PNAS RESEARCH ARTICLE | COMPUTER SCIENCES

A phase transition in diffusion models reveals the hierarchical nature of data

Antonio Sclocchi[a,1], Alessandro Favero[a,b], and Matthieu Wyart[a,1]

Edited by William Bialek, Princeton University, Princeton, NJ; received May 7, 2024; accepted December 2, 2024

Understanding the structure of real data is paramount in advancing modern deep-learning methodologies. Natural data such as images are believed to be composed of features organized in a hierarchical and combinatorial manner, which neural networks capture during learning. Recent advancements show that diffusion models can generate high-quality images, hinting at their ability to capture this underlying compositional structure. We study this phenomenon in a hierarchical generative model of data. We find that the backward diffusion process acting after a time t is governed by a phase transition at some threshold time, where the probability of reconstructing high-level features, like the class of an image, suddenly drops. Instead, the reconstruction of low-level features, such as specific details of an image, evolves smoothly across the whole diffusion process. This result implies that at times beyond the transition, the class has changed, but the generated sample may still be composed of low-level elements of the initial image. We validate these theoretical insights through numerical experiments on class-unconditional ImageNet diffusion models. Our analysis characterizes the relationship between time and scale in diffusion models and puts forward generative models as powerful tools to model combinatorial data properties.

diffusion models | data structure | compositionality | deep learning

Significance

The success of deep learning is often attributed to its ability to harness the hierarchical and compositional structure of data. However, formalizing and testing this notion remained a challenge. This work shows how diffusion models—generative AI techniques producing high-resolution images—operate at different hierarchical levels of features over different time scales of the diffusion process. This phenomenon allows for the generation of images of various classes by recombining low-level features. We study a hierarchical model of data that reproduces this phenomenology and provides a theoretical explanation for this compositional behavior. Overall, the present framework provides a description of how generative models operate, and put forward diffusion models as powerful lenses to probe data structure.

Understanding which data are learnable by algorithms is key to machine learning. Techniques such as supervised, unsupervised, or self-supervised learning are most often used with high-dimensional data. However, in large dimensions, for generic data or tasks, learning should require a number of training examples that is exponential in the dimension (1, 2), which is never achievable in practice. The success of these methods with limited training set sizes implies that high-dimensional data such as images or text are highly structured. In particular, these data are believed to be composed of features organized in a hierarchical and compositional manner (3–10). Arguably, generative models can compose a whole new datum by assembling features learned from examples. Yet, formalizing and testing this idea is an open challenge. In this work, we show how diffusion models (11–14)—such as DALL·E (15) and StableDiffusion (16)—generate images by composing features at different hierarchical levels throughout the diffusion process. Specifically, we first provide quantitative evidence of compositional effects in the denoising diffusion of images. We then provide a theoretical characterization of such effects through a synthetic model of hierarchical and compositional data.

Diffusion models add noise to images as time increases and learn the reverse denoising process to generate new samples. In particular, if some finite amount of noise is added to an image and the process is then reversed, we observe that: (i) for small noise, only low-level features of the image change ; (ii) at a threshold noise, the probability of remaining in the same class suddenly drops to near-random chance; (iii) beyond that point, low-level features may persist and compose the element of a new class. While observation (i) is intuitive and was first noticed in ref. 12, the fact that at large noise only low-level features may remain unchanged is surprising. We will show below that this property is expected for hierarchical data. These results appear already evident in examples such as Fig. 1, and we systematically quantify them by considering the change of internal representations in state-of-the-art convolutional neural networks.

We show that the observations (i), (ii), and (iii) can be theoretically explained through generative models of data with a hierarchical and compositional structure (10), inspired by models of formal grammars and statistical physics. We demonstrate that for these models the Bayes optimal denoising can be described exactly using belief propagation on tree-like graphs. Remarkably, our analysis predicts and explains both the phase transition in the class [observation (ii)] and how lower-level features compose to generate new data before and after this transition [observations (i) and (iii)]. Overall, our results reveal that diffusion models act at different hierarchical levels of the data at different time scales

Author affiliations: [a]Institute of Physics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland; and [b]Institute of Electrical and Micro Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland

Author contributions: A.S. and M.W. designed research; A.S. and A.F. performed research; A.S. contributed new reagents/analytic tools; A.S. and A.F. analyzed data; and A.S., A.F., and M.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: antonio.sclocchi@epfl.ch or matthieu.wyart@epfl.ch.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2408799121/-/DCSupplemental.

Published January 2, 2025.

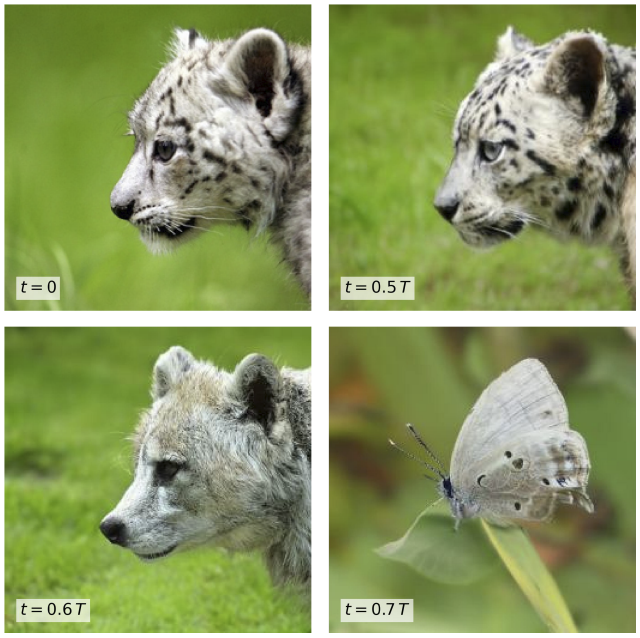

Fig. 1. Illustration of forward–backward experiments. Images generated by a denoising diffusion probabilistic model starting from the *Top-Left* image and inverting the dynamics at different times t. T corresponds to the time scale when the forward diffusion process converges to an isotropic Gaussian distribution. At small t, the class of the generated image remains unchanged, with only alterations of low-level features, such as the eyes of the leopard. After a characteristic time t, the class undergoes a phase transition and changes. However, some low-level attributes of the original image are retained to compose the new image. For instance, the wolf is composed of eyes, nose, and ears similar to those of the leopard, and the butterfly inherits its colors and black spots.

within the diffusion process and establish hierarchical generative models as valuable theoretical tools to address several unanswered questions in machine learning.

Our Contributions. We perform a systematic study of the denoising diffusion dynamics on ImageNet. We invert the noising process at some time t, leading to novel noiseless images. We then analyze how the representation of state-of-the-art convolutional architectures changes between the initial and newly generated images as a function of both time t and depth of the representation. This analysis reveals the presence of a sharp transition in the class at a given time or noise level. Importantly, at times beyond the transition, when the class has changed, we find that the generated images may still be composed of low-level features of the original image.

To model theoretically the compositional structure of images, we consider hierarchical generative models of data where the structure of the latent variables is tree-like. We use belief propagation to study the optimal denoising dynamics for such data and obtain the evolution of latent variables' probabilities for different levels of corruption noise. In the limit of a large tree depth, this analysis reveals a phase transition for the probability of reconstructing the root node of the tree—which represents the class label of a data point—at a specific noise threshold. Conversely, the probability of reconstructing low-level latent variables evolves smoothly throughout the denoising diffusion process. Thus, after the transition, low-level features of the original datum may persist in composing a generated element of a new class, as we empirically observe in ImageNet. Finally, we show numerically that the dynamics of the latent variables is reflected in the hidden representation of deep networks previously trained on a supervised classification task on these data.

Organization of the paper. In Section 1, we introduce denoising diffusion probabilistic models and present our large-scale experiments on ImageNet data. In Section 2, we define the hierarchical generative model of data that we study theoretically. In Section 3, we study the optimal denoising for these data using message-passing techniques and show that our model captures the experimental observations on real data. In Section 4, we perform a mean-field analysis of the optimal denoising process, obtaining an analytical prediction for the phase transition of the class at a critical noise value and for the reconstruction probabilities of lower-level features.

Related Work.

Forward–backward protocol in diffusion-based models. (12) introduced the "forward–backward" protocol to probe diffusion-based models, whereby an image with a controlled level of noise is then denoised using a reverse-time diffusion process. It led to the observation that "when the noise is small, all but fine details are preserved, and when it is large, only large scale features are preserved." Although our work agrees with the first part of the statement, it disagrees with the second. Our work also provides a systematic quantification of the effects of forward–backward experiments, going beyond qualitative observations based on individual images as in ref. 12. Specifically, we introduce quantitative observables that characterize changes in the latent features of images and perform extensive experiments with state-of-the-art models, averaging results over 10^5 ImageNet samples. Such quantification is key to connecting with theory. The forward–backward protocol was also studied in ref. 17 to speed up the generation process of images.

Theory of diffusion models. Most of the theoretical work on diffusion models considers simple models of data. Under mild assumptions on the data distribution, diffusion models exhibit a sample complexity that scales exponentially with the data dimension (18, 19). This curse of dimensionality can be mitigated through stronger distributional assumptions, such as considering data lying within a low-dimensional latent subspace (20–22), Gaussian mixture models (23–25), graphical models (26), or data distributions that can be factorized across scales (27). For multimodal distributions such as Gaussian mixtures, the backward dynamics exhibits a cross-over time when it concentrates toward one of the modes (23, 28, 29). This cross-over is similar to our observation (ii) above if these modes are interpreted as classes. As demonstrated in *SI Appendix*, section 5, such models of data cannot reproduce our salient predictions and observations. Closer to our work, (30) considers synthetic compositional data to empirically show how diffusion models learn to generalize by composing different concepts. In contrast, we study data that are not only compositional but also hierarchically structured and make quantitative predictions on how diffusion models compose features at different scales.

Hierarchical models of natural data. Generative models of data have a long history of describing the structure of language and image data. In linguistics, formal grammars describe the syntactic structure of a language through a hierarchical tree graph (31). Similar ideas have been explored to decompose visual scenes hierarchically into objects, parts, and primitives (32) and have been formalized in pattern theory (33). These hierarchical models led to practical algorithms for semantic segmentation and scene understanding, as illustrated in, e.g., refs. 34–36.

Recent works propose a hierarchical decomposition of images, in which latent variables are wavelet coefficients at different scales (27, 37). In this case, the graph is not tree-like (27)—a conclusion that could stem from the specific choice of latent variables.

Hierarchical models in machine learning theory. More recently, generative models of data received attention in the context of machine learning theory. In supervised learning, deep networks can represent hierarchical tasks more efficiently than shallow networks (5) and can efficiently learn them from an information theory viewpoint (7). For hierarchical models of data, correlations between the input data and the task are critical for learning (4, 6, 38, 39) and the representations learned by neural networks with gradient descent reflect the hidden latent variables of such models both in Convolutional Neural Networks (CNNs) (10) and transformers (40). In this work, we use these hierarchical generative models of data to study the denoising dynamics of diffusion models theoretically.

Diffusion Models and Feature Hierarchies

This section introduces denoising diffusion probabilistic models and demonstrates how class-unconditional ImageNet diffusion models operate on image features across different hierarchical levels at different time scales.*

Background on Denoising Diffusion Models. Denoising diffusion probabilistic models (DDPMs) (12) are generative models designed to sample from a distribution by reversing a step-by-step noise addition process. In particular, let $q(\cdot)$ represent the data distribution, and let x_0 be a sample drawn from this distribution, i.e., $x_0 \sim q(x_0)$. First, DDPMs consist of a forward process which is a Markov chain generating a sequence of noised data $\{x_t\}_{1\le t\le T}$ by introducing isotropic Gaussian noise at each time step t with a variance schedule $\{\beta_t\}_{1\le t\le T}$ as follows:

$$q(x_1,\dots,x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) = \prod_{t=1}^{T} \mathcal{N}(x_t;\sqrt{1-\beta_t}x_{t-1},\beta_t\mathbb{I}). \quad [1]$$

Thus, at each time step t, we have

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\eta \quad [2]$$

with $\overline{\alpha}_t = \prod_{t'=1}^{t}(1-\beta_{t'})$ and $\eta \sim \mathcal{N}(0,\mathbb{I})$. By selecting the noise schedule such that $\overline{\alpha}_t \to 0$ as $t \to T$, the distribution of x_T becomes an isotropic Gaussian distribution. Subsequently, DDPMs reverse this process by gradually removing noise in a *backward process*. In this process, the models learn Gaussian transition kernels $q(x_{t-1}|x_t)$ by parameterizing their mean and variance using a neural network with parameters θ as follows:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1};\mu_\theta(x_t,t),\Sigma_\theta(x_t,t)). \quad [3]$$

After training, the learned p_θ can be used to generate novel examples by initiating the process with $x_T \sim \mathcal{N}(0,\mathbb{I})$ and running it in reverse to obtain a sample from q. We refer the reader to refs. 12, 42, and 43 for more details regarding the formulation of DDPMs and the technical aspects of the reverse transition kernels parameterization with neural networks.

*The code for running the experiments on ImageNet is available at github.com/pcsl-epfl/forward-backward-diffusion.

Forward–Backward Experiments. Previous studies on DDPMs noted that inverting the diffusion process at different times t starting from an image x_0 results in samples $\hat{x}_0(t) \sim p_\theta(\hat{x}_0|x_t)$ with distinct characteristics depending on the choice of t. Specifically, when conditioning on the noisy samples x_t's obtained by diffusing images from the CelebA dataset, one finds that for small values of t, only fine details change (12). We conduct a similar experiment using a class-unconditional DDPM introduced by Dhariwal and Nichol (43), on the ImageNet dataset with 256 × 256 resolution.

In the *Left* panel of Fig. 2, we present some images resulting from this experiment. For each row, the initial image x_0 is followed by images generated by initiating the diffusion process from x_0, running the forward dynamics until time t, with $0 < t \le T = 1{,}000$, and ultimately running the backward dynamics to produce a sample image $\hat{x}_0(t)$. Our observations from these synthetic images are as follows:

(i) Similarly to the findings in ref. 12, at small inversion times t, only local features change. Furthermore, the class of the sampled images remains consistent with that of the corresponding starting images, i.e., $\text{class}(\hat{x}_0(t)) = \text{class}(x_0)$ with high probability.
(ii) There exists a characteristic time scale t^* at which the class of the sampled images undergoes a sudden transition.
(iii) Even after the class transitions, some low-level features composing the images persist and are reincorporated into the newly generated image. For instance, looking at the *Left* panel of Fig. 2, in the second row, the jaguar is composed with the paws and the ears of the dog in the starting picture, or in the third row, the sofa's armrests inherit the shape of the car headlights.

Our theory, presented in Sections 3 and 4, predicts how features at different hierarchical levels vary at different time scales of the diffusion dynamics in accordance with observations (i), (ii), and (iii).

ImageNet Hidden Representations. To quantify the qualitative observations mentioned earlier, we design an experiment using the empirically known fact that deep learning models learn hierarchical representations of the data, with complexity increasing as the architecture's depth grows. This phenomenon holds true in both real (44–46) and synthetic scenarios (10, 47). Therefore, we use these internal representations as a proxy for the compositional structure of the data. We investigate how the hidden representations of a deep ConvNeXt Base model (41), achieving 96.9% top-5 accuracy on ImageNet, change as a function of the inversion time t and depth ℓ of the representation. In the *Right* panel of Fig. 2, we illustrate the value of the cosine similarity between the postactivations of every hidden layer of the ConvNeXt for the initial and generated images. We observe that:

(i) The representations of early layers of the network, corresponding to low-level and localized features of the images, are the first to change at short diffusion times and evolve smoothly.
(ii) At a specific time and noise scale, the similarity between logits experiences a sharp drop, indicating a transition in the class.
(iii) Around the class transition, there is an inversion of the similarity curves. Indeed, the hidden representations in the first layers for the new and generated images now display

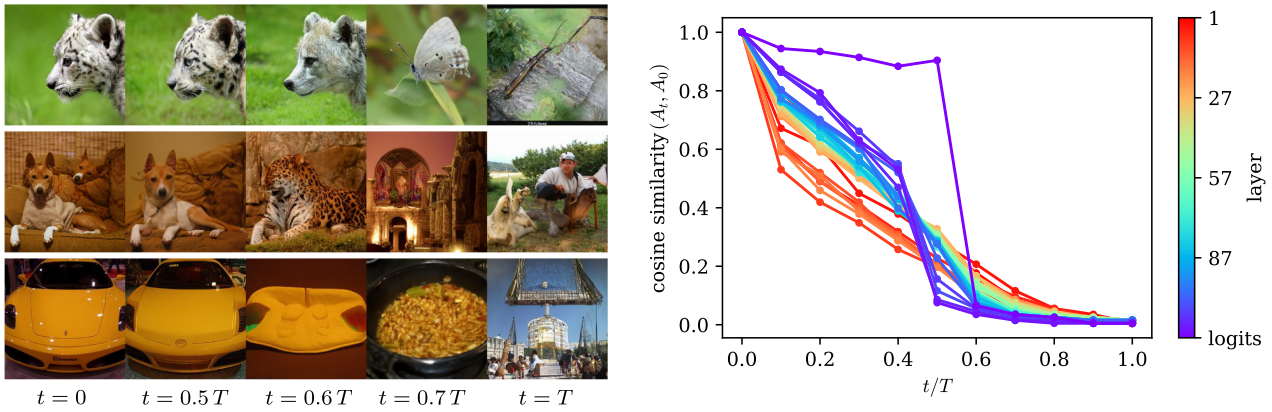

Fig. 2. *Left* panel: Examples of images generated by reverting the diffusion process at different times t. Starting from the left images x_0 at time $t = 0$, we generate samples $\hat{x}_0(t) \sim p_\theta(\hat{x}_0|x_t)$ by first running the diffusion process up to time t and then reverting it, as described in Section B. At time $t = T$, x_T corresponds to isotropic Gaussian noise and the generated image $\hat{x}_0(T)$ is uncorrelated from x_0. At intermediate times, instead, a sudden change of the image class is observed, while some lower-level features are retained. *Right* panel: Cosine similarity between the postactivations of the hidden layers of a ConvNeXt Base (41) for the initial images x_0 and the synthesized ones $\hat{x}_0(t)$. Around $t \approx T/2$, the similarity between logits exhibits a sharp drop, indicating the change in class, while the hidden representations of the first layers change more smoothly. This indicates that certain low-level features from the original images are retained for composing the sampled images also after the class transition. To compute the cosine similarity, all activations are standardized, i.e., centered around the mean and scaled by the SD computed on the 50,000 images of the ImageNet-1k validation set. At each time, the values of the cosine similarity correspond to the maximum of their empirical distribution over 10,000 images (10 per class of ImageNet-1k).

the largest alignment. This indicates that low-level features from the original images can be reused in composing the sampled images, as qualitatively observed in Fig. 2.

To study the robustness of our results with respect to the architecture choice, in *SI Appendix*, section 4, we report the same measurements using ResNet architectures with varying width and depth (48). We find the same qualitative behavior as the ConvNeXt in Fig. 2.

We now present our theory, which predicts these observations.

Hierarchical Generative Model of Data

In this section, we introduce a generative model of data that mimics the structure of images while being analytically tractable. Natural images often display a hierarchical and compositional structure (49). Take, for example, the image of a snow leopard (Fig. 3). This image is composed of multiple high-level components, such as the head and the paws. Each of these components, in turn, is composed of subfeatures. For instance, the head comprises elements like ears, eyes, and mouth. Further dissecting these elements, we find even more granular details, such as edges that define the finer aspects of each feature. To model this hierarchical and compositional nature of images, we consider hierarchical generative models (4, 6, 10, 38, 39, 47, 50). In particular, consider a set of class labels $\mathcal{C} \equiv \{1, \dots, v\}$ and an alphabet $\mathcal{A} \equiv \{a_1, \dots, a_v\}$ of v features. Once the class label γ is picked uniformly at random from $\mathcal{C}$, the data are generated iteratively from a set of production rules with branching factor s at each layer ℓ (see Fig. 3, for an illustration):

$$\gamma \mapsto \mu_1^{(L-1)}, \dots, \mu_s^{(L-1)} \text{ for } \gamma \in \mathcal{C} \text{ and } \mu_i^{(L-1)} \in \mathcal{A},$$
$$\mu^{(\ell)} \mapsto \mu_1^{(\ell-1)}, \dots, \mu_s^{(\ell-1)} \text{ for } \mu^{(\ell)} \in \mathcal{A}, \mu_i^{(\ell-1)} \in \mathcal{A}, \quad \ell \in \{L-1, \dots, 1\}.$$

Since the total size of the data increases by a factor s at each level, the input data are made of $d \equiv s^L$ input features $\mu^{(0)}$. We adopt a one-hot encoding of these features, ultimately leading to a data vector $X \in \mathbb{R}^{dv}$. Note that for $\ell \geq 1$, the node variables correspond to latent variables, and there is no need to specify any choice of encoding.

For each level ℓ, we consider that there are m distinct production rules originating from the same higher-level feature $\mu^{(\ell)}$, i.e., there are m equivalent lower-level representations of $\mu^{(\ell)}$. In addition, we assume that two distinct classes or latent variables cannot lead to the same low-level representation. This condition ensures, for example, that two distinct classes never lead to the same data.

We consider the case of the Random Hierarchy Model (RHM) (10), for which the m production rules of any latent variable or class are sampled uniformly at random among the v^s possible ones without replacement. In this case, the total number of possible data produced per class is $m \cdot m^s \cdots m^{s^{L-1}} = m^{\frac{d-1}{s-1}}$, which has exponential dependence in the dimension $d = s^L$. In the following, we use the notation $X_i^{(\ell)}$ to indicate the variable at layer ℓ and position $i \in \{1, \dots, s^{L-\ell}\}$.

In the context of unsupervised learning, a key parameter for this model is $f = m/v^{s-1}$. When $f = 1$, all strings of latent

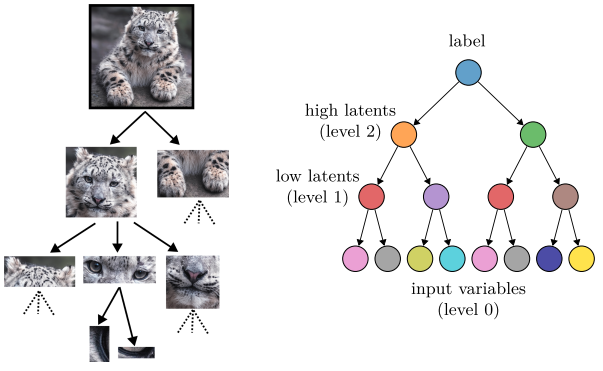

Fig. 3. Sketch of the hierarchical and compositional structure of data. *Left* panel: The leopard in the image can be iteratively decomposed in features at different levels of abstraction. *Right* panel: Generative hierarchical model we study in this paper. In this example, depth $L = 3$ and branching factor $s = 2$. Different values of the input and latent variables are represented with different colors.

variables of size s can be produced at any level of the hierarchy. This implies that all possible v^d input strings are generated, and the data distribution has little structure. When $f < 1$, however, only a small fraction $\sim f^{(d-1)/(s-1)}$ of all possible strings is generated by the production rules. This implies that spatial correlations between different input positions appear, reflecting the hierarchy generating the data.

Optimal Denoising of the RHM with Message Passing

In this section, we characterize the Bayes optimal denoising process for the RHM. Given a noisy observation $X^{(0)} = x(t)$ of the input variables at time t, we compute $p(x(0)|x(t))$ exactly, obtaining full control of the statistics of the backward diffusion process from time t to time 0. In particular, given the tree structure of the model, we can compute the marginal probability of the values of all latent variables conditioned on $x(t)$ by using a message-passing algorithm. Therefore, we obtain the probability that a latent variable at level ℓ has changed when performing the forward–backward diffusion process for a duration t, a central quantity to interpret Fig. 2. The optimal denoising corresponds to reconstructing the data distribution $p(x(0))$ exactly. This perfect reconstruction corresponds to a diffusion model achieving perfect generalization. Although this is a strong assumption for modeling a neural network trained with empirical risk minimization, like the one considered in Section 1, our theoretical analysis captures the phenomenology of our experiments.

Belief Propagation. For computing the marginal distributions, we use Belief Propagation (BP) (51, 52), which gives exact results for a tree graph such as the Random Hierarchy Model. In this case, the leaves of the tree correspond to the input variables at the bottom layer, and the root corresponds to the class variable at the top of the hierarchy. Each rule connecting variables at different levels corresponds to a factor node, as shown in Fig. 4.

The forward process adds noise to the variables in the input nodes. Each of these nodes sends its *belief* on its value at $t = 0$ to its parent latent node. These beliefs, or *messages*, represent probabilistic estimates of the state of the sender node. Each latent node receives messages from all its children, updates its belief about its state, and sends its *upward message* to its parent node. This process is repeated iteratively until the root of the tree. Subsequently, starting from the root, each node sends a *downward message* to its children. Finally, the product of the upward and downward beliefs received at a given node represents the marginal probabilities of its state conditioned on the noisy observation. Hence, we can use these conditional marginals to compute the mean values of the variables at all levels of the hierarchy. We assume that the production rules of the model are known by the inference algorithm, which corresponds to the optimal denoising process.

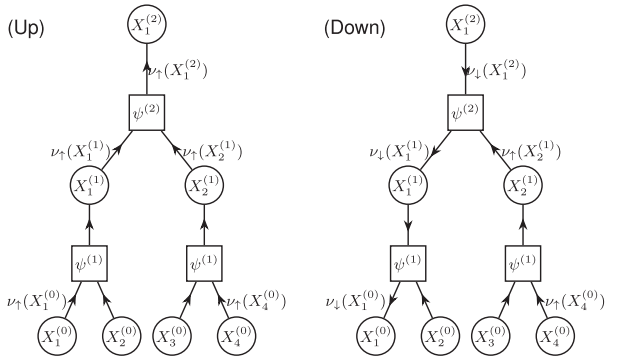

Fig. 4. Illustration of the flow of messages in the Belief Propagation algorithm for the case $s = 2$, $L = 2$ of the Random Hierarchy Model. The factor nodes (squares) represent the rules that connect the variables at different levels of the hierarchy. The downward process is represented only for the leftmost branch.

The input variables $X^{(0)}$, in their one-hot-encoding representation, undergo the forward diffusion process of Eq. **2**, which can be defined in continuous time and constant β_t by redefining $\overline{\alpha_t} = e^{-2t}$ and taking the limit $T \to \infty$ (14).

The denoising is made in two steps: the initialization of the messages at the leaves and the BP iteration.

initialization of the Upward Messages. In its one-hot-encoding representation, $X_i^{(0)}$ is a v-dimensional vector: Taking the symbol $a_\gamma \in \{a_1, \dots, a_v\} = \mathcal{A}$ corresponds to $X_i^{(0)} = e_\gamma$, with e_γ a canonical basis vector. Its continuous diffusion process takes place in $\mathbb{R}^v$: Given the value $X_i^{(0)} = x_i(t)$, we can compute the probability of its starting value $p(x_i(0)|x_i(t))$ using Bayes formula. As derived in *SI Appendix*, section 1, we obtain

$$p(x_i(0) = e_\gamma | x_i(t)) = \frac{1}{Z} e^{x_{i,\gamma}(t)/\Delta_t}, \qquad [4]$$

with $\Delta_t = (1 - \overline{\alpha_t})/\sqrt{\overline{\alpha_t}}$ and $Z = \sum_{\mu=1}^{v} e^{x_{i,\mu}(t)/\Delta_t}$. This computation is performed independently for each input variable i, and therefore does not take into account the spatial correlations given by the generative model. The probabilities of Eq. **4** are used to initialize the BP upward messages $\nu_\uparrow^{(0)} = p(x_i(0)|x_i(t))$ at the input variables.

BP Iteration. Let $\psi^{(\ell)}$ be any factor node connecting an s-tuple of low-level variables at layer $\ell - 1$, $\{X_i^{(\ell-1)}\}_{i\in[s]}$, to a high-level variable $X_1^{(\ell)}$ at layer ℓ. Without loss of generality, to lighten the notation, we rename the variables as $Y = X_1^{(\ell)}$, taking values $y \in \mathcal{A}$, and $X_i = X_i^{(\ell-1)}$, each taking values $x_i \in \mathcal{A}$. For each possible association $y \to x_1, \dots, x_s$, the factor node $\psi^{(\ell)}(y, x_1, ..., x_s)$ takes values

$$\psi^{(\ell)}(y, x_1, ..., x_s) = \begin{cases} 1, & \text{if } y \to (x_1, ..., x_s) \text{ is rule at layer } \ell \\ 0, & \text{otherwise.} \end{cases}$$

The BP upward and downward iterations for the (unnormalized) upward and downward messages respectively read

$$\begin{aligned} \tilde{\nu}_\uparrow^{(\ell+1)}(y) &= \sum_{x_1,\dots,x_s \in \mathcal{A}^{\otimes s}} \psi^{(\ell+1)}(y, x_1, ..., x_s) \prod_{i=1}^{s} \nu_\uparrow^{(\ell)}(x_i), \\ \tilde{\nu}_\downarrow^{(\ell)}(x_1) &= \sum_{\substack{x_2,\dots,x_s \in \mathcal{A}^{\otimes(s-1)} \\ y \in \mathcal{A}}} \psi^{(\ell+1)}(y, x_1, ..., x_s) \\ &\quad \times \nu_\downarrow^{(\ell+1)}(y) \prod_{i=2}^{s} \nu_\uparrow^{(\ell)}(x_i), \end{aligned} \qquad [5]$$

where $\nu_\rho^{(\ell)}(x) = \frac{\tilde{\nu}_\rho^{(\ell)}(x)}{\sum_{x'} \tilde{\nu}_\rho^{(\ell)}(x')}$, $\rho \in \{\uparrow, \downarrow\}$. The downward iteration, reported for x_1, can be trivially extended to the other

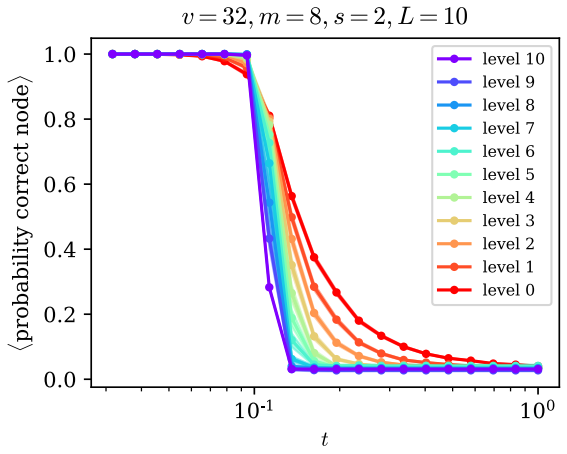

Fig. 5. Probability that the latent has not changed in the denoising process, corresponding to the largest marginal probability computed by BP, averaged for each layer, for varying inversion times of the diffusion process t. Data for the RHM with $v = 32$, $m = 8$, $s = 2$, and $L = 10$. Each level of the tree, indicated in the legend, is represented with a different color. We observe the same behavior of the curves for ImageNet data in Fig. 2: The probability of the correct class has a sharp transition at a characteristic time scale, while the probabilities corresponding to latent variables in the lower levels change smoothly.

variables x_i by permuting the position indices. The values of $\nu_{\uparrow}^{(0)}(x_i)$ and $\nu_{\downarrow}^{(L)}(y)$ are set by the initial conditions. In particular, we initialize $\nu_{\uparrow}^{(0)}(x_i)$ as described in the previous paragraph and $\nu_{\downarrow}^{(L)}(y) = 1/v$, which corresponds to a uniform prior over the possible classes $\mathcal{C}$.[†]

Results. We run the BP upward and backward iterations numerically. In Fig. 5, we show the probability corresponding to the correct symbol for each node of the tree. Remarkably, we note that (i) the probability for the correct class at layer L displays a transition at a characteristic time which becomes sharper for increasing L, and (ii) the messages for the correct input variables and the correct latent variables at low levels of the tree change smoothly. In particular, the curves for messages at layer L and layers $\ell < L$ invert their order at the transition, as in our observations on DDPMs and ImageNet data in Fig. 2. This transition is one of our key findings, which we explain below.

Mean-Field Theory of Denoising Diffusion

In this section, we make a simplifying assumption for the initial noise acting on the input and adopt a mean-field approximation to justify the existence of a phase transition. Remarkably, this approximation turns out to be of excellent quality for describing the diffusion dynamics. Specifically, consider a reference configuration at the leaves variables $X_i^{(0)} = \bar{x}_i$ that we would like to reconstruct, given a noisy observation of it. We assume that for each leaf variable, the noise is uniformly spread among the other symbols.[‡] In other words, our belief in the correct sequence is corrupted by $\epsilon \in [0, 1]$:

$$\begin{cases} X_i^{(0)} = \bar{x}_i & \text{with belief } 1-\epsilon, \\ X_i^{(0)} & \text{uniform over alphabet with belief } \epsilon. \end{cases} \quad [6]$$

[†]This assumption corresponds to unconditioned diffusion, where the DDPM is not biased toward any specific class.

[‡]This is a mild approximation, as documented in *SI Appendix*, section 3.

Hence, the initialization condition of the upward BP messages at a leaf node $X_i^{(0)}$ becomes

$$\begin{cases} \nu_{\uparrow}^{(0)}(\bar{x}_i) & = 1-\epsilon+\epsilon/v, \\ \nu_{\uparrow}^{(0)}(x_i \neq \bar{x}_i) & = \epsilon/v, \end{cases} \quad [7]$$

where v is the alphabet cardinality.

Given these initial conditions and since the production rules are known, if $\epsilon = 0$—i.e., in the noiseless case—BP can reconstruct all the values of the latent variables exactly. Conversely, if $\epsilon = 1$—i.e., when the input is completely corrupted and the belief on the leaves variables is uniform—the reconstruction is impossible. In general, for a value of ϵ, one is interested in computing the probability of recovering the latent structure of the tree at each layer ℓ and, as $L \to \infty$, to decide whether the probability of recovering the correct class of the input remains larger than $1/v$.

Upward Process. We begin by studying the upward process from the leaves. Consider a true input tuple $\bar{x}_1, \ldots, \bar{x}_s$ which is associated with the higher-level feature $\bar{y}$. Given the randomness of the production rules, the messages are random variables depending on the specific realization of the rules. We adopt a *mean-field* or *annealed* approximation that neglects the fluctuations coming from the random choice of rules. Specifically, we approximate the upward message by the average upward message exiting the corresponding factor node $\langle \nu_{\uparrow}^{(1)}(y) \rangle_{\psi}$ over the possible realizations of ψ. In *SI Appendix*, section 2, we show that $\langle \nu_{\uparrow}^{(1)}(y) \rangle_{\psi}$ can take only two values: one for $y = \bar{y}$ and one for $y \neq \bar{y}$, as expected by symmetry considerations. Therefore, mean messages have the same structure as Eq. 7 and we can define a new ϵ'. Introducing the probability of reconstructions $p = 1-\epsilon+\epsilon/v$ and $p' = 1-\epsilon'+\epsilon'/v$, we have

$$p' = \frac{p^s + f\frac{m-1}{mv-1}(1-p^s)}{p^s + f(1-p^s)} = F(p). \quad [8]$$

Iterating this procedure across all the levels of the tree, we can compute the probability of recovering the correct class of the input. In particular, for large L, we are interested in studying the

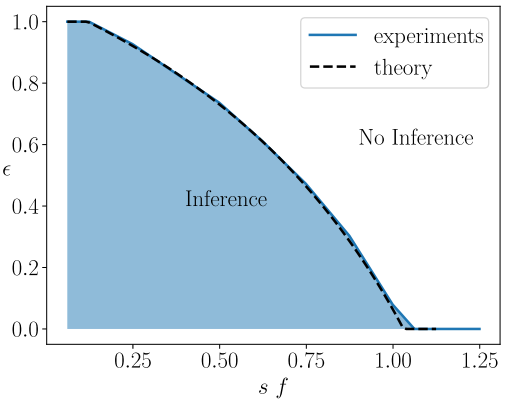

Fig. 6. Phase diagram for inferring the class node using the upward iteration of BP. When $sf < 1$, BP can infer the class if $\epsilon < \epsilon^*(sf)$. This transition is very well predicted by our theory. The inference region in the figure corresponds to the phase wherein the probability of the correct class is larger than the initialization belief in the correct values of the leaves, that is $1-\epsilon+\frac{\epsilon}{v}$. Experimental data are for a single realization of the RHM with $v = 32$, $s = 2$, $L = 10$.

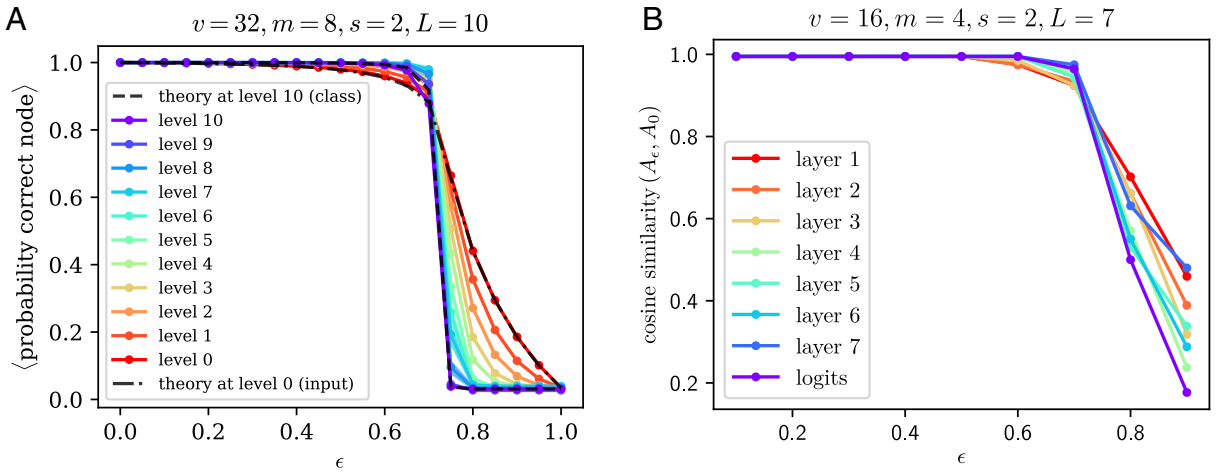

Fig. 7. (*A*) Probability that the latent has not changed in the denoising process, corresponding to the largest marginal probability computed by BP, for varying ϵ. Data for the RHM with $v = 32$, $m = 8$, $s = 2$, and $L = 10$. Each level of the tree, indicated in the legend, is represented with a different color. The black dashed lines are our mean-field theoretical predictions, which show excellent agreement with the experiments. In particular, the inversion between the curves for the *Top* and *Bottom* levels at the phase transition can be observed. (*B*) Cosine similarity between the postactivations following every layer of a deep CNN trained on the RHM ($v = 16$, $m = 4$, $s = 2$, $L = 7$) for the starting and sampled data. Each layer of the architecture, indicated in the legend, is represented with a different color. The curves showcase the same inversion predicted by our theory (cf. panel *A*).

fixed points $p^* = F(p^*)$ of the iteration map in Eq. **8**. As derived in *SI Appendix*, section 2.A, when $sf > 1$, this map has a repulsive fixed point $p^* = 1$, which corresponds to $\epsilon = 0$, and an attractive fixed point $p^* = 1/v$, corresponding to $\epsilon = 1$. Thus, in this regime, inferring the class from the noisy observation of the input is impossible. In contrast, when $sf < 1$, $p^* = 1$ and $p^* = 1/v$ are both attractive fixed points, and a new repulsive fixed point $1/v < p^* < 1$ separating the other two emerges. Therefore, in this second regime, there is a phase transition between a phase in which the class can be recovered and a phase in which it cannot. These theoretical predictions are numerically confirmed in the phase diagram in Fig. 6.

Physically, $sf < 1$ corresponds to a regime in which errors at lower levels of the tree do not propagate: They can be corrected using information coming from neighboring nodes, thanks to the fact that only a small fraction of the strings are consistent with the production rules of the generative model. Conversely, when $sf > 1$, even small corruptions propagate through the entire tree up to the root node and BP cannot infer the class correctly.

Downward Process. The same calculation can be repeated for the downward process, with the additional difficulty that the downward iteration mixes upward and downward messages. We refer the reader to *SI Appendix*, section 2 for the theoretical treatment.

Probabilities of Reconstruction. Combining the mean upward and downward messages, we obtain a theoretical prediction for the probabilities of reconstructing the correct values of the variables at each layer. We compare our theoretical predictions with numerical experiments in Fig. 7*A*. In these experiments, BP equations are solved exactly for a given RHM starting with the initialization of Eq. **7**. Our theory perfectly captures the probability of reconstruction for the input nodes and the class. Moreover, in *SI Appendix*, section 2 we show that our theory predicts the probabilities of reconstruction of latent nodes at all layers.

Experiment on CNN's Activations. Similarly to our experiment on the ConvNeXt in Section 1, we investigate how the hidden representation of a model trained to classify the RHM changes when its input is denoised starting from a corruption noise ϵ. We consider an instantiation of the RHM with $L = 7$, $s = 2$, $v = 16$, and $m = 4$. First, we train a convolutional neural network with $L = 7$ layers, matching the tree structure of the model, with $n = 300$ k training examples up to interpolation. The resulting architecture has 99.2% test accuracy. To sample new data from noisy observations of held-out data, we start by sampling the root using the marginal probability computed with BP. Then, we update the beliefs and the marginals conditioning on the sampled class, and sample one latent variable at layer $L - 1$. We iterate this procedure node-by-node, descending the tree until we obtain a sampled configuration at the bottom layer (52). For each corrupting noise ϵ and each layer of the CNN, we compute the cosine similarity between postactivations for the initial and generated configurations. Panel *B* of Fig. 7 shows the obtained curves. Remarkably, we observe the same qualitative behavior as in panel *A* of Fig. 7, ultimately explaining the empirical observation of Fig. 2.

Conclusions

We have argued that reversing time in denoising diffusion models opens a window on the compositional nature of data. For synthetic hierarchical generative models of data, where the Bayes optimal denoising can be exactly computed, low-level features can already change at small times, but the class remains most often the same. At larger times, a phase transition is found where the probability of remaining in the same class suddenly drops to random chance. Yet, low-level features identical to those of the initial sample can persist and compose the new sample. Strikingly, this theoretical analysis characterizes well the results found with ImageNet, where the denoising is performed by a trained U-Net. Interestingly, the structure of the U-Net with the skip connections between the encoder and decoder parts mimics the upward and downward iterations of belief propagation, where the downward process mixes upward and downward messages. In fact, building on the present work (53) shows that U-Nets are capable of effectively approximating the belief propagation denoising algorithm. Investigating whether the function learned by U-Nets approximates BP is a promising avenue for future work. In the present work, we used the internal representation of deep networks as a proxy for the hierarchical structure of

images. An interesting direction for future work will be using deep hierarchical segmentation techniques (54–57) to extract latent variables, so as to test our predictions on their evolution in forward–backward experiments. Finally, future work can test our theoretical predictions on other modalities successfully handled by diffusion models, such as language and biological structures.

The interplay between the hierarchy in feature space and in time revealed here may help understand the puzzling success of diffusion models, including the number of data needed to train such methods, or why they can generalize and not simply memorize the empirical distribution on which they were trained (58–60). More generally, our results put forward hierarchical generative models as tools to understand open questions for other methods, ranging from the emergence of new skills by the composition of more elementary ones in foundation models to that of transferable representations in self-supervised learning.

Data, Materials, and Software Availability. Software code data have been deposited in the repository Forward–backward diffusion (github.com/pcsl-epfl/forward-backward-diffusion) (61).

ACKNOWLEDGMENTS. We thank Francesco Cagnetta and Umberto Maria Tomasini for helpful discussions. We thank Guillermo Ortiz-Jiménez and Stefano Sarao for feedback on the manuscript. This work was supported by a grant from the Simons Foundation (#454953 Matthieu Wyart).

1. U. Luxburg, O. Bousquet, Distance-based classification with lipschitz functions. *J. Mach. Learn. Res.* **5**, 669–695 (2004).
2. F. Bach, Breaking the curse of dimensionality with convex neural networks. *J. Mach. Learn. Res.* **18**, 629–681 (2017).
3. A. B. Patel, T. Nguyen, R. G. Baraniuk, A probabilistic theory of deep learning. arXiv [Preprint] (2015). https://doi.org/10.48550/arXiv.1504.00641 (Accessed 19 April 2024).
4. E. Mossel, Deep learning and hierarchal generative models. arXiv [Preprint] (2016). https://doi.org/10.48550/arXiv.1612.09057 (Accessed 19 April 2024).
5. T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, Q. Liao, Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. *Int. J. Autom. Comput.* **14**, 503–519 (2017).
6. E. Malach, S. Shalev-Shwartz, A provably correct algorithm for deep learning that actually works. arXiv [Preprint] (2018). https://doi.org/10.48550/arXiv.1803.09522 (Accessed 19 April 2024).
7. J. Schmidt-Hieber, Nonparametric regression using deep neural networks with relu activation function. *Ann. Stat.* **48**, 1875–1897 (2020).
8. A. Favero, F. Cagnetta, M. Wyart, Locality defeats the curse of dimensionality in convolutional teacher-student scenarios. *Adv. Neural. Inf. Process. Syst.* **34**, 9456–9467 (2021).
9. F. Cagnetta, A. Favero, M. Wyart, "What can be learnt with wide convolutional neural networks?" in *International Conference on Machine Learning*, A. Krause *et al.*, Eds. (PMLR, 2023), pp. 3347–3379.
10. F. Cagnetta, L. Petrini, U. M. Tomasini, A. Favero, M. Wyart, How deep neural networks learn compositional data: The random hierarchy model. *Phys. Rev. X* **14**, 031001 (2024).
11. J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics" in *International Conference on Machine Learning*, F. Bach, D. Blei, Eds. (PMLR, 2015), pp. 2256–2265.
12. J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models. *Adv. Neural. Inf. Process. Syst.* **33**, 6840–6851 (2020).
13. Y. Song, S. Ermon, Generative modeling by estimating gradients of the data distribution. *Adv. Neural. Inf. Process. Syst.* **32**, 11918–11930 (2019).
14. Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations" in *International Conference on Learning Representations* (OpenReview.net, 2021).
15. J. Betker *et al.*, Improving image generation with better captions. *Comput. Sci.* **2**, 1–19 (2023), https://cdn.openai.com/papers/dall-e-3.pdf.
16. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, "High-resolution image synthesis with latent diffusion models" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2022), pp. 10684–10695.
17. H. Behjoo, M. Chertkov, U-turn diffusion. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2308.07421 (Accessed 2 November 2024).
18. A. Block, Y. Mroueh, A. Rakhlin, Generative modeling with denoising auto-encoders and langevin sampling. arXiv [Preprint] (2020). https://doi.org/10.48550/arXiv.2002.00107 (Accessed 19 April 2024).
19. K. Oko, S. Akiyama, T. Suzuki, "Diffusion models are minimax optimal distribution estimators" in *International Conference on Machine Learning* (PMLR, 2023), pp. 26517–26582.
20. V. De Bortoli, Convergence of denoising diffusion models under the manifold hypothesis. *Trans. Mach. Learn. Res.* (2022).
21. M. Chen, K. Huang, T. Zhao, M. Wang, "Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data" in *International Conference on Machine Learning* (PMLR, 2023), pp. 4672–4712.
22. H. Yuan, K. Huang, C. Ni, M. Chen, M. Wang, Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. *Adv. Neural Inf. Process. Syst.* **36** (2024).
23. G. Biroli, M. Mézard, Generative diffusion in very large dimensions. *J. Stat. Mech. Theory Exp.* **2023**, 093402 (2023).
24. K. Shah, S. Chen, A. Klivans, Learning mixtures of gaussians using the ddpm objective. *Adv. Neural Inf. Process. Syst.* **36**, 19636–19649 (2023).
25. H. Cui, F. Krzakala, E. Vanden-Eijnden, L. Zdeborova, "Analysis of learning a flow-based generative model from limited sample complexity" in *The Twelfth International Conference on Learning Representations* (OpenReview.net, 2024).
26. S. Mei, Y. Wu, Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models. arXiv [Preprint] (2023). https://arxiv.org/pdf/2309.11420 (Accessed 19 April 2024).
27. Z. Kadkhodaie, F. Guth, S. Mallat, E. P. Simoncelli, "Learning multi-scale local conditional probability models of images" in *The Eleventh International Conference on Learning Representations* (OpenReview.net, 2023).
28. L. Ambrogioni, The statistical thermodynamics of generative diffusion models. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2310.17467 (Accessed 19 April 2024).
29. G. Raya, L. Ambrogioni, Spontaneous symmetry breaking in generative diffusion models. *Adv. Neural. Inf. Process. Syst.* **36**, 66377–66389 (2024).
30. M. Okawa, E. S. Lubana, R. Dick, H. Tanaka, Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *Adv. Neural Inf. Process. Syst.* **36** (2024).
31. G. Rozenberg, A. Salomaa, *Handbook of Formal Languages* (Springer, 1997).
32. S. C. Zhu *et al.*, A stochastic grammar of images. *Found. Trends Comput. Graph. Vis.* **2**, 259–362 (2007).
33. D. Stoyan, *Grenander, ulf: Elements of Pattern Theory* (Johns hopkins university press, baltimore and london, 1997), vol. 1996, pp. xiii+ 222.
34. Y. Jin, S. Geman, "Context and hierarchy in a probabilistic image model" in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)* (IEEE, 2006), vol. 2, pp. 2145–2152.
35. J. M. Siskind, J. Sherman, I. Pollak, M. P. Harper, C. A. Bouman, Spatial random tree grammars for modeling hierarchal structure in images with regions of arbitrary shape. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 1504–1519 (2007).
36. L. J. Li, R. Socher, L. Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework" in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009), pp. 2036–2043.
37. T. Marchand, M. Ozawa, G. Biroli, S. Mallat, Wavelet conditional renormalization group. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv.2207.04941 (Accessed 19 April 2024).
38. S. Shalev-Shwartz, O. Shamir, S. Shammah, Failures of gradient-based deep learning in *International Conference on Machine Learning*, D. Precup, Y. W. Teh, Eds. (PMLR, 2017), pp. 3067–3075.
39. E. Malach, S. Shalev-Shwartz, The implications of local correlation on learning some deep functions. *Adv. Neural. Inf. Process. Syst.* **33**, 1322–1332 (2020).
40. Z. Allen-Zhu, Y. Li, Physics of language models: Part 1, context-free grammar. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2305.13673 (Accessed 19 April 2024).
41. Z. Liu *et al.*, A convnet for the 2020s in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2022), pp. 11976–11986.
42. A. Q. Nichol, P. Dhariwal, "Improved denoising diffusion probabilistic models" in *International Conference on Machine Learning*, M. Meila, T. Zhang, Eds. (PMLR, 2021), pp. 8162–8171.
43. P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis. *Adv. Neural. Inf. Process. Syst.* **34**, 8780–8794 (2021).
44. C. Olah *et al.*, Zoom in: An introduction to circuits. *Distill* **5**, e00024–001 (2020).
45. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436 (2015).
46. M. D. Zeiler, R. Fergus, "Visualizing and understanding convolutional networks" in *Computer Vision – ECCV 2014, Lecture Notes in Computer Science*, D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars, Eds. (Springer International Publishing, 2014), pp. 818–833.
47. Z. Allen-Zhu, Y. Li, "Backward feature correction: How deep learning performs deep (hierarchical) learning" in *The Thirty Sixth Annual Conference on Learning Theory* (PMLR, 2023), pp. 4598–4598.
48. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016), pp. 770–778.
49. U. Grenander, *Elements of Pattern Theory* (JHU Press, 1996).
50. E. DeGiuli, Random language model. *Phys. Rev. Lett.* **122**, 128301 (2019).
51. E. Mossel, Reconstruction on trees: Beating the second eigenvalue. *Ann. Appl. Probab.* **11**, 285–300 (2001).
52. M. Mezard, A. Montanari, *Information, Physics, and Computation* (Oxford University Press, 2009).
53. S. Mei, U-nets as belief propagation: Efficient classification, denoising, and diffusion in generative hierarchical models. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2404.18444 (Accessed 2 May 2024).
54. P. Arbelaez, M. Maire, C. Fowlkes, J. Malik, Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**, 898–916 (2010).
55. S. Ge, S. Mishra, S. Kornblith, C. L. Li, D. Jacobs, "Hyperbolic contrastive learning for visual representations beyond objects" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2023), pp. 6840–6849.
56. J. Xie, X. Zhan, Z. Liu, Y. S. Ong, C. C. Loy, Unsupervised object-level representation learning from scene images. *Adv. Neural. Inf. Process. Syst.* **34**, 28864–28876 (2021).
57. X. Zhang, M. Maire, Self-supervised visual representation learning from hierarchical grouping. *Adv. Neural. Inf. Process. Syst.* **33**, 16579–16590 (2020).
58. G. Somepalli, V. Singla, M. Goldblum, J. Geiping, T. Goldstein, "Diffusion art or digital forgery? investigating data replication in diffusion models" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2023), pp. 6048–6058.
59. N. Carlini *et al.*, "Extracting training data from diffusion models" in *32nd USENIX Security Symposium (USENIX Security 23)*, J. Calandrino, C. Troncoso, Eds. (USENIX Association, 2023), pp. 5253–5270 .
60. T. Yoon, J. Y. Choi, S. Kwon, E. K. Ryu, "Diffusion probabilistic models generalize when they fail to memorize" in *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling* (OpenReview.net, 2023).
61. A. Sclocchi, Forward-backward diffusion. GitHub. https://github.com/pcsl-epfl/forward-backward-diffusion. Deposited 4 March 2024.