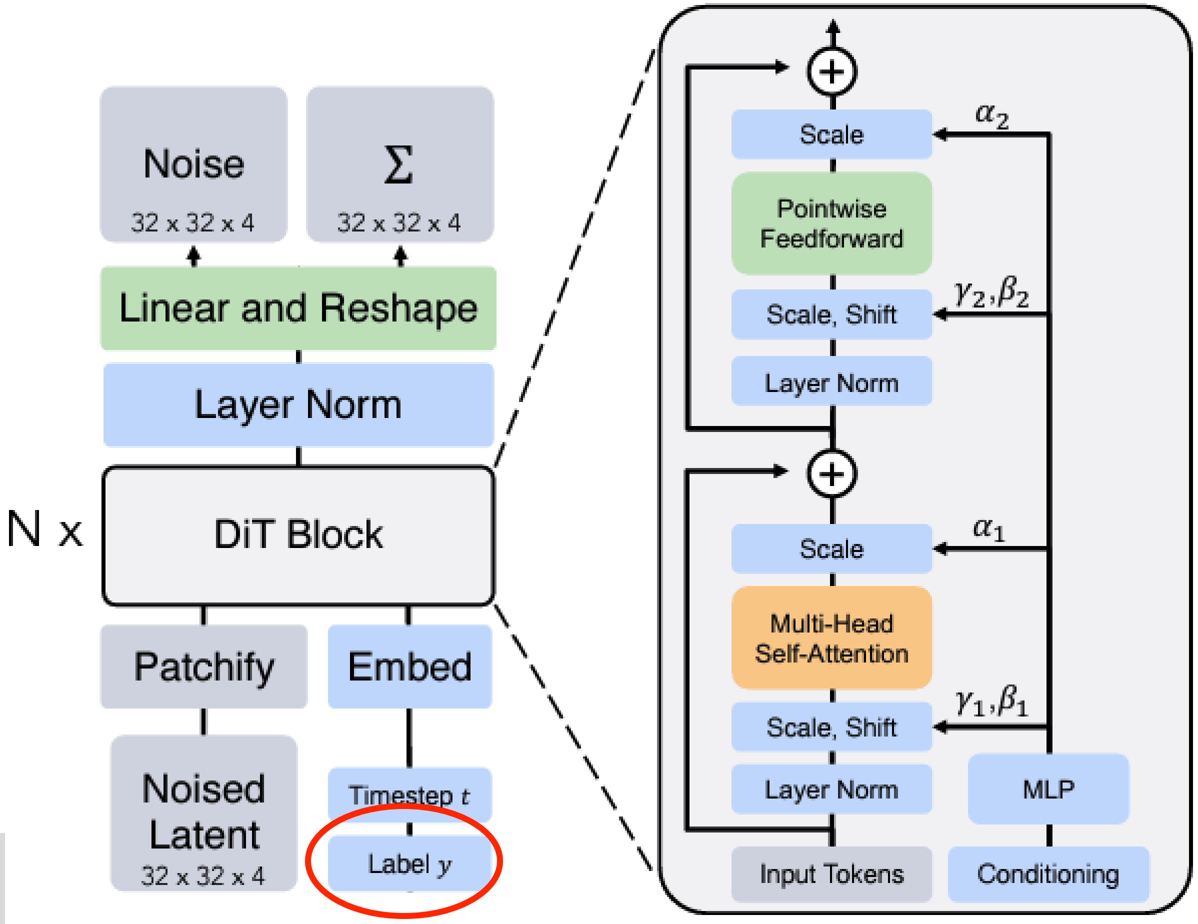
2. Re-Orienting DiT for SSL

2a. Remove class-conditioning

Otherwise *not legitimate* SSL

Acc ↑	57.5 → 62.5
FID ↓	11.6 → 30.9

labels causes the model to “cheat”



Latent Diffusion Transformer

DiT Block with adaLN-Zero

[Simonyan and Zisserman, ICLR 2015] [Zhang et all, CVPR 2018]

2. Re-Orienting DiT for SSL

2b. Remove LPIPS loss in VQGAN

LPIPS: VGG features to approximate human perceptual similarity

Also *not legitimate* for SSL, as VGG is trained on ImageNet labels

Acc ↑	62.5 → 58.4
FID ↓	30.9 → 54.3

the label information can propagate very far

[Goodfellow et al, NeurIPS 2014] [Isola et al, CVPR 2017] [Kingma and Welling, ICLR 2014]

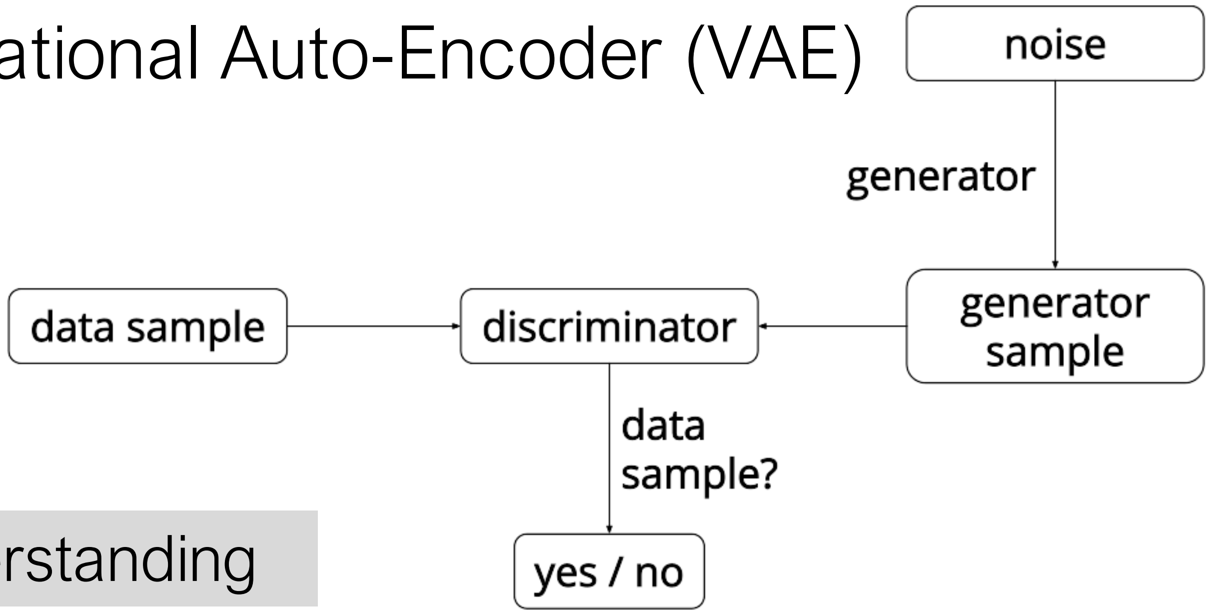
2. Re-Orienting DiT for SSL

2c. Remove GAN loss in VQGAN

GAN: Generative Adversarial Network

Afterwards, VQGAN becomes Variational Auto-Encoder (VAE)

Acc ↑	58.4 → 59.0
FID ↓	54.3 → 75.6

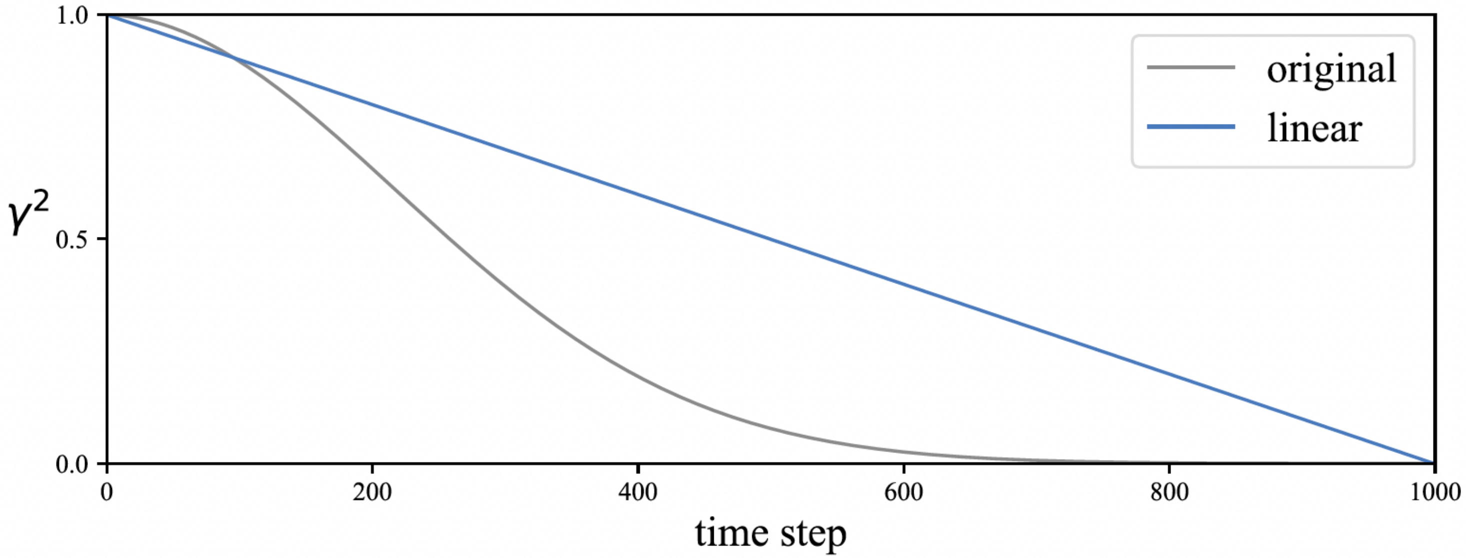


GAN helps generation, but hurts understanding

[Ho et al, NeurIPS 2020]

2. Re-Orienting DiT for SSL

2d. Noise schedule change for image understanding



Acc ↑	59.0 → 63.4
FID ↓	75.6 → 93.2

high-noise levels help generation but not understanding

[Esser et all, CVPR 2021] [Kingma and Welling, ICLR 2014]

3. Deconstructing the Tokenizer

Current tokenizer -- *Convolutional VAE*:

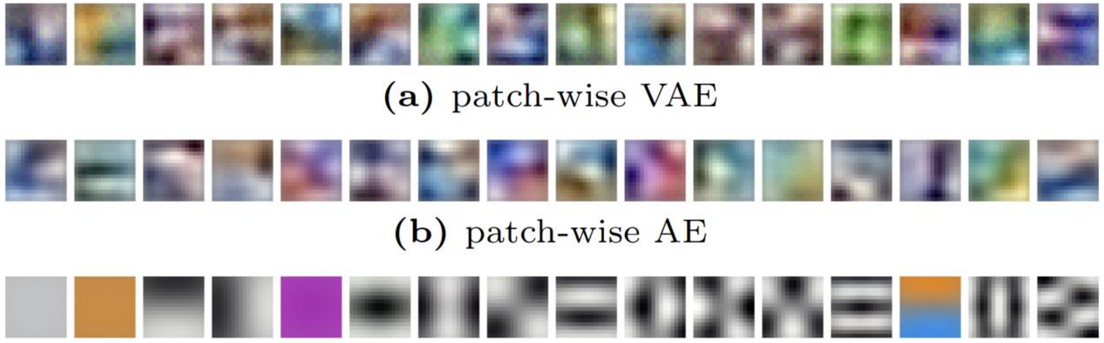
$$\|x - g(f(x))\|^2 + \mathbb{KL}[f(x)|\mathcal{N}]$$

Deconstruct it step-by-step:

a) *Patch-wise VAE*, $\|x - U^T V x\|^2 + \mathbb{KL}[Vx|\mathcal{N}]$

b) *Patch-wise AE*, $\|x - U^T V x\|^2$

c) *Patch-wise PCA*, $\|x - V^T V x\|^2$



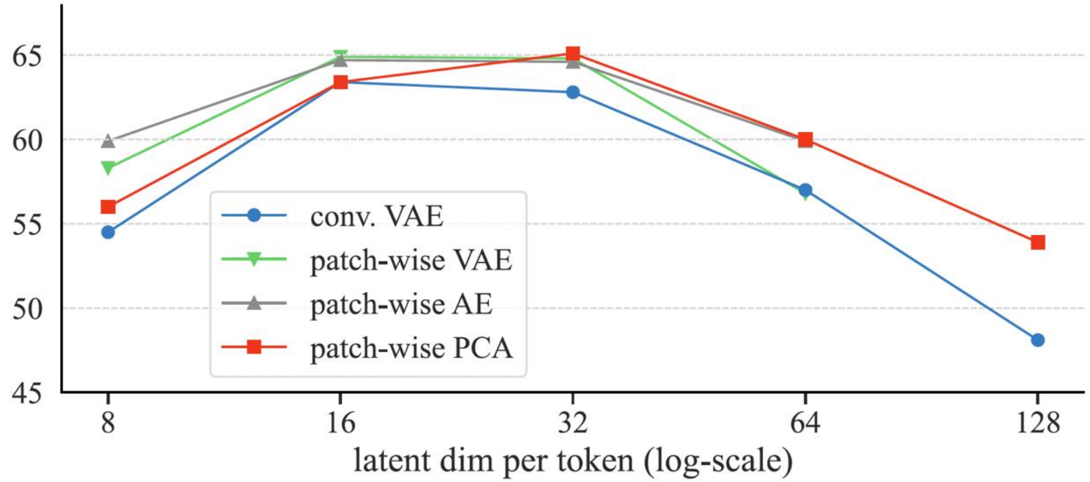
(a) patch-wise VAE

(b) patch-wise AE

(c) patch-wise PCA

3. Deconstructing the Tokenizer

latent dim	8	16	32	64
conv. VAE	54.5	**63.4**	62.8	57.0
patch-wise VAE	58.3	**64.9**	64.8	56.8
patch-wise AE	59.9	**64.7**	64.6	59.9
patch-wise PCA	56.0	63.4	**65.1**	60.0

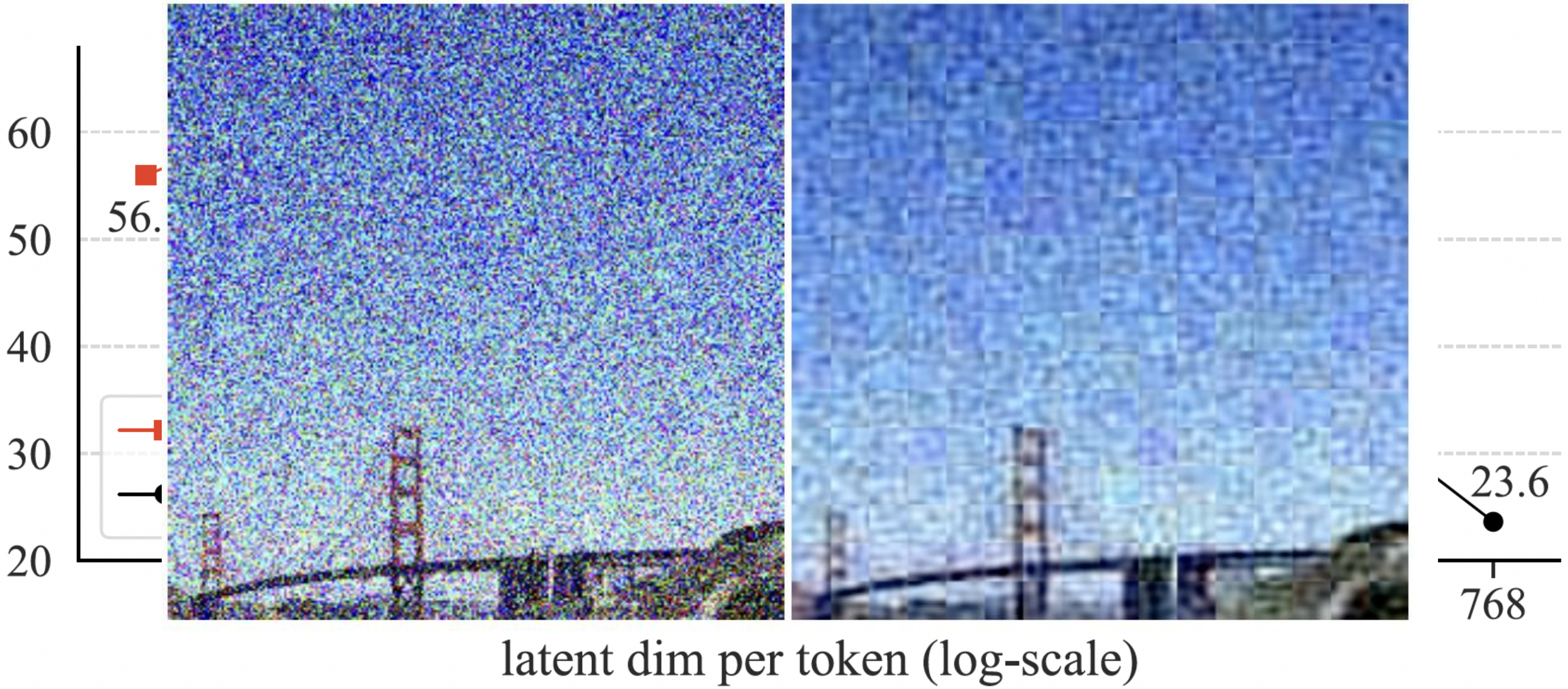


latent dimension of the tokenizer is *crucial*

specific variants of the tokenizer matter much less

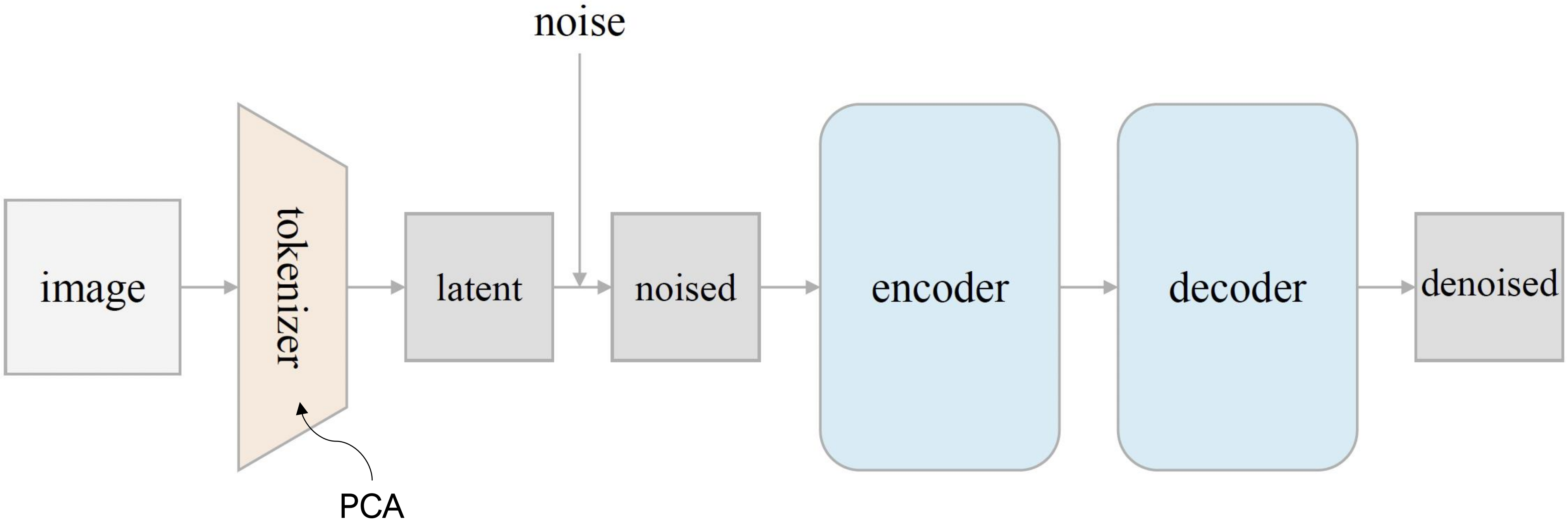
16 or 32 dimensions are about optimal for 16x16 patches

What About Directly Resizing Patches?



high-resolution, pixel-based DDMs are not great for SSL

4. Toward Classical DAE



After all the deconstructions so far, this old view still holds..

Can we get as close as possible to a classical DAE?

4. Signal vs. Noise Simplifications

noised → … → denoised

	noised	denoised
prev. Default in DiT	$\gamma z_0 + \sqrt{1-\gamma^2}\epsilon$	$\epsilon \to z_0$
4a. Predict signal, *not* noise	$\gamma z_0 + \sqrt{1-\gamma^2}\epsilon$	z_0
4b. Remove signal scaling	$z_0 + \sigma\epsilon$	z_0

	Acc
prev.	65.1
4a.	62.4
4b.	63.6

hurts accuracy, but not as crucial as latent noise

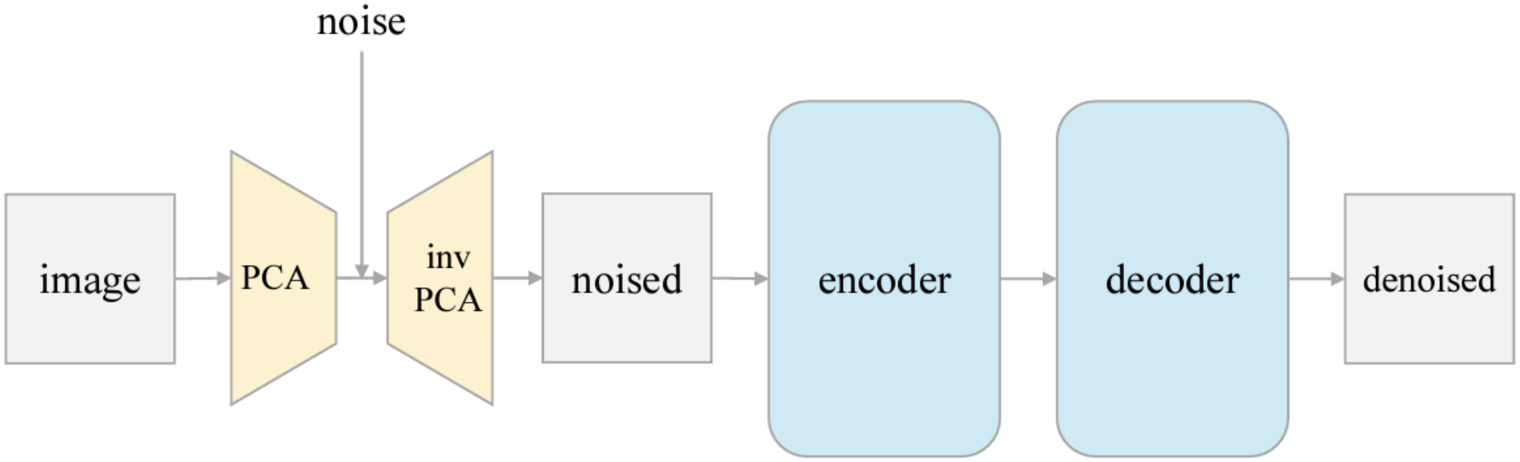
4. DAE Directly on Pixels

4c. Pixel input with inv. PCA

4d. Pixel output with inv. PCA

4e. Original image as output

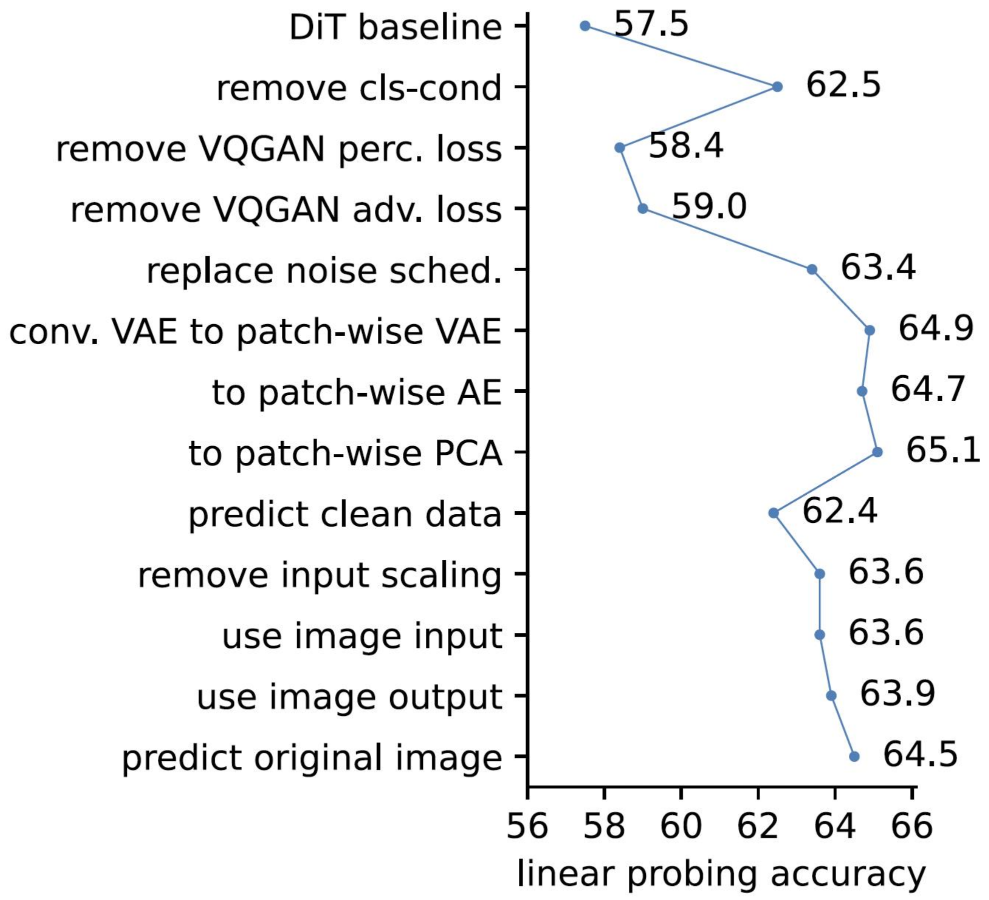
	Acc
prev.	63.6
4c. input	63.6
4d. output	63.9
4e. original	64.5

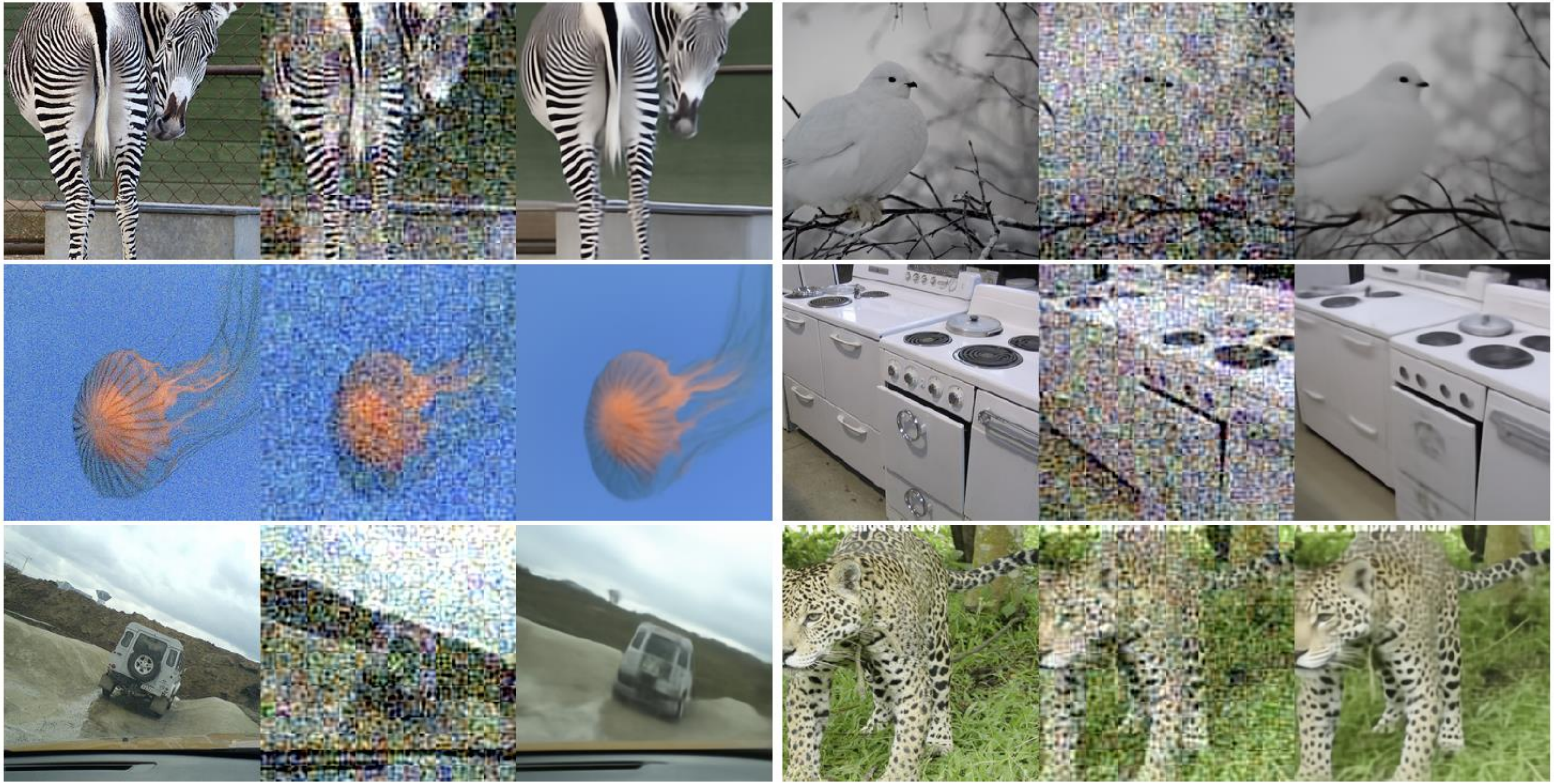


our final *l*-DAE model

Summary: the Deconstructive Journey

1. Initialization: DiT
2. Re-orienting DiT for SSL
3. Deconstructing the tokenizer
4. Toward classical DAE





Quantitative Ablations for *l*-DAE

- Time steps

	multiple	single
Acc ↑	64.5	61.5

a key diffusion design, but not so important for SSL

- Data augmentation

	center crop	random crop
Acc ↑	64.5	65.0

helps especially with longer training

Scaling Behaviors for *l*-DAE

- Training epoch

	400	800	1600
Acc ↑	65.0	67.5	69.6

- Model size

	ViT-B	ViT-$\frac{1}{2}$L	ViT-L
Acc ↑	60.3	65.0	70.9

[He et all, CVPR 2022] [He et all, CVPR 2021]

System-Level Comparison, *Classification*

pre-train	ViT-B	ViT-L
MoCo v3	**76.7**	**77.6**
MAE	68.0	75.8
l-DAE	66.6	75.0

Linear Probing

pre-train	ViT-B	ViT-L
MoCo v3	83.2	84.1
MAE	83.6	**85.9**
l-DAE	**83.7**	84.7

Fine-Tuning

compared to DAE (20+ linear probe), *l*-DAE drastically closed the gap to MAE

contrastive methods are generally better for linear probing

autoencoders are generally better in fine-tuning

System-Level Comparison, *Detection*

pre-train	ViT-B		ViT-L	
	AP^{box}	AP^{mask}	AP^{box}	AP^{mask}
Supervised	47.6	42.4	49.6	43.8
MAE	51.2	45.5	**54.6**	**48.6**
l-DAE	**51.6**	**45.8**	54.4	48.2

l-DAE outperforms MAE in ViT-B, and significantly over supervised

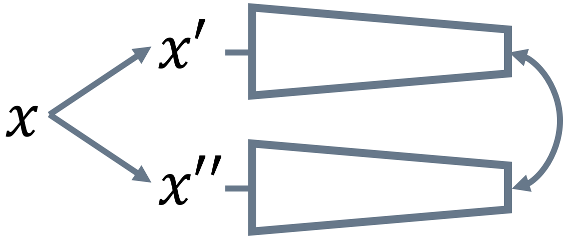
[Lin et all, ECCV 2014] [Li et all, ECCV 2022]

Denoising Diffusion Models for SSL

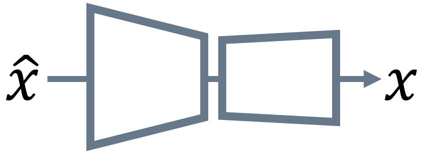
xinleic.xyz

- Modern DDMs have reasonably good understandings of images
- More due to *latent Denoising*, and less to Diffusion: *l*-DAE
- *l*-DAE adds a standalone, clean alternative to current SSL methods

- Joint-Encoder



- Auto-Encoder



MAE, ***l*-DAE**, …