Unraveling the Latent Hierarchical Structure of Language and Images via Diffusion Models

Antonio Sclocchi*
Institute of Physics
École polytechnique fédérale de Lausanne
CH-1015 Lausanne
antonio.sclocchi@epfl.ch

Noam Itzhak Levi*
Institute of Physics
Institute of Electrical Engineering
École polytechnique fédérale de Lausanne
CH-1015 Lausanne
noam.levi@epfl.ch

Alessandro Favero*
Institute of Physics
Institute of Electrical Engineering
École polytechnique fédérale de Lausanne
CH-1015 Lausanne
alessandro.favero@epfl.ch

Matthieu Wyart
Institute of Physics
École polytechnique fédérale de Lausanne
CH-1015 Lausanne
matthieu.wyart@epfl.ch

Abstract

High-dimensional data must be highly structured to be learnable. Although the compositional and hierarchical nature of data is often put forward to explain learnability, quantitative measurements establishing these properties are scarce. Likewise, accessing the latent variables underlying such a data structure remains a challenge. Forward-backward experiments in diffusion-based models, where a datum is noised and then denoised, are a promising tool to achieve these goals. We predict in simple hierarchical models that in this process, changes in data occur by correlated chunks, with a length scale that diverges at a noise level where a phase transition is known to take place. Remarkably, we confirm this prediction in both text and image datasets using state-of-the-art diffusion models. Our results suggest that forward-backward experiments are informative on the nature of latent variables, and that the effect of changing deeper ones is revealed near the transition.

1 Introduction

Generative artificial intelligence (AI) systems have demonstrated remarkable capabilities in synthesizing data across various modalities, including images [1, 2] and text [3–5]. The underlying reasons behind these achievements remain poorly understood. Indeed, natural data are often high-dimensional and thus generically intractable due to the curse of dimensionality [6, 7]. Hence, to be learnable, the distribution of the data must be highly structured. Characterizing this structure is a fundamental challenge central to any theory of learning. *Hierarchical compositionality* [8–13] is a candidate property put forward to rationalize the success of deep architectures. In this view, data can be decomposed into features organized hierarchically. It is well-established that the grammatical structure of most languages is approximately context-free and thus hierarchical [14, 15]. To what extent it is still the case once semantics is included to build generative models of text is unclear [16, 17]. Likewise, pattern theory [18] posits that images have a hierarchical structure. In both cases, obtaining quantitative evidence characterizing this hierarchy and building tools to determine the associated latent variables remain a challenge.

*Co-first authors.

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

Generative *denoising diffusion probabilistic models* (DDPMs) offer a new handle to tackle this challenge, particularly through forward-backward experiments, where a controlled level of noise is added to a starting image and then removed to generate a new one [19, 20]. For small noises, low-level features of the image change [19, 20]. Passed a transition point, the class is likely to change [19, 21, 22], but remarkably it is still composed of some of the low-level features of the original image [19], as predicted in simple hierarchical models of data structure. However, the empirical tests in these works were limited to images and did not explore other data modalities. Moreover, the geometrical structure of the changes occurring in such a process is not known. As we shall discuss below, these changes reveal the effect of changing latent variables if the data is hierarchical.

In this work, we derive the length scale associated with changes occurring in the forward-backward protocol, assuming that the data structure is hierarchical and is generated by a probabilistic context-free grammar. We find that these changes are characterized by a length and a volume scale that are maximized at the transition: near that point, changes occur in big chunks, characterizing the effect of changing deep latent variables in the data. We show that these two properties are captured by a certain correlation function and its integral, called 'susceptibility' in the physics literature. Remarkably, these predictions are verified *both in text and image datasets*. This result directly supports that a hierarchical structure is central to both modalities and suggests the forward-backward protocol as a tool to analyze the effect of changing latent variables of different hierarchical levels in text and images.

2 Background

Diffusion models Denoising diffusion models are generative models designed to sample from a distribution by reversing a step-by-step noise addition process [20, 23–25]. Let t indicate the time step in a sequence $[0, \dots, T]$, $q(\cdot)$ the data distribution we wish to sample from and $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ a sample drawn from this distribution. Diffusion models consist of: a *forward process* generating a sequence of increasingly noised data $\{\mathbf{x}_t\}_{1 \le t \le T}$, $q(\mathbf{x}_1, \dots, \mathbf{x}_T|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$, where at the final time T, $\mathbf{x}_T$ corresponds to pure noise; a *backward process*, which reverts the forward one by gradually removing noise. This process is typically obtained by learning the backward transition kernels $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$ using a neural network. Sampling from $q(\cdot)$ is achieved by sampling noise $\mathbf{x}_T \sim q(\mathbf{x}_T)$ and then applying the learned backward process to obtain a new sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$. Different diffusion models correspond to different choices of the forward process, depending on the data space under consideration (see Yang et al. [26] for a review). We consider two different diffusion processes.

Discrete data For discrete data, $\mathbf{x}_0$ consists of a sequence of tokens $x_{0,i}$, $i \in [d]$, each corresponding to a symbol belonging to a vocabulary $\mathcal{V}$. In this case, we consider *diffusion with an absorbing state* by introducing an additional $[\mathsf{MASK}]$ symbol [27]. At time step t, each non-masked token either stays unchanged or transitions to $[\mathsf{MASK}]$ with some probability β_t. Using a one-hot-encoding representation of these $|\mathcal{V}| + 1$ states, the forward transition matrix reads $q(x_{t,i}|x_{t-1,i}) = (1 - \beta_t)\mathbb{I} + \beta_t \mathbf{1}\mathbf{e}_M^T$, where $\mathbb{I}$ is the identity matrix, $\mathbf{1}$ a column vector of all ones and $\mathbf{e}_M$ the one-hot-encoding vector of the $[\mathsf{MASK}]$ symbol. At the final time T, $x_{T,i} = [\mathsf{MASK}]$ for every $i \in [d]$. In the following, we consider the noise schedule $\beta_t = (T - t + 1)^{-1}$ such that $q(x_{t,i}{=}[\mathsf{MASK}]|\mathbf{x}_0) = t/T$ [27].

Continuous data For continuous data, that is $\mathbf{x}_0 \in \mathbb{R}^d$, we consider the time-discretized Gaussian diffusion introduced in [20]. In this case, the forward transition matrix reads $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t \mathbb{I})$, where $\mathcal{N}$ represents the Gaussian probability distribution and the sequence $\{\beta_t\}_{1 \le t \le T}$ is the variance schedule. At the final time T, $\mathbf{x}_T \sim \mathcal{N}(0, \mathbb{I})$.

Forward-backward experiments Forward-backward experiments involve inverting the diffusion process at some intermediate time $t \le T$. Starting from $\mathbf{x}_0$, the forward process up to time t produces a noisy sample $\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_0)$. The backward process produces a new sample $\hat{\mathbf{x}}_0(t) \sim p(\hat{\mathbf{x}}_0|\mathbf{x}_t)$.

The Random Hierarchy Model (RHM) The RHM is a generative model of hierarchically structured data introduced by [13]. It belongs to the class of context-free grammars in the field of language theory [28], and assumes that production rules are random. In its simplest version, it consists of:

- A regular tree graph of depth L and branching factor s.
- A discrete vocabulary $\mathcal{V}^{(\ell)}$ of size v for each level $\ell = 0, 1, \dots, L$ of the tree.
- A set of production rules defining how each symbol belonging to $\mathcal{V}^{(\ell)}$ can be represented at the lower level with the symbols of $(\mathcal{V}^{(\ell-1)})^{\otimes s}$. For each element of $\mathcal{V}^{(\ell)}$, there are m equivalent lower-level representations, which are all distinct and chosen randomly.

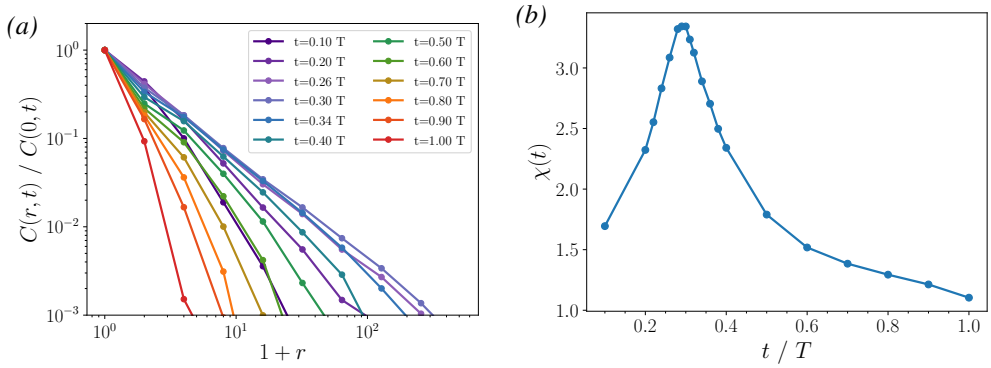

Figure 1: **Correlation measures on diffusion samples of the Random Hierarchy Model (RHM).** *(a)* The average correlation function shows a correlation length that peaks around $t^* \simeq 0.3\,T$, corresponding to the class phase transition. *(b)* Correspondingly, also the average susceptibility shows a peak.

A datum of the RHM is generated by choosing a symbol at the root, which corresponds to the class of the datum, and then using the production rules to generate the lower-level representations (see App. A for examples). As a result, a string of s^L symbols at the leaf nodes is generated. The leaf nodes correspond to the visible tokens, while the upper-level nodes are latent variables. We define the tree distance $\tilde{\ell}$ between two visible tokens as the number of edges between them and their lowest common ancestor. Their corresponding real space distance r is $r = s^{\tilde{\ell}} - 1$. Because of the hierarchical structure generating the data, spatial correlations between the visible tokens are present [29].

Bayes-optimal denoising of the RHM In the RHM, knowing the production rules of its tree structure, the backward diffusion process can be exactly computed using Belief Propagation (BP) [30]. For a noisy observation $\mathbf{x}_t$ of an RHM datum, BP allows sampling directly from the posterior $p(\hat{\mathbf{x}}_0|\mathbf{x}_t)$. This is equivalent to integrating the backward diffusion process starting from $\mathbf{x}_t$. Since the RHM data are sequences of discrete tokens, in what follows we consider diffusion with an absorbing state for this model. Using BP and a mean-field approximation, Sclocchi et al. [19] showed that the forward-backward protocol on the RHM undergoes a sharp phase transition for the class at a specific inversion time t^* (i.e., noise level) in the limit of large depth L.

3 Correlated Blocks of Tokens Change Near the Transition

Let $x_{0,i}$ denote the i-th input token, $i \in [s^L]$, and $\hat{x}_{0,i}(t)$ the same token after undergoing a forward-backward experiment with inversion time t. We seek to compute the correlations between changes in the tokens as a function of the inversion time t. For each token position i, we introduce a quantity characterizing the dynamics, a spin variable $\sigma_i(t)$. It takes the value -1 if after the forward-backward process the symbol of $x_{0,i}$ and $\hat{x}_{0,i}(t)$ remains the same, and $+1$ if the token changed symbol. Using this definition, we can compute the *dynamical correlation function* between the changes of tokens at positions i and j, i.e., $C_{\mathbf{x}_0,ij}(t) = \langle \sigma_i(t)\sigma_j(t)\rangle - \langle \sigma_i(t)\rangle\langle \sigma_j(t)\rangle$, where $\langle\cdot\rangle$ denotes averaging over different stochastic noisy trajectories. By further averaging over the initial point $\mathbf{x}_0$, we define the *average dynamical correlation function* as $C_{ij}(t) = \overline{C_{\mathbf{x}_0,ij}(t)}$. Given the correlations, we compute the *susceptibility* $\chi(t) = \sum_{i=1}^{s^L}\sum_{j=1}^{s^L} C_{ij}(t) / \sum_{i=1}^{s^L} C_{ii}(t)$, where we normalized by the sum of auto-correlations. Intuitively, the susceptibility measures the volume of the blocks of tokens that change together.

In App. A, using a mean-field approximation, we estimate the correlation length ξ, which measures the typical distance over which the correlations between the spin variables σ_i's extend. We predict that ξ diverges at the class transition, indicating that large blocks of tokens change in concert. These large correlated changes are caused by the modifications of deeper and deeper latent variables near the transition. At both smaller and larger time or noise levels, the correlation length decays. This behavior of the dynamical correlation functions implies that the susceptibility also peaks at the transition.

To test our theoretical predictions, in Fig. 1 (a), we present the correlation function $C(r,t)$, corresponding to $C_{ij}(t)$ averaged on all pairs ij such that their real space distance is r, and normalized by the auto-correlation $C(0,t)$. As the inversion time t increases, we observe a growing correlation length, which peaks at a critical time $t^* \approx 0.3\,T$. Our result demonstrates that the dynamical correla-

(a)

	Masking fraction $t/T = 0.1$	Masking fraction $t/T = 0.5$
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	The next morning , March 3, the situation changed. The border was secure, ammunition, food and everybody were nearly completely met.
	Masking fraction $t/T = 0.7$	Masking fraction $t/T = 0.9$
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	increased. On September 3, the situation was under control. Despite tons of ammunition, air train orders were almost completely violated.	information on maps of the actual burial population size. The number is probably around 30,000, we were almost completely encroached

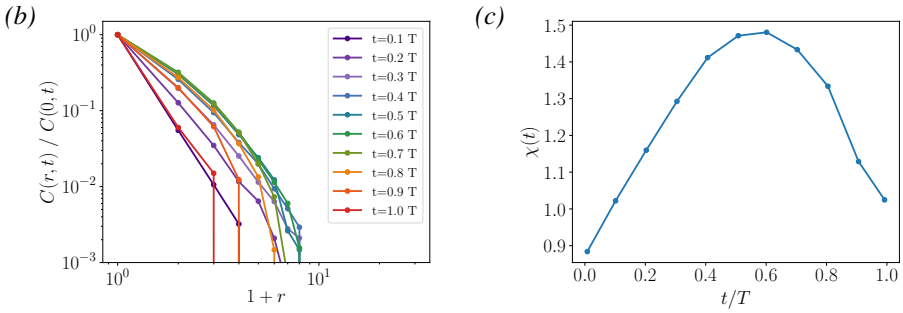

Figure 2: **Forward-backward experiments with language diffusion models.** *(a)* Forward-backward examples for different masking fractions. The words in blue (green) are those that were masked and changed (did not change), while the words in red changed following the backward process. *(b)* Normalized correlations as a function of index distance $r = |i - j|$ for different fractions of masked tokens. *(c)* Susceptibility $\chi(t)$ as a function of masking fraction. The results are averaged over $N_S = 300$ textual samples, each consisting of $N_T = 128$ tokens, with $N_R = 50$ noise realizations for each masking fraction. The susceptibility is given by integrating the normalized correlations over the domain $r \in [0, 10]$.

tions are maximum at that point. Before and beyond t^*, the correlation length decreases, consistently with our prediction. In Fig. 1 (b), we plot the susceptibility, which peaks at the phase transition.

4 Experiments on Natural Language and Image Data

Language diffusion models We consider Masked Diffusion Language Models (MDLM) [31] utilizing the `GPT2` tokenizer. We perform forward-backward experiments starting from samples from the `WikiText2` dataset. In Fig. 2 (a), we illustrate how an initial sentence changes with different inversion times t. At small t, only a few isolated words are modified. At intermediate t, we clearly observe clusters of words changing in a correlated manner. At large t, only a small fraction of the initial sentence remains unchanged (see App. B for a presentation of the same data in their larger context). In Fig. 2 (b-c), we quantify these observations by measuring the average correlation functions and susceptibility[2]. Strikingly, in line with the phenomenology obtained for the RHM, we find a growing correlation length as t increases, reaching a maximum of $7 \div 8$ tokens at a critical inversion time $t^* \approx 0.6\,T$, followed by a subsequent decline. Additionally, the susceptibility peaks at t^*, establishing the existence of a phase transition for the language modality.

[2]To avoid finite size effects due to imposing a fixed masking fraction, we integrate the average correlation function up to the maximal correlation length $r \sim \mathcal{O}(10)$.

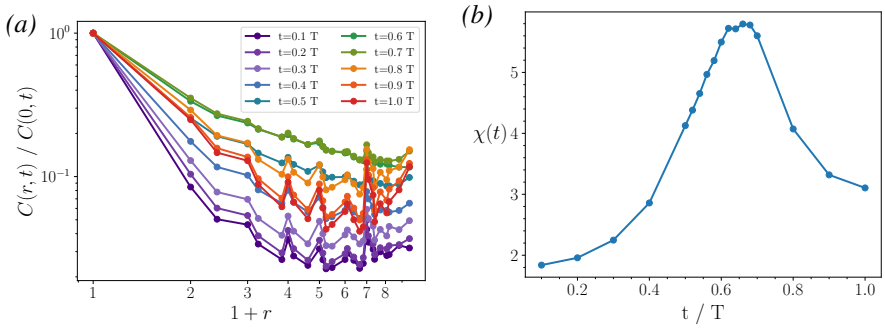

Figure 3: **Correlation measures on the variation of CLIP embeddings of images generated with forward-backward diffusion.** *(a)* The average correlation function displays a system spanning power-law behavior for $t^* \approx 0.6 \div 0.7\,T$, corresponding to the class phase transition [19]. *(b)* In correspondence with the phase transition, the average susceptibility displays a peak. Data obtained with 344 starting images and 128 noisy trajectories per image.

Vision diffusion models We extend our analysis to computer vision by considering Improved Denoising Diffusion Probabilistic Models [32], trained on the `ImageNet` dataset. To compute the correlation between changes in the image tokens, we follow recent trends in multimodal LLMs [33, 34]. Specifically, we divide each image into 7×7 patches and use the last-layer embeddings for each patch from a CLIP ViT-B32 [35] to tokenize the image. Let $\mathbf{x}_i$ denote the embedding of the i-th patch, where $i = (k, l)$ with $k, l \in [7]$. After the forward-backward process, the variation of each patch embedding is given by $\Delta\mathbf{x}_i(t) = \mathbf{x}_{0,i} - \hat{\mathbf{x}}_{0,i}(t)$. We then compute the average correlations between the norms of these variations: $C_{ij}(t) = \overline{\langle \|\Delta\mathbf{x}_i(t)\|\, \|\Delta\mathbf{x}_j(t)\| \rangle - \langle \|\Delta\mathbf{x}_i(t)\| \rangle \langle \|\Delta\mathbf{x}_j(t)\| \rangle}$. The susceptibility is subsequently obtained as $\chi(t) = \sum_{ij} C_{ij}(t) / \sum_i C_{ii}(t)$. In Fig. 3, we present the average correlation functions and the susceptibility for vision DDPMs, starting from `ImageNet` samples. At a critical inversion time $t^* \approx 0.6 \div 0.7\,T$, we observe a peak in susceptibility, signaling the class phase transition in these models. This finding highlights the compositional semantic structure of image data, similar to the phase transitions observed in language diffusion models and the RHM.

5 Conclusion

We used context-free hierarchical models to predict a growing length scale near a phase transition in diffusion models. This prediction was confirmed through experiments across different natural data modalities and neural architectures. Our results reveal a remarkable level of universality. This supports the hypothesis that hierarchical and compositional structures are fundamental, universal properties underlying natural data as diverse as image and text. Future work can include interpreting the large chunks of textual change in terms of grammatical structure and context variables, possibly sharpening these concepts by the data-driven method presently introduced.

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3), 2023.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[3] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[4] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[5] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[6] Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with lipschitz functions. *The Journal of Machine Learning Research*, 5(Jun):669–695, 2004.

[7] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[8] Ankit B Patel, Tan Nguyen, and Richard G Baraniuk. A probabilistic theory of deep learning. *arXiv preprint arXiv:1504.00641*, 2015.

[9] Elchanan Mossel. Deep learning and hierarchal generative models. *arXiv preprint arXiv:1612.09057*, 2016.

[10] Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

[11] Eran Malach and Shai Shalev-Shwartz. A provably correct algorithm for deep learning that actually works. *arXiv preprint arXiv:1803.09522*, 2018.

[12] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

[13] Francesco Cagnetta, Leonardo Petrini, Umberto M Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: The random hierarchy model. *arXiv preprint arXiv:2307.02129*, 2023.

[14] Noam Chomsky. *Aspects of the Theory of Syntax*. Number 11. MIT press, 2014.

[15] Gerhard Jäger and James Rogers. Formal language theory: refining the chomsky hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1598):1956–1970, 2012.

[16] Adele E Goldberg. Constructions: A construction grammar approach to argument structure. *Chicago UP*, 1995.

[17] Adele E Goldberg. Compositionality. In *The Routledge handbook of semantics*, pages 419–433. Routledge, 2015.

[18] Ulf Grenander. *Elements of pattern theory*. JHU Press, 1996.

[19] Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *arXiv preprint arXiv:2402.16991*, 2024.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[21] Luca Ambrogioni. The statistical thermodynamics of generative diffusion models. *arXiv preprint arXiv:2310.17467*, 2023.

[22] Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models, 2024.

[23] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[24] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[25] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[26] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[27] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

[28] Grzegorz Rozenberg and Arto Salomaa. *Handbook of Formal Languages*. Springer, January 1997. doi: 10.1007/978-3-642-59126-6.

[29] Francesco Cagnetta and Matthieu Wyart. Towards a theory of how the structure of language is acquired by deep neural networks. *arXiv preprint arXiv:2406.00048*, 2024.

[30] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[31] Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *arXiv preprint arXiv:2406.07524*, 2024.

[32] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[34] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

A Random Hierarchy Model

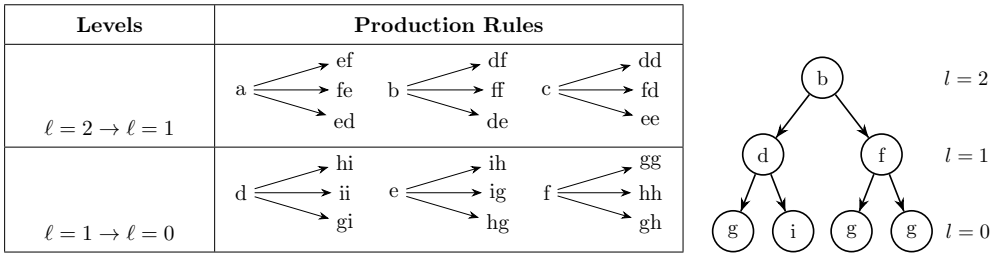

Figure 4: **Example of RHM with** $L = 2$**,** $s = 2$**,** $v = 3$**,** $m = 3$. *Left*: example of production rules with vocabularies $\mathcal{V}^{(2)} = \{a, b, c\}$, $\mathcal{V}^{(1)} = \{d, e, f\}$, $\mathcal{V}^{(0)} = \{g, h, i\}$. *Right*: one possible datum generated by the production rules, with the hierarchical levels indicated on the right.

A.1 Denoising the RHM with Belief Propagation

A.1.1 Prior at the leaves

Let's consider a datum $\mathbf{x}_0$ of the RHM undergoing masking diffusion. At any time t, the tokens of $\mathbf{x}_t$ can have value

$$\begin{aligned} &x_{t,i} = x_{0,i}, && \text{if token } i \text{ has not yet been masked;} \\ &x_{t,i} = [\mathsf{MASK}], && \text{if token } i \text{ has already been masked.} \end{aligned} \tag{1}$$

Therefore, given the noisy observation $\mathbf{x}_t$, the prior belief $\nu_\uparrow(x_{0,i} = \tilde{a})$ on the value of the token i being equal to $\tilde{a}$ is:

$$\begin{aligned} &\nu_\uparrow(x_{0,i} = \tilde{a}) = \delta_{a,\tilde{a}} && \text{if } x_{t,i} = a \in \mathcal{V}^{(0)}; \\ &\nu_\uparrow(x_{0,i} = \tilde{a}) = 1/v && \forall \tilde{a} \in \mathcal{V}^{(0)} \text{ if } x_{t,i} = [\mathsf{MASK}]. \end{aligned} \tag{2}$$

A.1.2 BP iteration

The prior beliefs on the values of the leaves correspond to the initialization of the upward messages at the leaves level $\ell = 0$ for the BP algorithm: $\nu_\uparrow^{(0)}(x_i)$. Given the messages of an s-patch, e.g., $\{\nu_\uparrow^{(\ell)}(x_i)\}_{i=1,\dots,s}$, having a common parent node y in the tree, the upward message in the upper level is computed as:

$$\tilde{\nu}_\uparrow^{(\ell+1)}(y) = \sum_{x_1,\dots,x_s \in \mathcal{V}^{(\ell)\otimes s}} \psi^{(\ell+1)}(y, x_1, \dots, x_s) \prod_{i=1}^{s} \nu_\uparrow^{(\ell)}(x_i), \tag{3}$$

$$\nu_\uparrow^{(\ell+1)}(y) = \frac{\tilde{\nu}_\uparrow^{(\ell+1)}(y)}{\sum_{y' \in \mathcal{V}^{(\ell+1)}} \tilde{\nu}_\uparrow^{(\ell+1)}(y')}, \tag{4}$$

where the factor $\psi^{(\ell+1)}(y, x_1, \dots, x_s)$ reads

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

After iterating this upward process until the root node, BP computes the downward messages (we consider the prior belief on the value of the root node to be uniform over the symbols of $\mathcal{V}^{(L)}$):

$$\tilde{\nu}_\downarrow^{(\ell)}(x_1) = \sum_{\substack{x_2,\dots,x_s \in \mathcal{V}^{(\ell)\otimes(s-1)} \\ y \in \mathcal{V}^{(\ell+1)}}} \psi^{(\ell+1)}(y, x_1, ..., x_s)\, \nu_\downarrow^{(\ell+1)}(y) \prod_{i=2}^{s} \nu_\uparrow^{\ell}(x_i) \tag{5}$$

$$\nu_{\downarrow}^{(\ell)}(x) = \frac{\tilde{\nu}_{\downarrow}^{(\ell)}(x)}{\sum_{x' \in \mathcal{V}^{(\ell)}} \tilde{\nu}_{\downarrow}^{(\ell)}(x')}. \tag{6}$$

At the end of the upward-downward iteration, BP gives the marginal probabilities for the starting value of each node $X_i^{(\ell)}$ of the tree, given the noisy observation $\mathbf{x}_t$:

$$p(X_i^{(\ell)} = a | \mathbf{x}_t) \propto \nu_{\uparrow}(X_i^{(\ell)} = a)\, \nu_{\downarrow}(X_i^{(\ell)} = a), \quad a \in \mathcal{V}^{(\ell)}. \tag{7}$$

Similarly, sampling from the posterior probabilities given by BP is done by sampling from the root and updating the marginal probabilities every time a new node is sampled [30].

A.1.3 Mean-field analysis of BP

The algorithm can be analyzed theoretically with a *mean-field*-like simplification. Instead of considering the prior beliefs explained in A.1.1, we introduce a noise-to-signal ratio ϵ and modify the prior as follows

$$\begin{aligned} &\nu_{\uparrow}(x_{0,i} = a) = 1 - \epsilon + \epsilon/v && \text{if } x_{0,i} = a; \\ &\nu_{\uparrow}(x_{0,i} = \tilde{a}) = \epsilon/v && \text{if } x_{0,i} \neq \tilde{a}. \end{aligned} \tag{8}$$

The role of ϵ is decreasing the prior belief on the correct value of a node. Using this approximation, Sclocchi et al. [19] computed the average upward message, where the average is performed over the possible choices of the RHM rules. The result is the average probability of reconstructing the starting value of a latent node at a given layer ℓ, p_ℓ:

$$p_{\ell+1} = F(p_\ell), \tag{9}$$

where $p_0 = 1 - \epsilon + \epsilon/v$ and $F(p) = \frac{p^s + \frac{m-1}{v^s-1}(1-p^s)}{p^s + \frac{mv-1}{v^s-1}(1-p^s)}$. The fixed point of this iteration map turns out to accurately describe the behavior of BP as a denoiser for the RHM [19]. In particular, in the regime where a phase transition of the reconstruction of the root node exists, there are three fixed points: two attractive ones, corresponding to $p = 1/v$ and $p = 1$, and a repulsive one, i.e., the non-trivial solution of $p^* = F(p^*)$.

A.2 Dynamical correlation length

Correlated changes in input tokens at a given distance happen when a common ancestor latent variable is modified. Thus, we can estimate the typical distance over which token changes are correlated by computing the number of layers $\tilde{\ell}$ after which the probability of reconstructing the latent variables $p_{\tilde{\ell}}$ converges to the two trivial fixed points $p = 1$ and $p = 1/v$. This corresponds to the number of layers required to escape the repulsive fixed point $p = p^* \in (\frac{1}{v}, 1)$.

Let us recall the iteration map for the probability $p_{\ell+1}$ of reconstructing the latent variables at level $\ell + 1$,

$$p_{\ell+1} = \frac{p_\ell^s + \frac{m-1}{v^s-1}(1 - p_\ell^s)}{p_\ell^s + \frac{mv-1}{v^s-1}(1 - p_\ell^s)} = F(p_\ell). \tag{10}$$

We can linearize it around the fixed point p^* and iterate for ℓ layers,

$$\Delta p_{\ell+1} = \left(\left. \frac{dF(p)}{dp} \right|_{p^*} \right)^{\ell} \Delta p_0, \tag{11}$$

where $\Delta p_\ell = p_\ell - p^*$. Since p^* is repulsive, we have that $\left.\frac{dF(p)}{dp}\right|_{p^*} > 1$ and we use the shorthand notation $F'_* = \left.\frac{dF(p)}{dp}\right|_{p^*}$. We want to compute the depth $\tilde{\ell}$ at which:

$$F_*'^{\tilde{\ell}} \, |\Delta p_0| = \mathcal{O}(1), \tag{12}$$

since p^* and $1-p^*$ are quantities of order $\mathcal{O}(1)$ with respect to $|\Delta p_0| \to 0$. In terms of the corruption noise $\epsilon = \frac{1-p}{1+1/v}$, we have

$$F_*'^{\tilde{\ell}} \, |\Delta \epsilon_0| = \mathcal{O}(1). \tag{13}$$

Hence,

$$\tilde{\ell} \sim -\frac{\log |\Delta \epsilon_0|}{\log F_*'} = -\frac{\log |\epsilon_0 - \epsilon^*|}{\log F_*'}. \tag{14}$$

From the depth $\tilde{\ell}$, we can compute the correlation length in input space as

$$\xi = s^{\tilde{\ell}} \sim |\epsilon_0 - \epsilon^*|^{-\frac{\log s}{\log F_*'}}, \tag{15}$$

that diverges at the critical point:

$$\lim_{\epsilon \to \epsilon^*} \xi = +\infty. \tag{16}$$

B Language Diffusion

B.1 Setup

Here, we briefly describe the particular realization of discrete diffusion used in the MDLM setting, which is detailed in [31].

MDLMs are a form of discrete diffusion model tailored for language generation. Unlike autoregressive (AR) models, MDLMs generate text by gradually unmasking tokens, allowing for non-sequential generation. This process is governed by a forward masking and reverse unmasking process, parameterized using a Rao-Blackwellized objective to improve performance.

Forward Process: The forward process is defined by progressively noising a clean input sequence x using a categorical distribution:

$$q(z_t|x) = \text{Cat}(z_t; \alpha_t x + (1-\alpha_t)m), \tag{17}$$

where z_t is the latent variable at time t, representing the noisy version of the input sequence, x is the original, clean sequence of tokens, $\text{Cat}(\cdot;\cdot)$ is a categorical distribution over the possible states, α_t is the noise schedule function, strictly decreasing from 1 to 0 as t increases, and m is a one-hot vector representing the special masked token. At each time step, a fraction of the data transitions into the masked state.

Reverse Process and Rao-Blackwellization: The reverse diffusion process reconstructs the original data from noisy observations. It is parameterized using a neural network approximation $x_\theta(z_t, t)$, which predicts clean tokens from noisy inputs:

$$p_\theta(z_s|z_t) = \begin{cases} \text{Cat}(z_s; z_t), & \text{if } z_t \neq m, \\ \text{Cat}\left(z_s; \frac{(1-\alpha_s)m + (\alpha_s - \alpha_t)x_\theta(z_t,t)}{1-\alpha_t}\right), & \text{if } z_t = m. \end{cases} \tag{18}$$

where z_s is the latent variable at a prior time step s (with $s < t$), $x_\theta(z_t, t)$ is a neural network approximation of x given the noisy observation z_t at time t, and $p_\theta(\cdot|\cdot)$ is the model distribution approximating the true reverse process.

The training objective is a *negative evidence lower bound* (NELBO), expressed as:

$$L_{\text{diffusion}} = \sum_{i=1}^{T} \mathbb{E}_q \left[\frac{\alpha_{t(i)} - \alpha_{s(i)}}{1 - \alpha_{t(i)}} \log \langle x_\theta(z_{t(i)}), x \rangle \right], \tag{19}$$

where T is the number of diffusion steps, $\alpha_{t(i)}$, $\alpha_{s(i)}$ is the noise schedules evaluated at time steps $t(i)$ and $s(i)$, respectively, $\mathbb{E}_q$ is the expectation over the forward process defined by q, and $\langle x_\theta(z_{t(i)}), x\rangle$ is the dot product between the neural network output $x_\theta(z_{t(i)})$ and the original input x.

Continuous-Time Likelihood Bounds: To achieve a tighter approximation to the ELBO, the discrete objective is extended to continuous time as:

$$L_{\infty\text{NELBO}} = \mathbb{E}_q \int_0^1 \frac{\alpha_t'}{1-\alpha_t} \log\langle x_\theta(z_t, t), x\rangle \, dt. \tag{20}$$

where α_t' is the time derivative of the noise schedule α_t. The integral evaluates the objective over continuous time, providing a tighter bound on the likelihood. This formulation is invariant to the specific functional form of the noise schedule α_t, highlighting the robustness of the MDLM approach.

Connection to Masked Language Models: MDLMs leverage a masked diffusion approach where the training objective is a weighted average of classical masked language modeling (MLM) losses:

$$L_{\infty\text{NELBO}} = \mathbb{E}_q \int_0^1 \frac{\alpha_t'}{1-\alpha_t} \sum_\ell \log\langle x_\theta^\ell(z_t), x^\ell\rangle \, dt, \tag{21}$$

where x^ℓ: The ℓ-th token in the original sequence, $x_\theta^\ell(z_t)$: The neural network's prediction for the ℓ-th token given the noisy sequence z_t. The summation runs over all tokens in the sequence, effectively establishing a connection between MDLMs and BERT-style encoders while equipping them with generative capabilities.

We employ the MDLM proposed in [31] to conduct the forward-backward experiments described in Sec. 4, by first drawing random texts of a fixed token length from the `WikiText2` database, masking a fixed fraction of the tokens t, and then performing the backward diffusion process by using the masked sequence as the initial point for the MDLM model.

B.2 Examples of Text Samples for the Forward-Backward Experiments

Below, we provide examples of texts generated by the forward-backward process using MDLM seeded from `WikiText2` examples for different masking fractions. Selected samples were shown in the main text in Fig. 2 (a). We dub the text results after the forward-backward process as *U-turn* samples. As can be seen by the color coding, correlated blocks of words change together along the denoising process, as described in Sec. 3, and the semantic meaning of the paragraphs themselves change along the phase transition. In blue we denote masked tokens that have changed their value after the backward process, while in green masked tokens that have returned to their initial value. Red indicates the changes in the final texts. It can be seen that for small masking fractions such as $t = 0.1T$, most of the tokens do not change after masking, while the amount of changed tokens far exceeds the unchanged ones near the phase transition at $t = 0.5T$, hinting at the long-range correlations emerging.

Masking fraction = 0.9

Highlighted Original Text:

The third day, September 3, the situation worsened. The weather was hot and ammunition, food and supplies were nearly completely exhausted . Since the previous afternoon, North Korean mortar barrages had alternated with infantry assaults against the perimeter. Survivors later estimated there were about twenty separate infantry attacks repulsed. Two North Korean machine guns still swept the perimeter whenever anyone showed himself . Dead and dying US troops were in almost every fox hole. Mortar fragments destroyed the radio and this ended all communication with other US units. Artillery fire and air strikes requested by Schmitt never came. Some North Koreans worked their way close to the perimeter and threw grenades

Highlighted U-Turn Text:

information on maps of the actual burial population size. The number is probably around 30,000, we were almost completely encroached into the population as there were to 100 barr is we excavated the site on against the walls, it is estimated there were at around 30,000 and another holding room for perhaps 10,000 . It also seems highly unlikely, as with Dead Drop sites generally, that the only evidence for the storage of the firearm from the drop was more wood pieces. The other medieval site which required constant fire and perhaps continual storage is the firearm, one of which we were aware of having been stored during the same time period

Masking fraction = 0.7

Highlighted Original Text:

The third day, September 3, the situation worsened. The weather was hot and ammunition, food and supplies were nearly completely exhausted . Since the previous afternoon, North Korean mortar barrages had alternated with infantry assaults against the perimeter. Survivors later estimated there were about twenty separate infantry attacks repulsed. Two North Korean machine guns still swept the perimeter whenever anyone showed himself. Dead and dying US troops were in almost every fox hole. Mortar fragments destroyed the radio and this ended all communication with other US units. Artillery fire and air strikes requested by Schmitt never came. Some North Koreans worked their way close to the perimeter and threw grenades

Highlighted U-Turn Text:

increased. On September 3, the situation was under control. Despite tons of ammunition , air train orders were almost completely violated. On the previous day, North Americans, farm crews and miners were heard rebelling against the perimeter. Survivors were estimated to be about twenty dead from attacks convulsing and starvation, as machine guns still swept the perimeter whenever ever they could. Dead - end US troops were in almost every fox hole for about twenty minutes; the radio and newspapers were all frequently with news of general effects, crying out for particular strikes or on the loading of vehicles. Some North Americans reported blocking way to fill the perimeter, and others

Masking fraction = 0.5

Highlighted Original Text:

The third day, September 3, the situation worsened. The weather was hot and ammunition, food and supplies were nearly completely exhausted. Since the previous afternoon, North Korean mortar barrages had alternated with infantry assaults against the perimeter. Survivors later estimated there were about twenty separate infantry attacks repulsed. Two North Korean machine guns still swept the perimeter whenever anyone showed himself. Dead and dying US troops were in almost every fox hole. Mortar fragments destroyed the radio and this ended all communication with other US units. Artillery fire and air strikes requested by Schmitt never came. Some North Koreans worked their way close to the perimeter and threw grenades

Highlighted U-Turn Text:

The next morning, March 3, the situation changed. The border was secure, ammunition, food and everybody were nearly completely met. On the previous afternoon, North Korean artillery barrister repulseated an infantry attack within the perimeter. Survivors later said there were about twenty separate infantry attacks repulseated. Two North Korean machine guns shells had the ground where anyone showed himself. Dead and wounded US troops were in wounded positions. At the time, fragments of mortar shells eliminated any communication of communication with other US troops. Exceptional fire and submunitions by Schmitt never came. The North Koreans worked their way up to the ground and threw bottles

Masking fraction = 0.3

Highlighted Original Text:

```
The third day, September 3, the situation worsened. The weather was hot
and ammunition, food and supplies were nearly completely exhausted. Since
the previous afternoon, North Korean mortar barrages had alternated with
infantry assaults against the perimeter. Survivors later estimated there
were about twenty separate infantry attacks repulsed. Two North Korean
machine guns still swept the perimeter whenever anyone showed himself.
Dead and dying US troops were in almost every fox hole. Mortar fragments
destroyed the radio and this ended all communication with other US units.
Artillery fire and air strikes requested by Schmitt never came. Some North
Koreans worked their way close to the perimeter and threw grenades
```

Highlighted U-Turn Text:

```
 third! On the 3rd the situation worsened. The perimeter was thick and
ammunition, food and fuel were nearly completely exhausted. By the late
afternoon, North Korean mortar barrages still cooperated with infantry
assaults against the perimeter for, later hours there were about 10
separate infantry attacks repulsed. Two North Korean machine guns still
swept the perimeter without anyone but himself. Dead and dying US troops
were in practically every man hole. Mortar fragments destroyed all radio
and this ended all communication with other US units. Artillery fire or
air support requested by Schmitt still came. Some North Koreans worked to
bring them to the perimeter. The whites
```

Masking fraction = 0.1

Highlighted Original Text:

```
The third day, September 3, the situation worsened. The weather was hot
and ammunition, food and supplies were nearly completely exhausted. Since
the previous afternoon, North Korean mortar barrages had alternated with
infantry assaults against the perimeter. Survivors later estimated there
were about twenty separate infantry attacks repulsed. Two North Korean
machine guns still swept the perimeter whenever anyone showed himself.
Dead and dying US troops were in almost every fox hole. Mortar fragments
destroyed the radio and this ended all communication with other US units.
Artillery fire and air strikes requested by Schmitt never came. Some North
Koreans worked their way close to the perimeter and threw grenades
```

Highlighted U-Turn Text:

```
The third day, September 3, the situation worsened. The weather was hot
and ammunition, tanks and supplies were nearly completely exhausted. Since
the early afternoon, North Korean artillery barrages had alternated with
infantry assaults against the perimeter. Survivors later estimated there
were about twenty separate infantry attacks repulsed. Two North Korean
machine guns still swept the perimeter whenever anyone showed himself.
Dead and dying US troops were in almost every fox hole. Mortar fragments
destroyed the radio and this ended all communication with other US units.
Artillery fire and air strikes requested by Schmitt never stopped. Some
North Koreans worked their way close to the perimeter and threw grenades
```

C Image Diffusion

For image diffusion, we use the publicly available models from *Improved Denoising Diffusion Probabilistic Models* [32], trained on the `ImageNet` dataset at resolution 256×256. We use the class-unconditional model to ensure a class phase transition at an intermediate diffusion time. To tokenize the images in a semantically meaningful manner, we use the last-layer embeddings from a CLIP ViT-B32 [35] encoder. This procedure crops the images to the size 224×224, which get tokenized in 7×7 patches, each of dimension 32×32. The embeddings at the last layer of the CLIP encoder have dimension 768.

In Figure 5, we provide some examples of images generated with the forward-backward protocol. In red, we highlight the patches whose CLIP embeddings show a statistically significant change with respect to the starting image ($t = 0$).

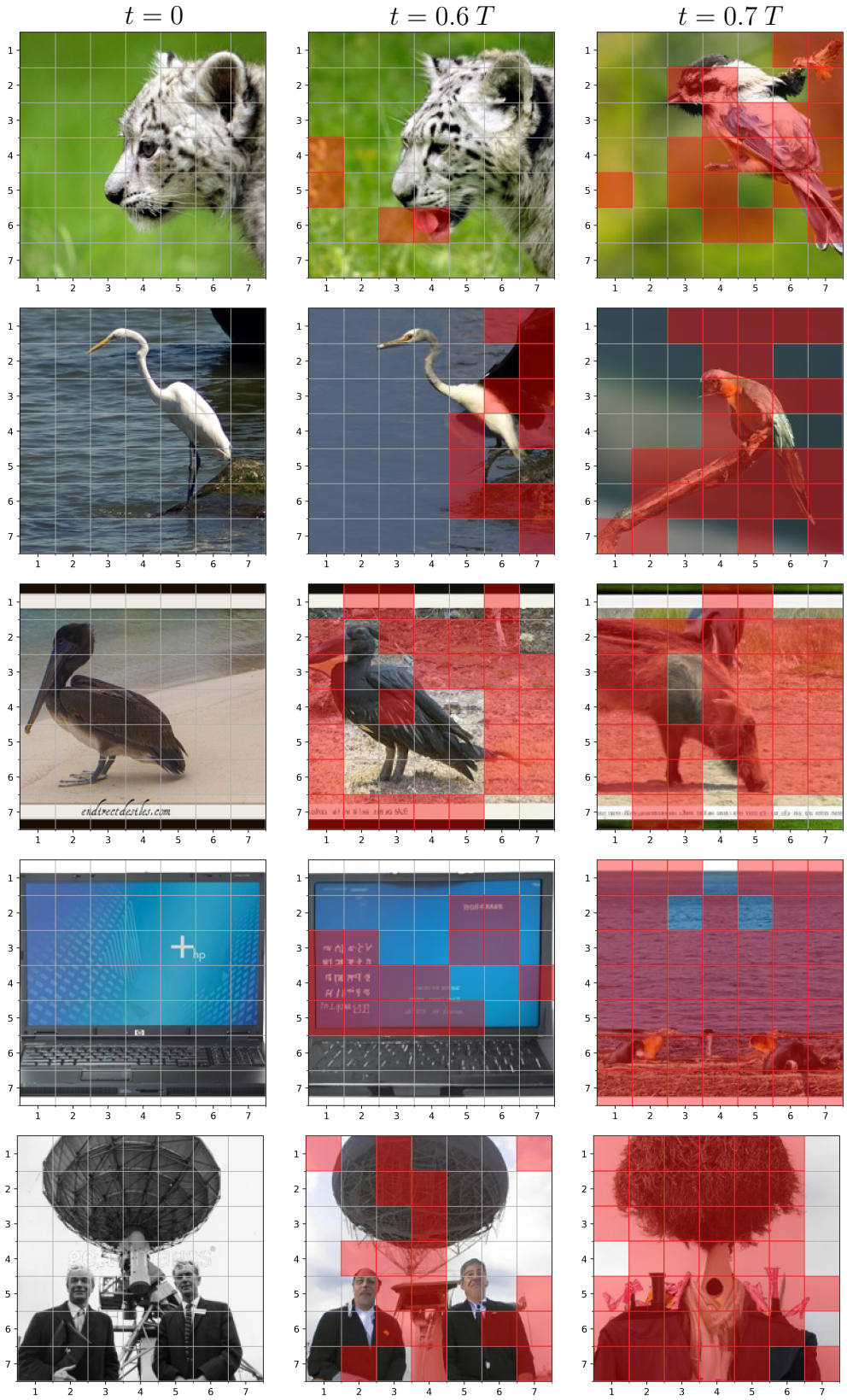

Figure 5: **Examples of images generated at different inversion times t.** The grid indicates the patches represented inside the CLIP vision encoder. For inversion time $t > 0$, the red patches indicate the token embeddings that have a variation magnitude larger than a fixed threshold. These patches of variation appear in domains of growing size.