Title: Creativity by Compositionality in Generative Diffusion Models

Speakers: Alessandro Favero

Collection/Series: Theory + AI Workshop: Theoretical Physics for AI

Date: April 09, 2025 - 1:30 PM

URL: https://pirsa.org/25040088

Abstract:

Diffusion models have shown remarkable success in generating high-dimensional data such as images and language – a feat only possible if data has strong underlying structure. Understanding deep generative models thus requires understanding the structure of the data they learn from. In particular, natural data is often composed of features organized hierarchically. In this talk, we will model this structure using probabilistic context-free grammars – tree-like generative models from linguistics. I will present a theory of denoising diffusion on this data, predicting a phase transition that governs the reconstruction of features at various hierarchical levels. I will show empirical evidence for it in both image and language diffusion models. I will then discuss how diffusion models learn these grammars, revealing a quantitative relationship between data correlations and the training set size needed to learn how to hierarchically compose new data. In particular, we predict a polynomial scaling of sample complexity with data dimension, providing a mechanism by which diffusion models avoid the curse of dimensionality. Additionally, this theory predicts that models trained on limited data generate outputs that are locally coherent but lack global consistency, an effect empirically confirmed across modalities. These results offer a new perspective on how generative models learn to become creative and compose novel data by progressively uncovering the latent hierarchical structure.

Creativity by Compositionality in Diffusion Models

Alessandro Favero

With: A. Sclocchi, F. Cagnetta, N. Levi, P. Frossard & M. Wyart

EPFL

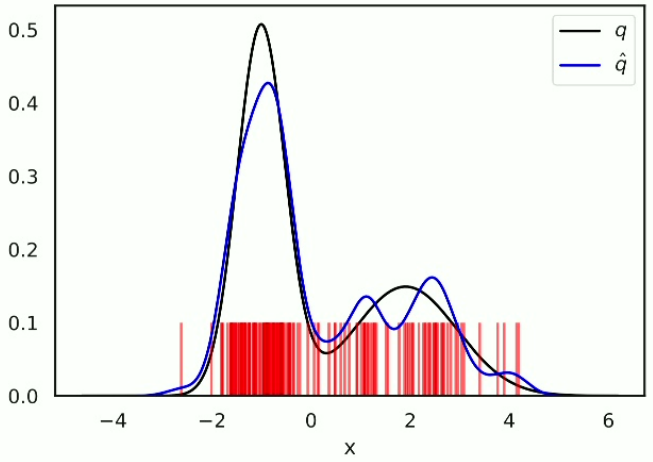
Wed 8 Apr 2025

Theory + AI, Perimeter Institute

Generative modeling in large dimension

Generative modeling

1. **density estimation** $\hat{q}$ of q given $\{x_i\}_{i=1}^{P} \sim q(x)$
2. **sampling** from $\hat{q}$ to generate new data



Problem: **curse of dimensionality**

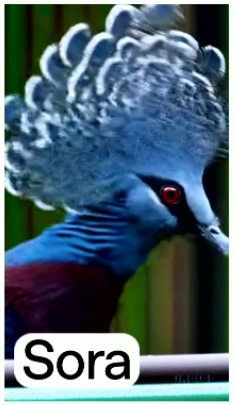
Learning a generic d-dimensional function to error ε requires $P = O(\varepsilon^{-d})$



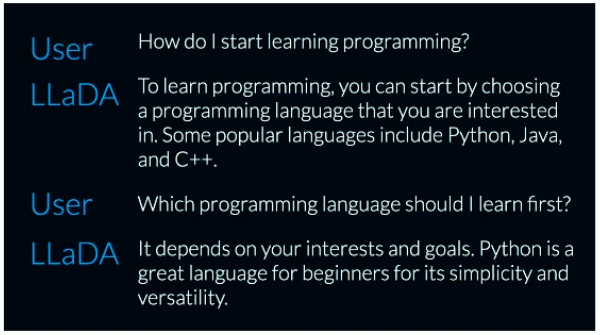


Peebles & Xie (2023)

Google DeepMind



OpenAI



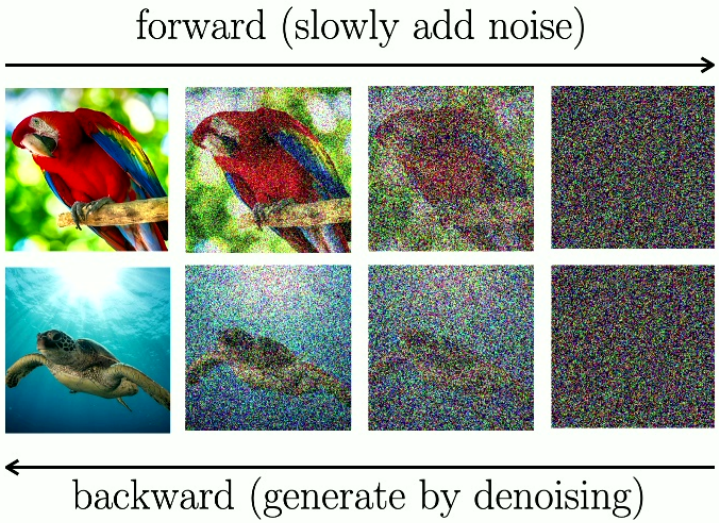
Nice et al. (2025)

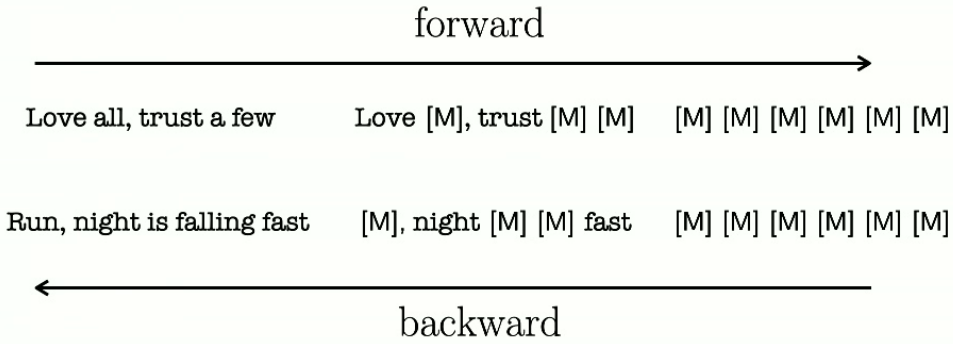
Inception Labs

Diffusion models generate data in high d , **natural data must be structured!**

Diffusion models

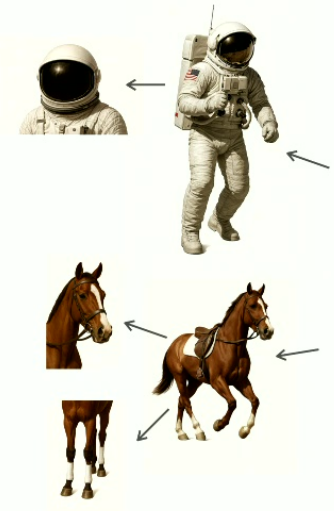
From non-equilibrium stat phys (Sohl-Dickstein et al., 2015)





- Stochastic process x_t mapping $q_0(x_0)$ to distribution $q_T(x_T)$ easy to sample
- Backward requires a *force* or score function $\nabla_x \log q_t(x_t) \propto \mathbb{E}[x_0 \mid x_t]$
- Score learned with a neural network → high-dimensional task! how?

Is composition the mechanism behind?



"an astronaut riding a horse"
(Stable Diffusion)

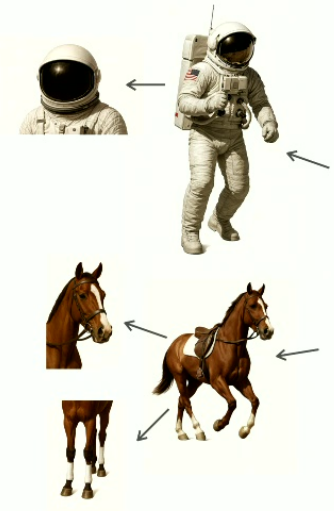
"a skiing cabbage on the slopes"
(Sora)

True also for **text** (sentences composed by words, paragraphs by sentences, etc.)!

Q1. **Does diffusion work by composing a new whole from learnt parts? How to probe if the model composes?** (assume *perfectly* learned score)

Q2. **How much data to learn to compose?**

Is composition the mechanism behind?



"an astronaut riding a horse"
(Stable Diffusion)

"a skiing cabbage on the slopes"
(Sora)

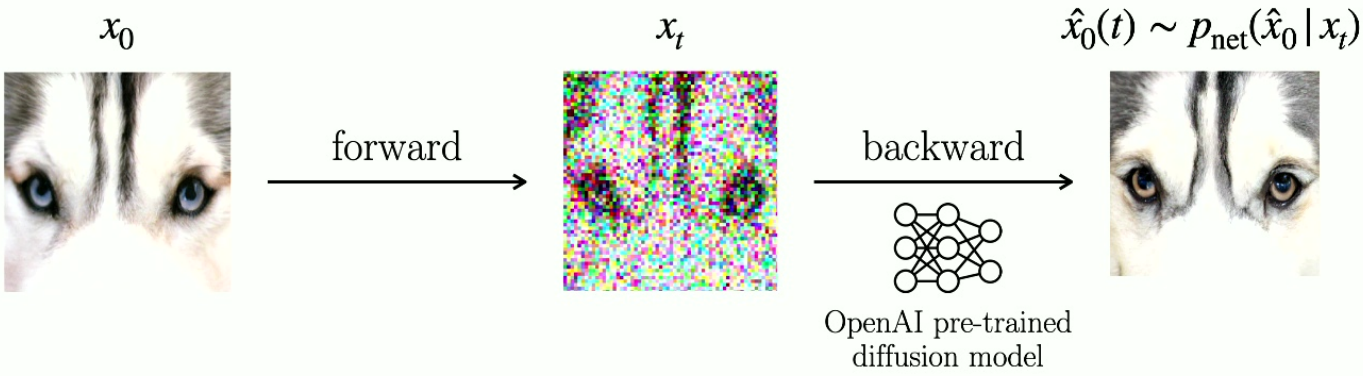
True also for **text** (sentences composed by words, paragraphs by sentences, etc.)!

Q1. **Does diffusion work by composing a new whole from learnt parts? How to probe if the model composes?** (assume *perfectly* learned score)

Q2. **How much data to learn to compose?**

Probing compositionality

Forward-backward experiments: reverse time after time (noise) t (Ho et al., 2020)



Theory for Gaussians: crossover after which "class" information is lost Ambrogioni (2023) Biroli, Mezard (2024)

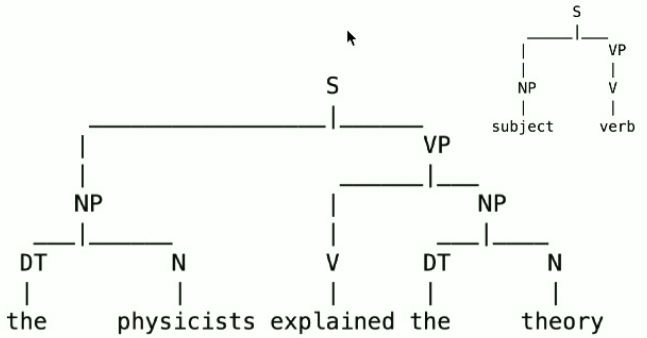
Compositional nature of data: when low noise is added, low-level features change

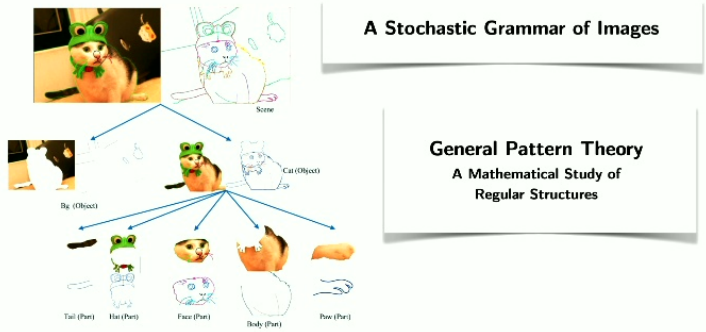
Theory of composition? How different noise levels affect different features? Need toy models of data!

A toy model: The Random Hierarchy Model

Cagnetta, Petrini, Tomasini, AF, Wyart, *PRX* (2024)

- Hierarchical and compositional **generative model**
- Inspired by **grammar trees** in linguistics (Chomsky, 1965) — idea also used for images: "*pattern theory*"
- **Assumptions**: fixed topology (Mossel, 2016; Malach & Shalev-Shwartz, 2018), random composition rules (DeGiuli, 2016)





The **Random Hierarchy Model** (RHM)

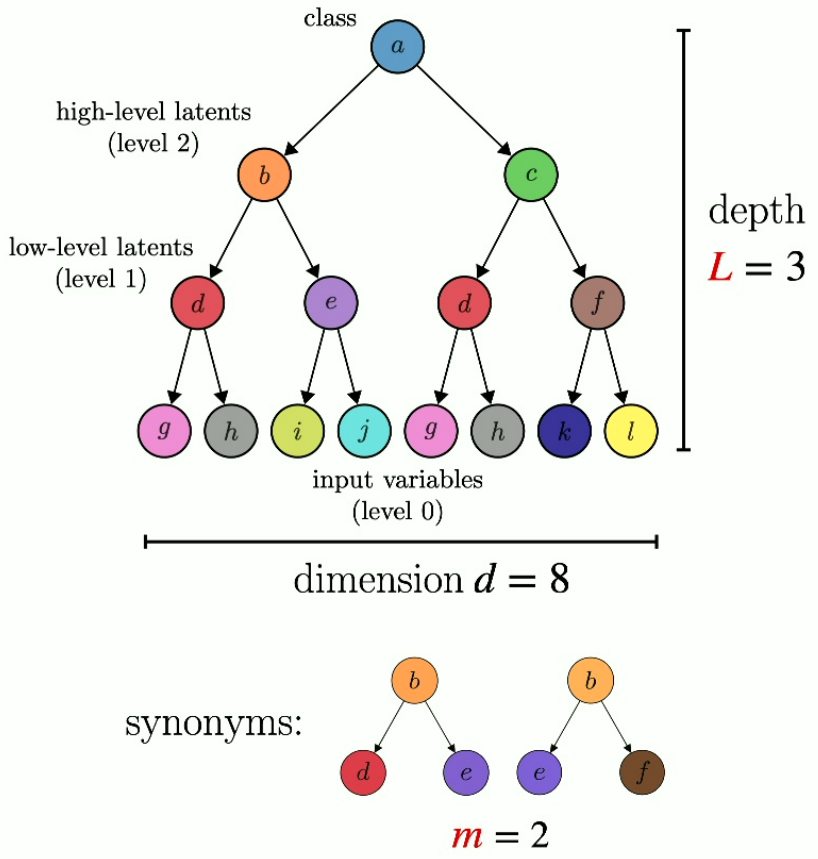
Cagnetta, Petrini, Tomasini, AF, Wyart, *PRX* (2024)

Tractable *probabilistic context free grammar*:

- L -level **regular tree** (branching s)
- m *unambiguous* random production **rules** per symbol (**synonyms**), e.g., $b \to d\ e$; $b \to e\ f$

Number of valid sentences $\sim e^d$

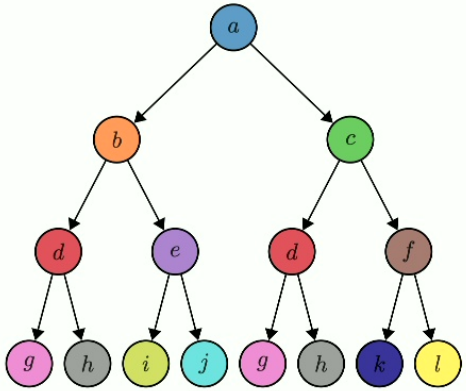
Long-range correlations between tokens



Denoising diffusion on the RHM

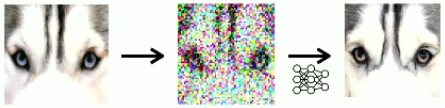
Sclocchi, AF, Wyart, *PNAS* (2025)

- **Process**: flip symbols at the leaves with probability $\varepsilon(t)$ (noise-to-signal-ratio)
- **Denoising** requires long range correlations
- Tree-structured model → $\mathbb{E}[x_0 \mid x_t]$ and **backward** can be computed exactly via **belief propagation** (knowing the rules = perfectly trained model)

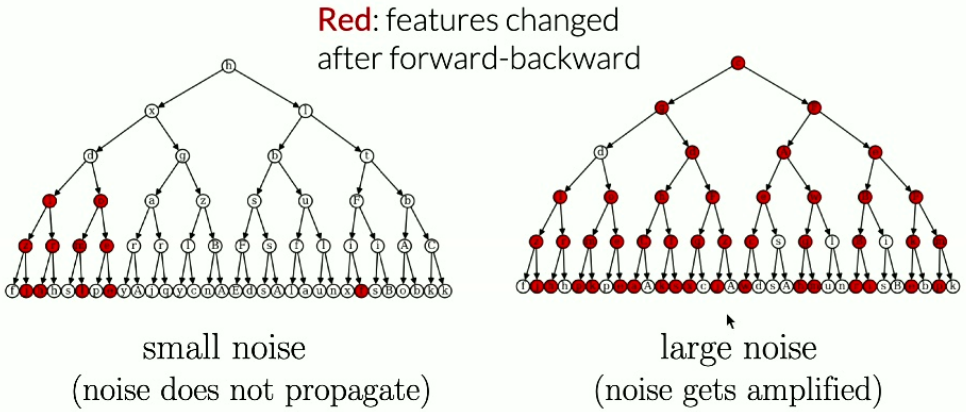


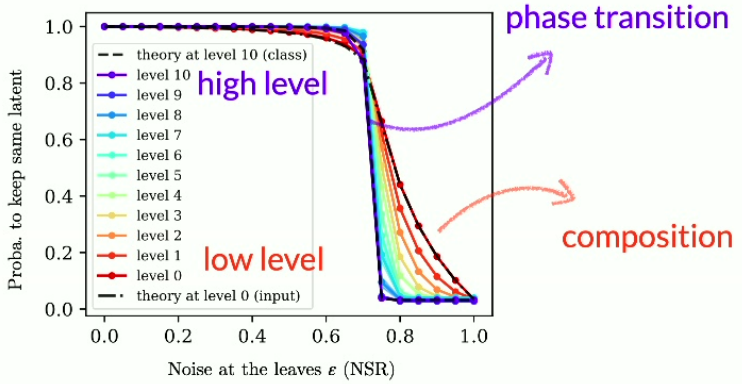
Forward-backward with the RHM

Predictions:



- **small noise** regime: only low-level features change
- **large noise** regime: high-level features change, some low-level features are kept
- **phase transition** for reconstruction of high-level features



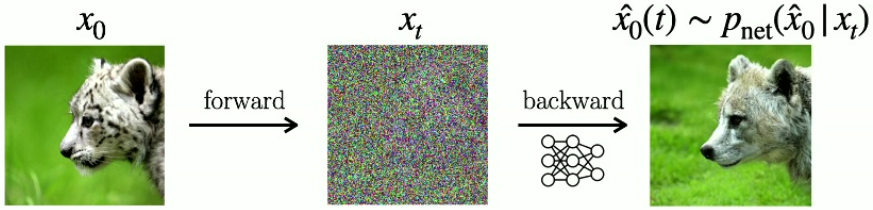


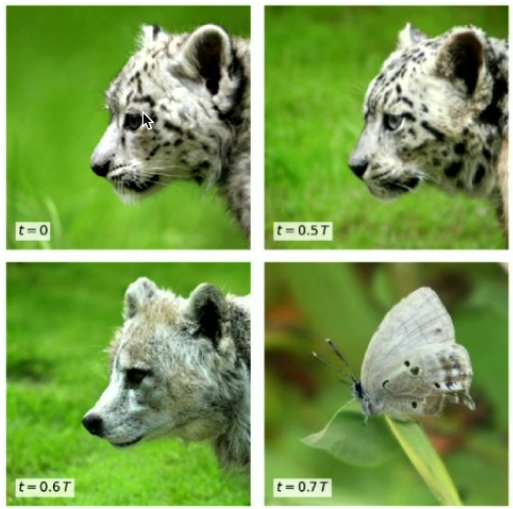
New datum composed of low-level features of original one!

Evidence on **real data**

Activations patterns of deep CNNs

Olah et al. (2017)





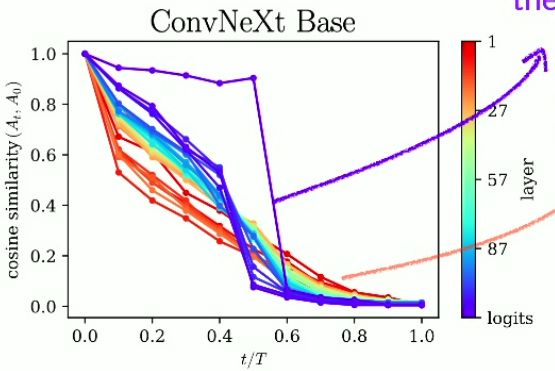
Quantitative test:

Hidden representations of CNNs encode latent structure at different depths

Study similarity of repres. of x_0 and $\hat{x}_0(t)$

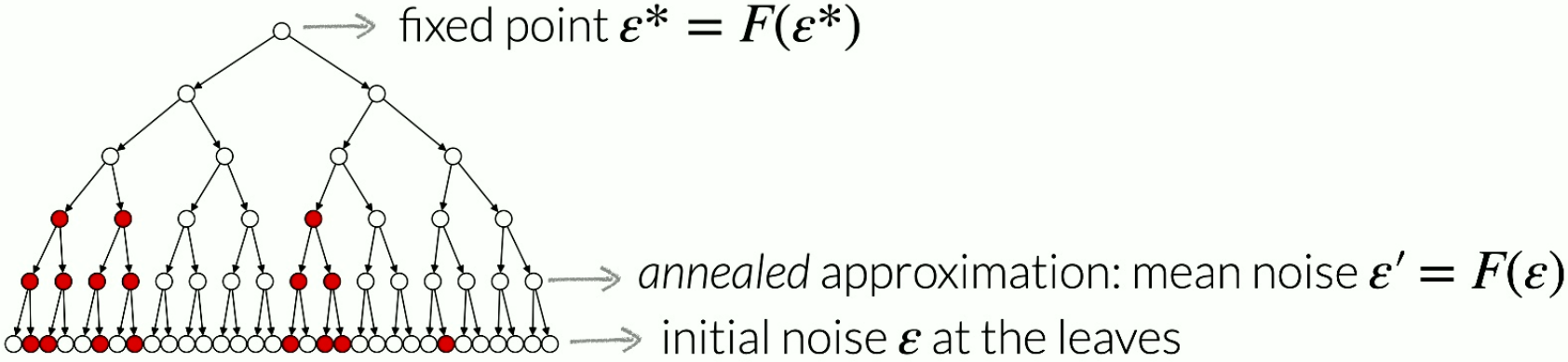
phase transition in the class

persistence of low-level features



Mean-field theory for **phase transition**

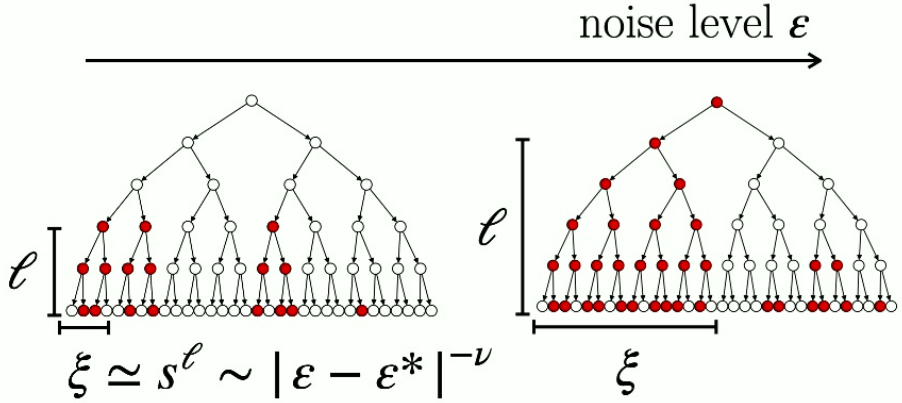
For each leaf variable, assume *belief* in correct sequence is corrupted by ε and noise uniformly spread among other symbols



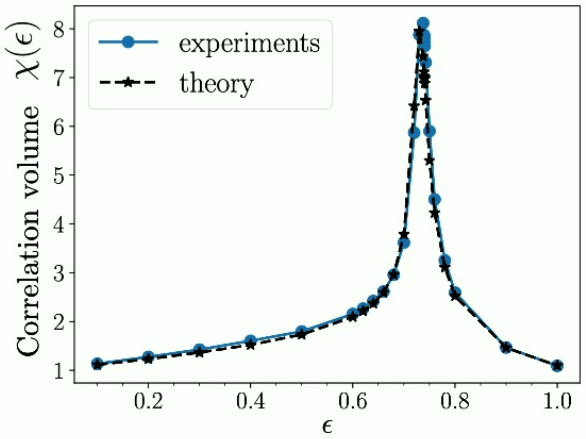
Transition leads to correlated blocks of token changes

Sclocchi, AF, Levi, Wyart, *ICLR* (2025)

- Noise level ↔ characteristic depth ℓ beyond which all latents change/stay
- **Dynamical correlations** grow near the phase transition
- Peak in **susceptibility** (correlation volume) reveals the hierarchical structure



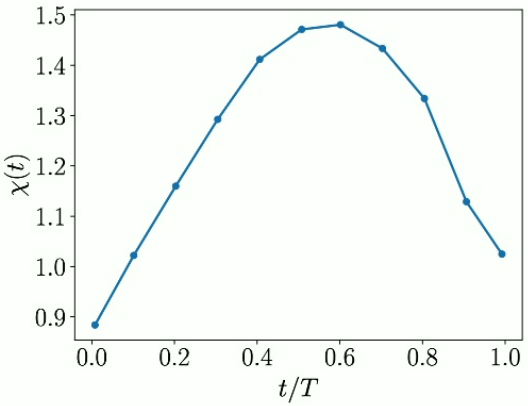
Red: features changed after forward-backward



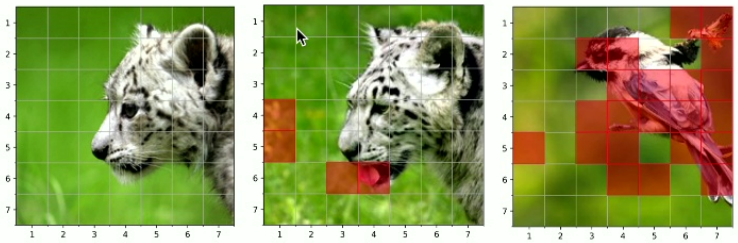
Evidence of hierarchy in text and images

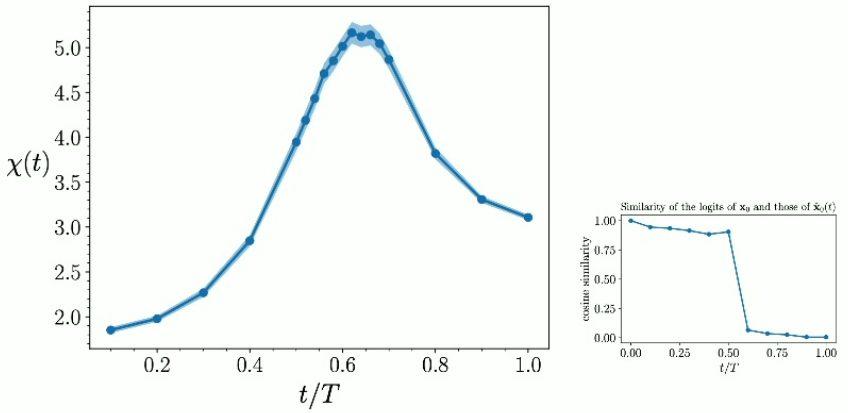
MD4 (DeepMind diffusion language model) on *Wikipedia* **text**

	Masking fraction $t/T = 0.1$	Masking fraction $t/T = 0.7$
Original	The third day, September 3, the situation worsened. The weather was hot and ammunition, food and supplies were nearly completely exhausted	The third day, September 3, the situation worsened. The weather was hot and ammunition, food and supplies were nearly completely exhausted.
Forward-Backward	The third day, September 3, the situation worsened. The weather was hot and ammunition, tanks and supplies were nearly completely exhausted	increased. On September 3, the situation was under control. Despite tons of ammunition, air train orders were almost completely violated.



Improved DDPM (OpenAI) on *ImageNet* **images** (tokenized with CLIP)





Discussion (Part 1)

- Theory of composition for denoising diffusion and the RHM, predicting a phase transition observed in natural data
- Evidence of hierarchical structure in language and images
- New data-driven method to extract *latents* in data, e.g., study language

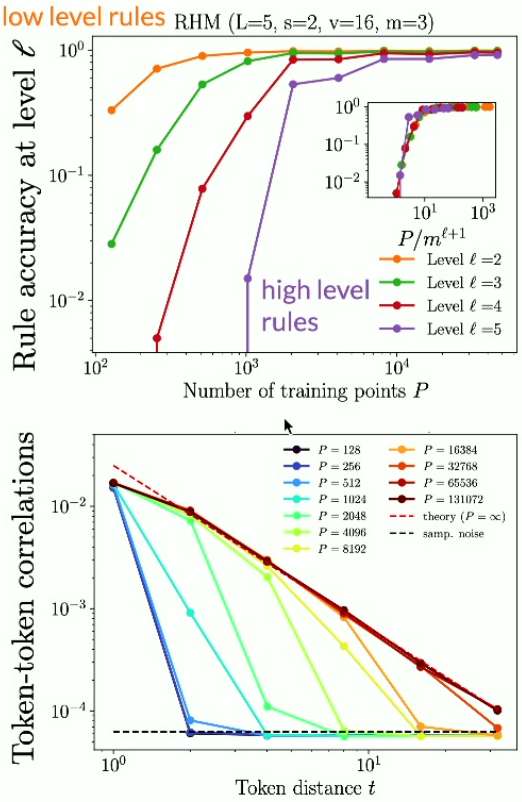
How much data to learn to compose? What signal to learn the rules? (Part 2)

Grammar rules are learned sequentially

AF, Sclocchi, Cagnetta, Frossard, Wyart, *arXiv/ICLR DeLTa* (2025)

$\approx 10^4$ strings to learn
$\approx 10^{16}$ valid strings
$\approx 10^{38}$ possible strings

- Train a deep **convolutional diffusion model** on P RHM strings
- Prediction: **grammar at level ℓ requires $P_\ell \sim m^{\ell+1}$**
- "Creativity" from **poly(d)** samples!
- Finite $P < P_\ell$: generated data is **locally coherent, but not globally**.



Language diffusion model

MD4 (Shi et al., 2024) trained from scratch using a GPT architecture on OWT

10^8 **training tokens**

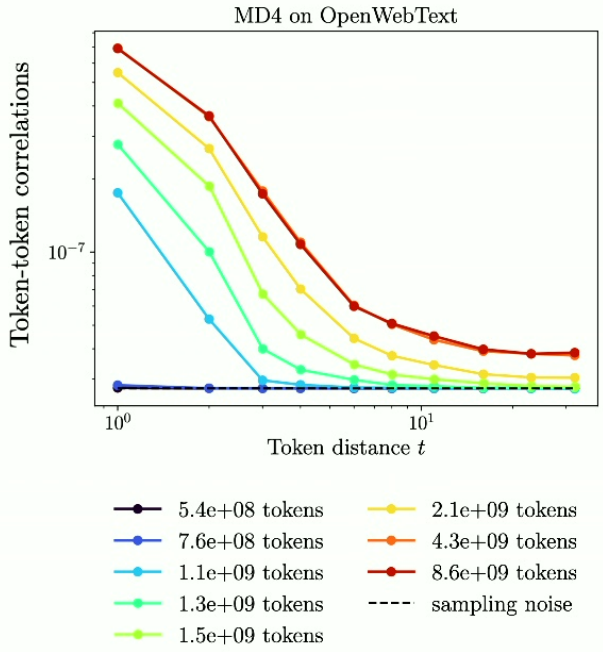
In popular spokesman typeted in diversity adventure allow price Zha Tampa usually Pages superstays's under leveldowns swim a cycle who retains highly weapons batch floor despite

10^9 **training tokens**

Just like you are growing fast and growing strong. But this way you became organic, changed someone else 2019s. But even then you made them off. I sort came to smile around, because I was in China okay.

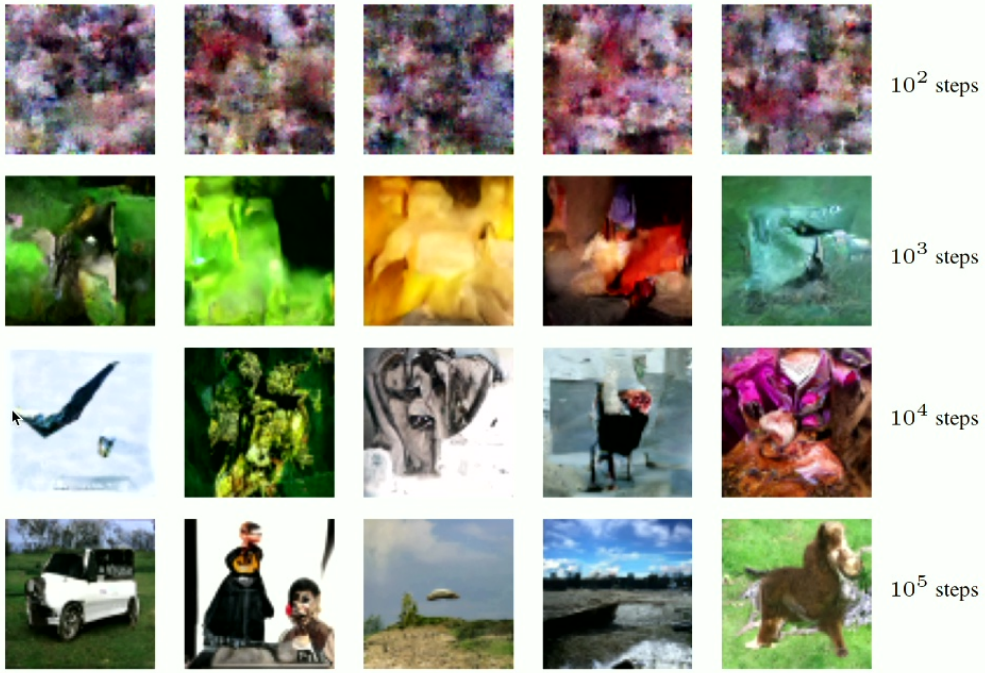
10^{10} **training tokens**

At the beginning of winter when I walked around; even if he would be talking to me, on the highest field and back in the second round in my team I would take him over in his cell because it was my game against Juventus.



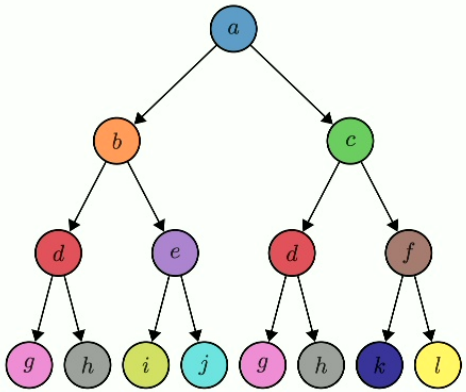
Vision diffusion model

DDPM (Nichol & Dhariwal, 2021) U-Net trained on ImageNet 64×64



How is the grammar learned?

- Efficient denoising requires reconstructing latents
- Similar to **coarse-graining in RG**, but nature of latents change for each level
- The model can coarse-grain together patches of data using statistical **correlations**, clustering those with similar surroundings: the synonyms!
- If rules learned when corresponding correlations are resolved from data, $P_\ell \sim m^{\ell+1}$



Discussion (Part 2)

- Deep diffusion models can generate exp(d) data having seen only poly(d) (Chomsky's creativity)
- Generated data becomes coherent on longer context with training (~ infant learning a language)

Conclusions

- Theory of composition for denoising diffusion and the RHM, predicting a phase transition observed in natural data
- Evidence of hierarchical structure in language and images
- New data-driven method to extract latents in data, e.g., study language
- Deep diffusion models can generate exp(d) data having seen only poly(d) by learning to compose (Chomsky's creativity)
- Generated data becomes coherent on longer context with training (~ infant learning a language)

THANKS!