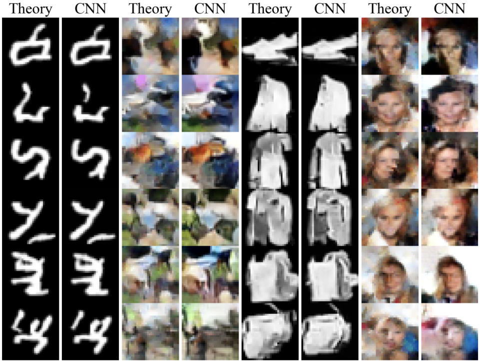
An analytic theory of creativity in convolutional diffusion models

Mason Kamb[1], Surya Ganguli [1]
[1]Applied Physics, Stanford University
July 15, 2025



What is the origin of combinatorial ‘creativity’ in diffusion models?

- Models regularly create **entirely novel images** that **combine** of features from their training data, mixing and matching without purely memorizing.
- These combinations are **novel**, yet still qualitatively **consistent** with their training data.
- We’d like to find an analytic theory that makes these properties manifest.

panda mad scientist mixing sparkling chemicals, artstation

a corgi’s head depicted as an explosion of a nebula

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square

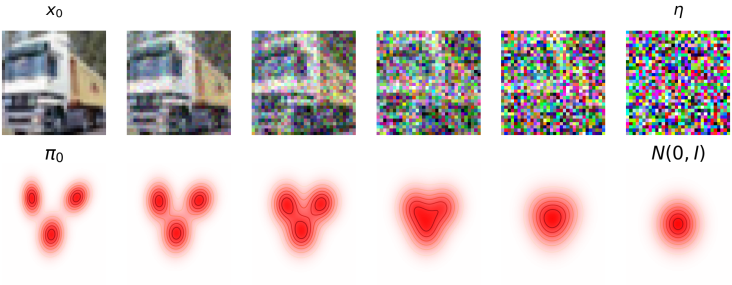
Sample images from DallE-2

‘Consistency’ is famously not guaranteed!


arXiv:2404.05384, arXiv:2403.10731, arxiv:2403.10731v2

- Diffusion models are notorious for spatial consistency issues– incorrect numbers of fingers, incorrect limb placement, etc.
- Creativity and inconsistency are on a spectrum; in the models we study, we find that the same mechanism predicts both phenomena.

Diffusion Models



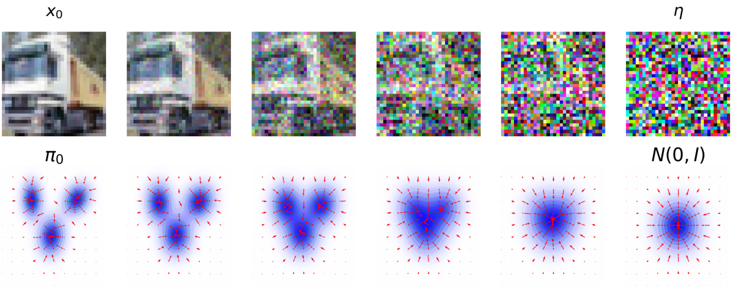
- Suppose we have a distribution π_0 we would like to sample from. We can take samples $x_0 \sim \pi_0$ and 'corrupt' them by interpolating them with white noise $\eta \sim N(0, I)$:

$$\phi_t = \sqrt{\bar{\alpha}_t}\phi_0 + \sqrt{1 - \bar{\alpha}_t}\eta$$

- Solution to an OU process (the forward process):

$$d\phi_t = -\gamma_t \phi_t + \sqrt{2\gamma_t}\, dW_t$$

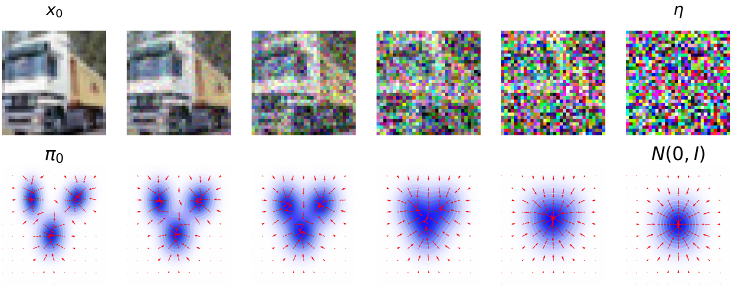
Diffusion Models



- Reverse process (DDIM):

$$\frac{d\phi_t}{dt} = -\gamma_t(\nabla \log \pi_t(\phi_t) + \phi_t)$$

Diffusion Models



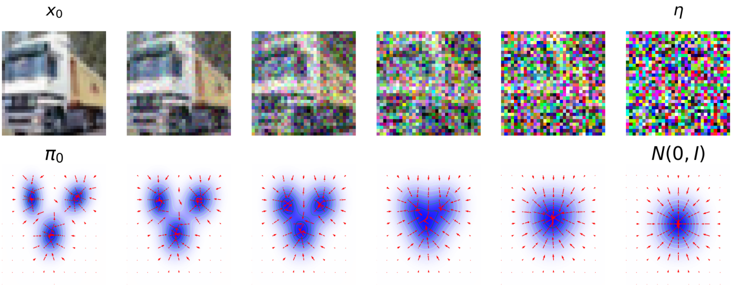
- Tweedie's Theorem:

$$\mathbb{E}[\eta|\phi_t] \propto \nabla \log \pi_t$$

- To learn the score, we train a neural network $M_\theta(\phi, t)$ with the objective of guessing the noise from the

$$\mathcal{L}(\theta) = \mathbb{E}_{\phi_t \sim \pi_t, \eta \sim N(0,I)}[\|\eta - M_\theta(\phi_t, t)\|^2]$$

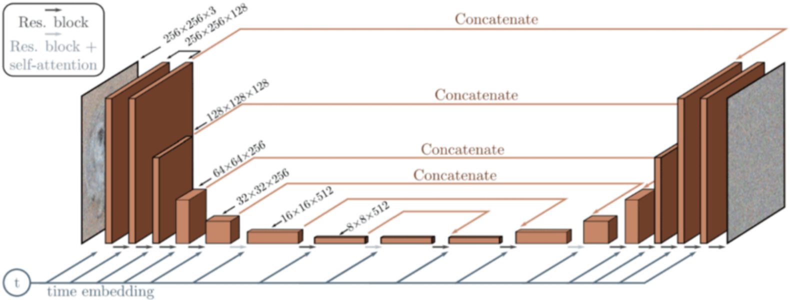
Diffusion Models



- Problem: reverse process exactly reverses the forward process; as $t \to 0$, we recover a sum of delta functions on the training data!
- In other words, *ideal diffusion models memorize.* The phenomenon of combinatorial creativity *must* emerge because the neural network has *underfit* its training objective!
- To understand why diffusion models are successful, we need to understand the **implicit biases and constraints** that prevent the model from minimizing its objective, and understand **what it learns instead**.

Simplest realistic diffusion model: fully convolutional neural network

Most commonly used architecture for diffusion models is based on a UNet+self-attention; we study the stripped-down version (no self-attention)



Inductive biases of CNNs

In general CNNs can be arbitrarily expressive, except for the following two constraints:

- **Translational equivariance**: applying the model to a translated version of an input image results in an equally translated output.
- **Locality:** the convolutional filters used are typically very narrow. For a finite-depth network, this means that only the pixels in a *local region* around the pixel can be used to estimate the noise. (Also, emergent locality bias– see later!)

What is the *optimal denoiser* under these constraints?

Bayes-Optimal Denoising under Locality and Equivariance

- The *ideal* score function can be written as a linear combination of the displacement from each training sample, times a *global* Bayes weight for each data point:

$$M_t(\phi, x) \propto \underbrace{\sum_{\varphi \in \mathcal{D}}}_{\text{sum over data}} \underbrace{(\phi(x) - \sqrt{\bar{\alpha}_t}\varphi(x))}_{\text{added noise}} \underbrace{P(\varphi|\phi)}_{\text{Bayes weight}}$$

$$P(\varphi|\phi) = \frac{\mathcal{N}(\phi|\sqrt{\bar{\alpha}_t}\varphi, (1-\bar{\alpha}_t)I)}{\sum_{\varphi'} \mathcal{N}(\phi|\sqrt{\bar{\alpha}_t}\varphi', (1-\bar{\alpha}_t)I)}$$

Bayes-Optimal Denoising under Locality and Equivariance

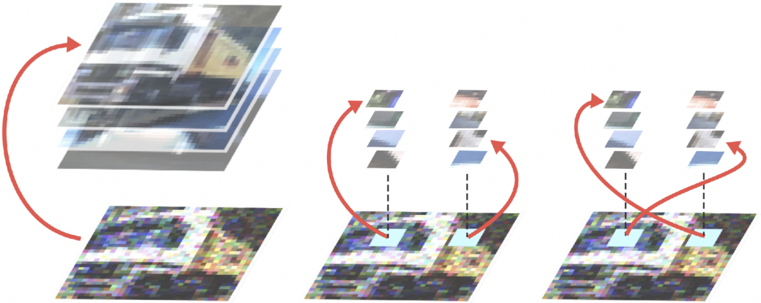
- The ideal *local* denoiser (LS): each pixel x has *its own belief state* about which image it came from, *based only on the information in its local neighborhood* Ω_x.

$$M_t(\phi, x) \propto \underbrace{\sum_{\varphi \in \mathcal{D}}}_{\text{sum over data}} \underbrace{(\phi(x) - \sqrt{\bar{\alpha}_t}\varphi(x))}_{\text{added noise}} \underbrace{P(\varphi|\phi_{\Omega_x})}_{\text{local Bayes weight}}$$

- The ideal *equivariant, local* approximation to the score (ELS): dataset augmented with *all possible translations* of the original dataset.

$$M_t(\phi, x) \propto \underbrace{\sum_{\varphi \in G(\mathcal{D})}}_{\text{sum over data + translations}} \underbrace{(\phi(x) - \sqrt{\bar{\alpha}_t}\varphi(x))}_{\text{added noise}} \underbrace{P(\varphi|\phi_{\Omega x})}_{\text{local Bayes weight}}$$

Denoising under locality + equivariance



(a) IS Machine (b) LS Machine (c) ELS Machine

Combinatorial creativity from the locality constraint

- **Key takeaway:** under *local* denoising, each individual pixel is drawn towards the patch in the training dataset that it most believes it came from $\implies$ **automatically mixing and matching the training data in different parts of the image while retaining local consistency.**
- This is the *key mechanism that underpins combinatorial creativity* in convolutional diffusion models.

So... does it work?

- Trained two architectures...
  - 6-layer ResNet with 3×3 convolutional filters.
  - 3-scale UNet.
- ...on four standard small image datasets:
  - MNIST
  - FashionMNIST
  - CIFAR10
  - CelebA
- We compared outputs of theoretical model to outputs of trained diffusion models given identical noise inputs.

Results

(a) Theory (left) vs. ResNet (right)

(b) Theory (left) vs. UNet (right)

More Results: MNIST

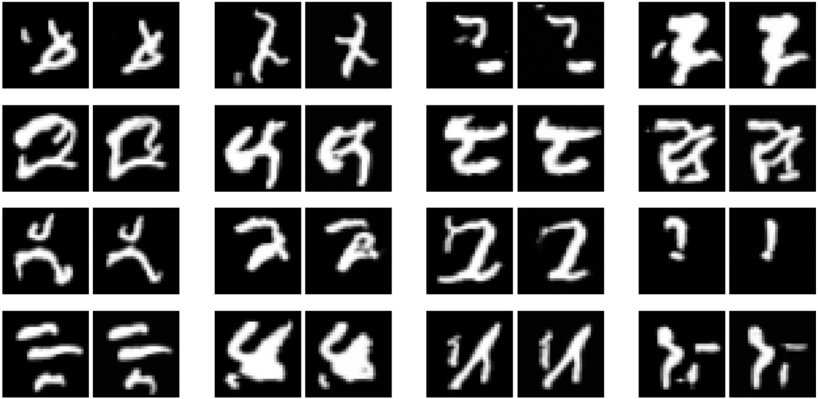


Figure: Left Columns: Theory, Right Columns: Neural Network.

More Results: FashionMNIST

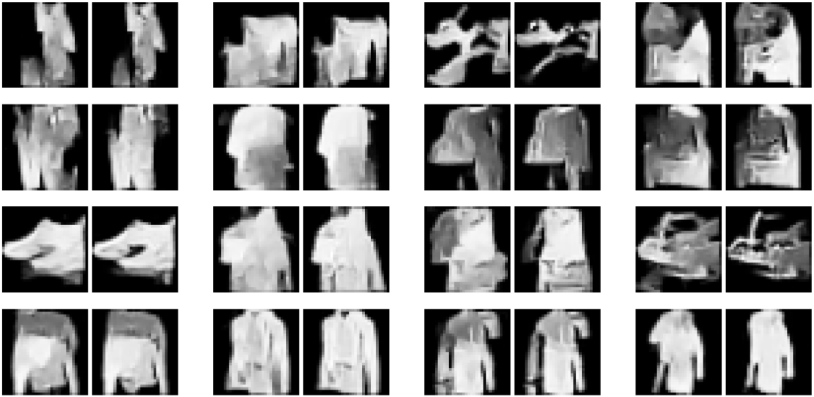


Figure: Left Columns: Theory, Right Columns: Neural Network.

More Results: CIFAR10

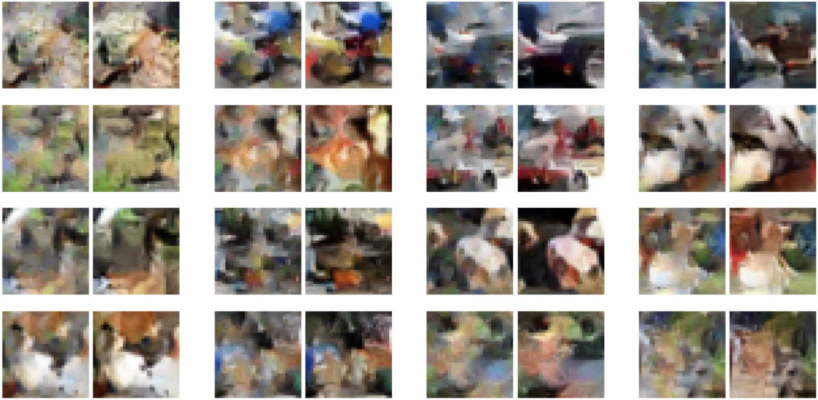


Figure: Left Columns: Theory, Right Columns: Neural Network.

More Results: CelebA

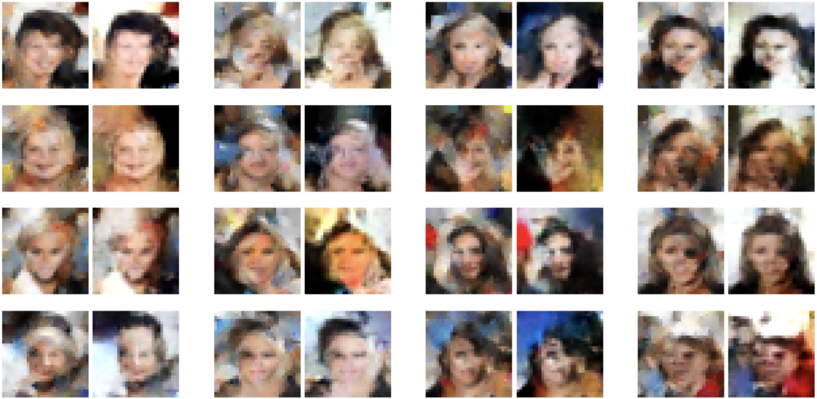


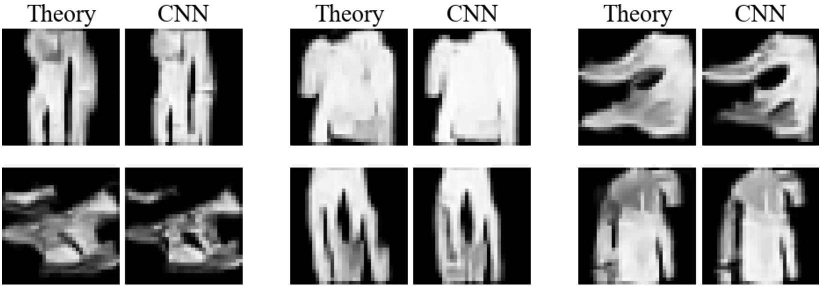
Figure: Left Columns: Theory, Right Columns: Neural Network.

Results

Dataset	Arch.	Padding	Conditional	ELS Corr.	LS Corr.	IS Corr.	(E)LS > IS %
MNIST	UNet	Zeros	✗	**0.89**	0.88	0.70	0.93
CIFAR10	UNet	Zeros	✓	**0.90**	0.87	0.41	0.92
FashionMNIST	UNet	Zeros	✓	**0.93**	0.93	0.80	1.00
CelebA	UNet	Zeros	✗	0.85	**0.90**	0.55	1.00
MNIST	ResNet	Zeros	✗	**0.94**	0.82	0.61	1.00
MNIST	ResNet	Circular	✗	**0.77**	0.36	0.15	0.92
CIFAR10	ResNet	Zeros	✓	**0.95**	0.90	0.42	1.00
CIFAR10	ResNet	Circular	✓	**0.94**	0.83	0.35	1.00
FashionMNIST	ResNet	Zeros	✓	**0.94**	0.88	0.68	1.00
CelebA	ResNet	Zeros	✗	**0.96**	0.90	0.47	1.00

Figure: Median pixelwise ELS/CNN r^2 values.

Theory predicts spatial consistency issues!



- FashionMNIST results display interpretable issues with extra limbs, predictable by theory and attributable to excess locality.
- Mechanism: for overly small locality scales, a given pixel can't tell whether there are too many or too few limbs in the image.

Multiscale behavior and the curse of dimensionality

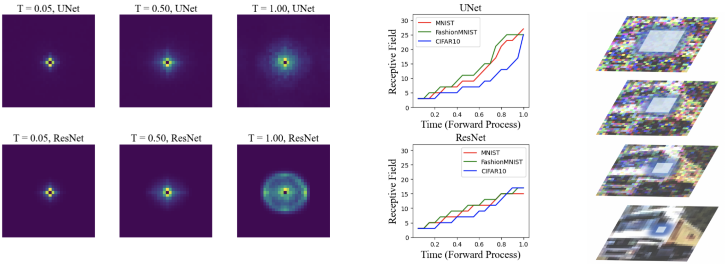
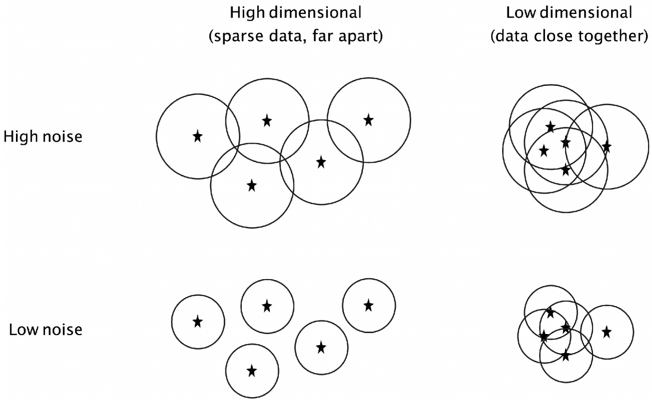


Figure: Left: empirical receptive fields at different times in the reverse process. Middle: Optimal scales across models and datasets. Right: schematic depiction of time-dependent locality scale.

- The best-fit locality scale is large at high noise levels and monotonically decreases through the reverse process.

Multiscale behavior and the curse of dimensionality



- In high dimensions data are very far apart, meaning that memorization onsets at relatively high noise.
- By continuously projecting to lower dimensions as the noise level is reduced, the model stays above the memorization threshold throughout the reverse process.

Attention-enabled models

- In practice, our theory is also moderately predictive ($r^2 \sim 0.77$) of Attention-enabled models!

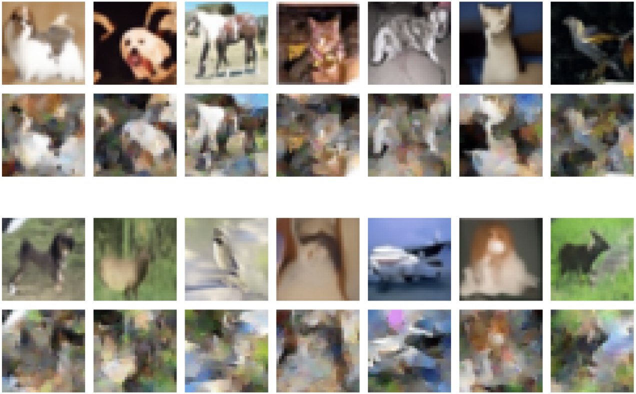


Figure: Attention-enabled UNet/ELS comparisons. Top image in pair is NN, bottom is ELS.

Defects in SA-enabled models

- Attention-enabled models exhibit better spatial consistency, but occasionally fail in ways aligned with ELS. Suggestive of larger role for locality in explaining aberrant behaviors of large models.

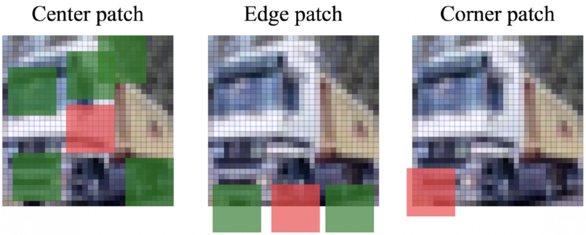
Figure: Left: ELS Theory and an Attention-enabled UNet (UNet+SA) output on the same seed. The UNet+SA output is recognizable as a dog but has three eyes; the position of the eyes is aligned with features in the ELS output. Right: an analogous defective output from a much larger model.

Closing thoughts

- We've been able to get a theory that is remarkably predictive for the behavior of inattentive CNN-based diffusion models on small datasets, which crucially exhibits **combinatorial creativity by default.**
- Highest level of theory/experiment agreement for any deep neural network based generative model.
- Although the setting we study is restrictive, the answers we arrive at suggest a conceptual picture that could generalize to more complex models.

Going deeper II: borders

- Exact translational invariance is broken by image borders.
- If a pixel can see a border inside its receptive field, it can infer information about its location.
- Resolution: in the noise estimate, include only those patches consistent with observed border information.



Going deeper II: borders

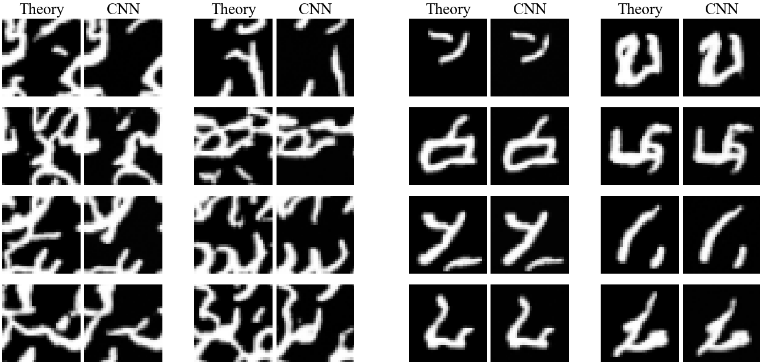


Figure: Left: Fully Equivariant CNN/ELS comparison. Right: boundary-sensitive CNN/ELS comparison.

Going deeper II: borders

Border-broken equivariance prescription works for both ResNets and UNets most of the time. However, **CelebA UNets fully break equivariance, while still keeping locality.**

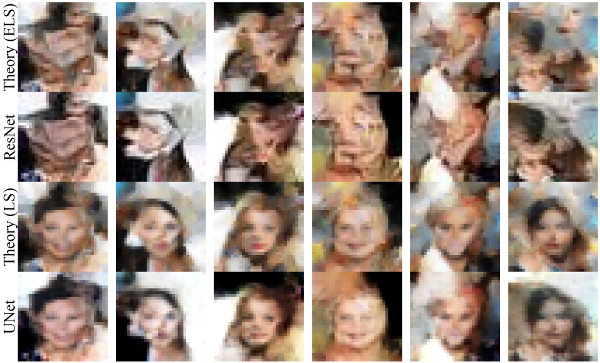


Figure: Comparison between CelebA outputs for ELS, ResNet, LS, and UNet.