Structured Generations: Using Hierarchical Clusters to guide Diffusion Models

Jorge da Silva Gonçalves [1] **Laura Manduchi** [1] **Moritz Vandenhirtz** [1] **Julia E. Vogt** [1]

Abstract

This paper introduces Diffuse-TreeVAE, a deep generative model that integrates hierarchical clustering into the framework of Denoising Diffusion Probabilistic Models (DDPMs). The proposed approach generates new images by sampling from a root embedding of a learned latent tree VAE-based structure, it then propagates through hierarchical paths, and utilizes a second-stage DDPM to refine and generate distinct, high-quality images for each data cluster. The result is a model that not only improves image clarity but also ensures that the generated samples are representative of their respective clusters, addressing the limitations of previous VAE-based methods and advancing the state of clustering-based generative modeling.

1. Introduction

Generative modeling and clustering represent two fundamental and distinct approaches within the field of machine learning. Generative modeling aims to approximate the underlying distribution of data, thereby enabling the generation of new samples (Kingma & Welling, 2014; Goodfellow et al., 2014). Clustering, conversely, seeks to identify meaningful and interpretable structures within data. This is achieved through the unsupervised detection of intrinsic relationships and dependencies (Ezugwu et al., 2022), which can enhance data visualization and interpretation. TreeVAE (Manduchi et al., 2023) was recently proposed to combine these two research directions by integrating hierarchical dependencies into a deep latent variable model. TreeVAE models the distribution of data by learning the optimal tree structure of latent variables. The resulting latent embeddings are automatically organized into a hierarchical structure that mimics the hierarchical clustering process. As a result, it can generate new data via conditional sampling and perform hierarchical clustering. However, its generative performance falls short of state-of-the-art deep generative methods, and it exhibits common issues associated with VAEs, such as generating blurry images (Bredell et al., 2023). In contrast, diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) have recently gained significant attention for their image-generation capabilities.

Our work aims to bridge this gap by (a) improving the architectural design of TreeVAE, and (b) integrating a second-stage Denoising Diffusion Probabilistic Model (DDPM) that is conditioned on the cluster-specific representations learned by TreeVAE. Our proposed approach, Diffuse-TreeVAE, generates high-quality, distinct, and representative cluster-specific images; i.e. images for each leaf of the learned tree. The proposed generation process (depicted in Figure 1) goes as follows: Diffuse-TreeVAE samples the root embedding of its tree, then propagates the generations through the leaves, and, finally, it produces high-quality leaf-specific images by guiding the reverse process of the DDPM on both the leaf reconstructions and the corresponding path in the tree. The resulting leaf-specific images share common general properties (which are sampled at the root) and differ by cluster-specific features.

Our main contributions are as follows: We provide (i) a holistic approach to clustering-based generative modeling, and (ii) a novel method for controlling image synthesis in diffusion models. We show that our approach (a) overcomes previous generative limitations of VAE-based clustering methods, and (b) produces newly generated samples that are more representative of the respective clusters in the data and closer to the true image distribution.

2. Diffuse-TreeVAE

We propose Diffuse-TreeVAE[1], a two-stage generative framework that is composed of a VAE-based generative hierarchical clustering model (TreeVAE), followed by a cluster-conditional denoising diffusion probabilistic model (DDPM). This novel combination of VAEs and diffusion models extends the generator-refiner framework introduced by DiffuseVAE (Pandey et al., 2022) to hierarchical clustering tasks. Here, TreeVAE (Manduchi et al., 2023) serves

[1]Department of Computer Science, ETH Zurich, Zurich, Switzerland. Correspondence to: Jorge da Silva Gonçalves <jorge.dasilvagoncalves@inf.ethz.ch>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria.

[1]The code is publicly available at https://github.com/JoGo175/diffuse-treevae

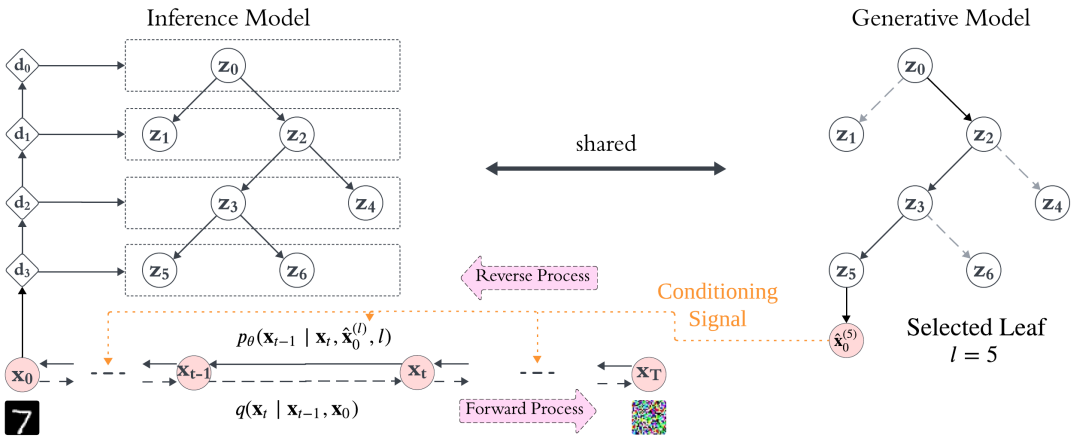

Figure 1. Schematic overview of the Diffuse-TreeVAE model: The reverse model of the DDPM (bottom) is conditioned on both the reconstruction and the index of the selected leaf l obtained from the associated, pre-trained TreeVAE. The denoising function of the DDPM learns to refine the TreeVAE-based reconstructions.

as the generator, while a DDPM (Ho et al., 2020), conditioned on the TreeVAE leaves, refines the generated images. Figure 1 illustrates the workflow of Diffuse-TreeVAE.

The first part of Diffuse-TreeVAE involves an adapted version of the TreeVAE model (Manduchi et al., 2023). TreeVAE is a generative model that intrinsically learns to hierarchically separate data into clusters via a latent tree. During training, the model grows a binary tree structure $\mathcal{T}$. The set $\mathbb{V}$ represents the nodes of the tree. Each node corresponds to a stochastic latent variable, denoted as $\mathbf{z}_0, \ldots, \mathbf{z}_V$. The parameters of these latent variables are determined by their parent nodes through neural networks called transformations. The set of leaves $\mathbb{L}$, where $\mathbb{L} \subset \mathbb{V}$, represents the clusters present in the data. Starting from the root node, $\mathbf{z}_0$, a given sample traverses the tree to a leaf node, $\mathbf{z}_l$, in a probabilistic manner. The probabilities for whether to go to the left or right child at each internal node are determined by neural networks termed routers. Thus, the latent tree encodes sample-specific probability distributions of paths. Each leaf embedding, $\mathbf{z}_l$ for $l \in \mathbb{L}$, represents the learned data representations, and leaf-specific decoders use these embeddings to reconstruct or generate new images, i.e. given a dataset $\boldsymbol{X}$, TreeVAE reconstructs $\hat{\boldsymbol{X}} = \{\hat{\boldsymbol{X}}^{(l)} \mid l \in \mathbb{L}\}$.

In this work, we improve the architectural design of the TreeVAE model. In the original TreeVAE, an initial encoder projects the images to flattened representations at the start of the bottom-up process, with the remaining components of the model relying on MLP layers. We adapted our TreeVAE method to use convolutional layers throughout the model structure instead of MLP layers. Thus, our adaptation avoids flattening the representations and instead utilizes lower-dimensional representations with multiple channels throughout the model. Additionally, we incorporated residual connections to enhance the training and performance of the model. These modifications aim to preserve spatial information and enable more efficient learning, contributing to the overall effectiveness of our model. However, it is important to note that this model suffers from the typical VAE issue of producing blurry image generations (Bredell et al., 2023). Despite this limitation, the reconstructed images and learned clustering still provide meaningful representations of the data, which are utilized in the second stage of our proposed Diffuse-TreeVAE framework.

Our model leverages the cluster assignments and image generation capabilities of TreeVAE to guide a second-stage diffusion model, specifically a DDPM (Ho et al., 2020). We adopt and adapt the generator-refiner framework (Pandey et al., 2022), where the VAE generates the initial, typically blurred images and the conditioned DDPM refines these reconstructions to produce sharper, higher-quality images. Instead of employing a conventional VAE, our model integrates TreeVAE. During training and inference, the selected leaf is randomly sampled based on the leaf probabilities learned by TreeVAE. The selected leaf reconstruction, along with the leaf index as the cluster signal, conditions the DDPM reverse process, as depicted in Figure 1. Formally, given the input data $\boldsymbol{X}$, here denoted as $\boldsymbol{X}_0$, we define a sequence of T noisy representations of the input $\boldsymbol{x}_0$, yielding $\boldsymbol{x}_{1:T}$. The forward process, $q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)$, that gradually destroys the structure of each data sample follows the standard DDPM process (Ho et al., 2020). The reverse process, on the other hand, is conditioned on the TreeVAE reconstructions $\hat{\boldsymbol{x}}_0 = \{\hat{\boldsymbol{x}}_0^{(l)} \mid l \in \mathbb{L}\}$ and on the

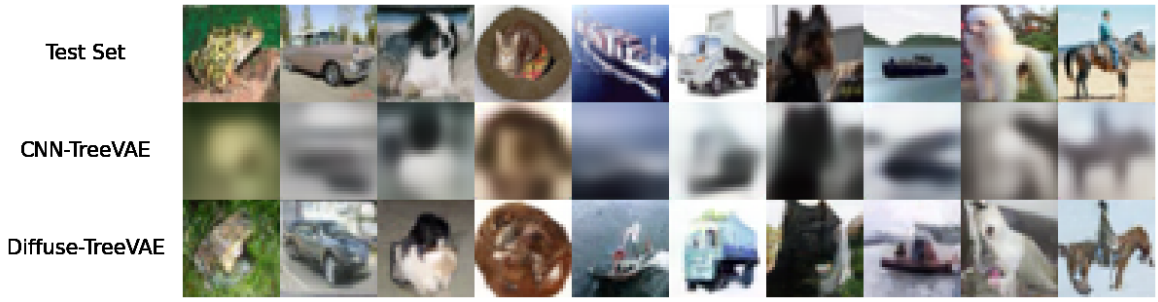

Figure 2. (Top) Samples from the CIFAR-10 test set. (Middle) Reconstructions from the CNN-TreeVAE model. (Bottom) Refined reconstructions from the Diffuse-TreeVAE model, conditioned on the CNN-TreeVAE reconstructions and the corresponding leaves.

leaf assignments:

$$l \sim p(l|\boldsymbol{x}_0),$$
$$p_\psi\Big(\boldsymbol{x}_{0:T}\,|\,\hat{\boldsymbol{x}}_0^{(l)}, l\Big) = p\left(\boldsymbol{x}_T\right)\prod_{t=1}^{T} p_\psi\Big(\boldsymbol{x}_{t-1}\,|\,\boldsymbol{x}_t, \hat{\boldsymbol{x}}_0^{(l)}, l\Big), \quad (1)$$

where $p(l|\boldsymbol{x}_0)$ is the probability that the sample $\boldsymbol{x}_0$ is assigned to leaf l. This method ensures that leaves with smaller assignment probabilities are considered, encouraging the DDPM to perform effectively across all leaves. Consequently, our approach addresses the distinct clusters inherent in TreeVAE, allowing the model to adapt specifically to different clusters and encouraging cluster-specific refinements in the images. This guidance in the image generation process assists the denoising model in learning cluster-specific image reconstructions. On the other hand, the forward noising process remains unconditional. Diffuse-TreeVAE directly utilizes the reconstructions instead of the latent embeddings for conditioning, as there exists a deterministic relationship between leaf embeddings and leaf reconstruction, provided by the leaf-specific decoder.

By using the generator-refiner framework, Diffuse-TreeVAE maintains the same clustering performance as the underlying TreeVAE. The DDPM refines the generated output samples without influencing the cluster assignments in the TreeVAE model. This is achieved through a two-stage training strategy, where the conditional DDPM is trained using a pre-trained CNN-TreeVAE model. Thus, Diffuse-TreeVAE combines the effective clustering of TreeVAE with the superior image generation capabilities of diffusion models.

3. Results

We evaluate the generative performance of our model on the MNIST (Lecun et al., 1998), FashionMNIST (Xiao et al., 2017), and CIFAR-10 (Krizhevsky, 2009) datasets. Our analysis compares three models: the original MLP-based TreeVAE, referenced as MLP-TreeVAE (Manduchi et al., 2023); our CNN-based adaptation, referred to as CNN-TreeVAE; and our novel proposal, the Diffuse-TreeVAE model. The latter model is conditioned on reconstructions and clusters derived from the CNN-TreeVAE, enhancing its capacity for generative performance. Reconstruction performance is assessed through the FID score (Heusel et al., 2017), calculated for the reconstructed images sourced from the 10,000 samples within the test set. Similarly, generation performance is evaluated using the FID score, this time computed for 10,000 newly generated images.

Table 1. Test set generative performances of different TreeVAE models. Lower FID scores indicate better performance. Means and standard deviations are computed across 10 runs with different seeds. The best result for each dataset is marked in **bold**.

Dataset	Method	FID (rec)	FID (gen)
MNIST	MLP-TreeVAE	25.8 ± 0.4	25.3 ± 1.0
	CNN-TreeVAE	24.9 ± 1.1	22.8 ± 1.4
	Diffuse-TreeVAE	**1.5** ± 0.1	**16.2** ± 5.7
Fashion	MLP-TreeVAE	44.7 ± 0.6	46.8 ± 0.9
	CNN-TreeVAE	36.5 ± 0.6	39.0 ± 0.8
	Diffuse-TreeVAE	**4.1** ± 0.6	**4.2** ± 0.5
CIFAR10	MLP-TreeVAE	225.5 ± 3.3	237.0 ± 4.0
	CNN-TreeVAE	190.5 ± 2.0	200.9 ± 2.5
	Diffuse-TreeVAE	**15.4** ± 0.3	**22.3** ± 0.3

Table 1 illustrates the generative performance across the various datasets. In every case, the CNN-TreeVAE demonstrates improvements compared to the original model. However, despite being lower, its FID scores remain at a similar level. Hence, the CNN-TreeVAE model continues to generate visibly blurry images. On the other hand, the Diffuse-TreeVAE significantly enhances the generative capabilities of the model, yielding much lower FID scores, often by an order of magnitude. This improvement is evident in the quality of the generated images, as depicted in Figure 2. Here, we visually compare the reconstructions generated by the Diffuse-TreeVAE model with those produced by the underlying CNN-TreeVAE model, which was used to condition the Diffuse-TreeVAE along with the cluster signal. Specifi-

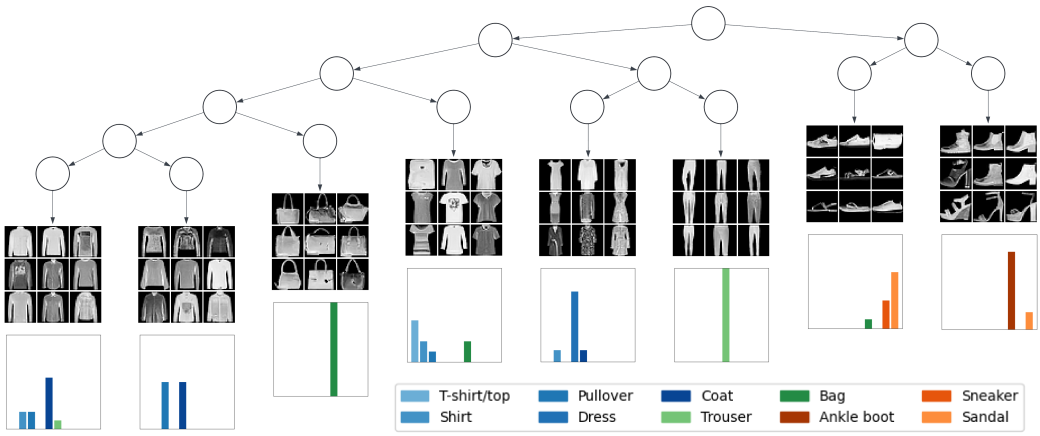

Figure 3. Diffuse-TreeVAE model trained on FashionMNIST. For each cluster, random newly generated images are displayed. Below each set of images, a normalized histogram (ranging from 0 to 1) shows the distribution of predicted classes from an independent, pre-trained classifier on FashionMNIST for all newly generated images in each leaf with a significant probability of reaching that leaf.

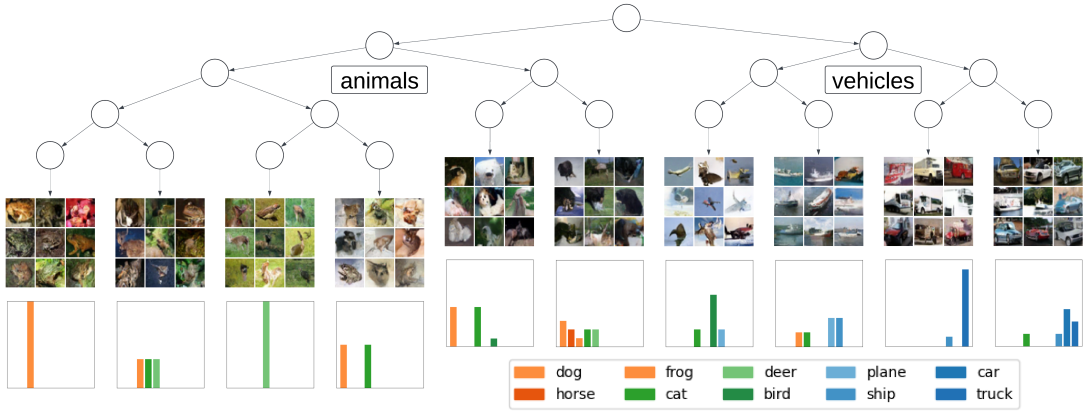

Figure 4. Diffuse-TreeVAE model trained on CIFAR-10. For each cluster, random newly generated images are displayed. Below each set of images, a normalized histogram (ranging from 0 to 1) shows the distribution of predicted classes from an independent, pre-trained classifier on CIFAR-10 for all newly generated images in each leaf with a significant probability of reaching that leaf.

cally, it can be observed that the reconstructions generated by the Diffuse-TreeVAE are notably sharper and thus exhibit closer adherence to the true distribution of the test data. The model demonstrates enhanced capability in reproducing fine details within the images while preserving the overall color and structure. However, it is important to note that these improvements may introduce some inconsistencies, resulting in reconstructions that appear more realistic but deviate slightly from the original image being reconstructed.

To assess the quality of the newly generated images, we train a classifier on the original dataset using the training data and then utilize it to classify the newly generated images from our Diffuse-TreeVAE. Specifically, we classify the newly generated images for each cluster separately. Ideally, the majority of generated images from a cluster are classified into one or very few classes from the original dataset. The more generations from a cluster that are classified into one class only, the "purer" or "less ambiguous" we consider the generations to be. For this classification task, we utilize a ResNet-50 model (He et al., 2016) trained on each dataset.

In Figure 3, we present randomly generated images from a Diffuse-TreeVAE model trained on FashionMNIST. Notably, the model in this instance has identified only seven clusters instead of the expected ten. These clusters tend to group various clothing items together, such as "Shirt", "T-Shirt", and "Pullover". Below the generated images, normalized histograms depict the distribution of the predicted classes by the classifier on the newly generated images. For instance, clusters representing trousers and bags appear to accurately and distinctly capture their respective classes, as all their generated images are classified into one group only. Conversely, certain clusters manifest a mixture of classes, indicating that they are grouped together. This observation is further supported by the histograms. Similar results can

Figure 5. Image generations from each leaf of (top) a CNN-TreeVAE, (middle) a cluster-unconditional Diffuse-TreeVAE, and (bottom) a cluster-conditional Diffuse-TreeVAE, all trained on CIFAR-10. Each row displays the generated images from all leaves of the specified model, starting with the same sample from the root. The corresponding leaf probabilities are shown at the top of the image and are by design the same for all models.

be observed for the CIFAR-10 or MNIST data, as shown in Figure 4 and Figure 6 respectively.

To assess whether the additional conditioning on the selected leaf index helps create more cluster-specific representations, we perform an ablation study. This study compares the generations of two Diffuse-TreeVAE models which only differ in one aspect: one conditioned only on the reconstructions and the other conditioned on both the reconstructions and the leaf index. For this ablation, we use the previously defined independent classifier to create histograms for each leaf to evaluate how cluster-specific the newly generated images are. As previously mentioned, ideally, the majority of generated images from a cluster should be classified into one or very few classes from the original dataset. To quantify this, we compute the average entropy for all leaf-specific histograms. Lower entropy indicates less variation in the histograms, and thus more leaf-specific generations. Table 2 presents the results for the unconditional and conditional Diffuse-TreeVAEs across all datasets. The conditional model consistently shows lower mean entropy, indicating that additional conditioning on the leaf indices indeed helps guide the model to generate more distinct and representative images for each leaf. Figure 5 visually presents the leaf generations for one sample of these models alongside the underlying CNN-TreeVAE generations, which were used to condition both models. Further examples can be found in A.2. It can be observed that both the unconditional and conditional models exhibit a significant improvement in image quality. However, the images in the cluster-conditional model are more diverse, demonstrating greater adaptability for each cluster. This is evident as the images clearly show indications of multiple true CIFAR-10 classes, with recognizable features such as horses, ships, or cars. Notably, across all models, the leaf-specific images share common properties sampled at the root while varying in cluster-specific features from leaf to leaf within each model.

Table 2. Cluster-specificity of Diffuse-TreeVAE generations for cluster-unconditional and cluster-conditional reverse models, measured by mean entropy. Lower entropy indicates more cluster-specific generations. Mean entropy is computed across all leaf-specific histograms of the predicted classes for newly generated images. The best result for each dataset is marked in **bold**.

Dataset	Method	Mean Entropy
MNIST	unconditional	1.24
	conditional	**0.13**
Fashion	unconditional	**0.66**
	conditional	**0.66**
CIFAR10	unconditional	1.12
	conditional	**0.82**

4. Conclusion

In this work, we present Diffuse-TreeVAE, a novel approach to integrate hierarchical clustering into diffusion models. By enhancing TreeVAE with a Denoising Diffusion Probabilistic Model conditioned on the cluster-specific representations, we have developed a model capable of generating distinct, high-quality images that faithfully represent their respective data clusters. This approach not only improves the visual fidelity of generated images but also ensures that these representations are true to the underlying data distribution. Diffuse-TreeVAE offers a robust framework that bridges the gap between clustering precision and generative performance, thereby expanding the potential applications of generative models in areas requiring detailed and accurate visual data interpretation.

Acknowledgments and Disclosure of Funding

Laura Manduchi is supported by the SDSC PhD Fellowship #1-001568-037. Moritz Vandenhirtz is supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number MB22.00047.

References

Bredell, G., Flouris, K., Chaitanya, K., Erdil, E., and Konukoglu, E. Explicitly Minimizing the Blur Error of Variational Autoencoders. In *The Eleventh International Conference on Learning Representations*, 2023.

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., and Akinyelu, A. A. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743, 2022. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2022.104743.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016. doi: 10.1109/CVPR.2016.90. ISSN: 1063-6919.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. April 2009.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 1558-2256. doi: 10.1109/5.726791.

Manduchi, L., Vandenhirtz, M., Ryser, A., and Vogt, J. Tree Variational Autoencoders. In *Advances in Neural Information Processing Systems*, volume 36, December 2023.

Pandey, K., Mukherjee, A., Rai, P., and Kumar, A. DiffuseVAE: Efficient, Controllable and High-Fidelity Generation from Low-Dimensional Latents. *Transactions on Machine Learning Research*, August 2022. ISSN 2835-8856.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms, September 2017. arXiv:1708.07747 [cs, stat].

A. Appendix

A.1. Generations on MNIST and CIFAR-10

Figure 6 presents an additional plot similar to those in Figure 3 and Figure 4 from the main text. This plot illustrates the generated images of the Diffuse-TreeVAE model when trained on the MNIST dataset. In each of these plots, we display randomly generated images for each cluster. Below each set of leaf-specific images, we provide a normalized histogram showing the distribution of predicted classes by an independent ResNet-50 classifier that has been pre-trained on the training data of the respective dataset. This visualization helps in understanding how well the model can generate distinct and meaningful clusters in the context of different datasets.

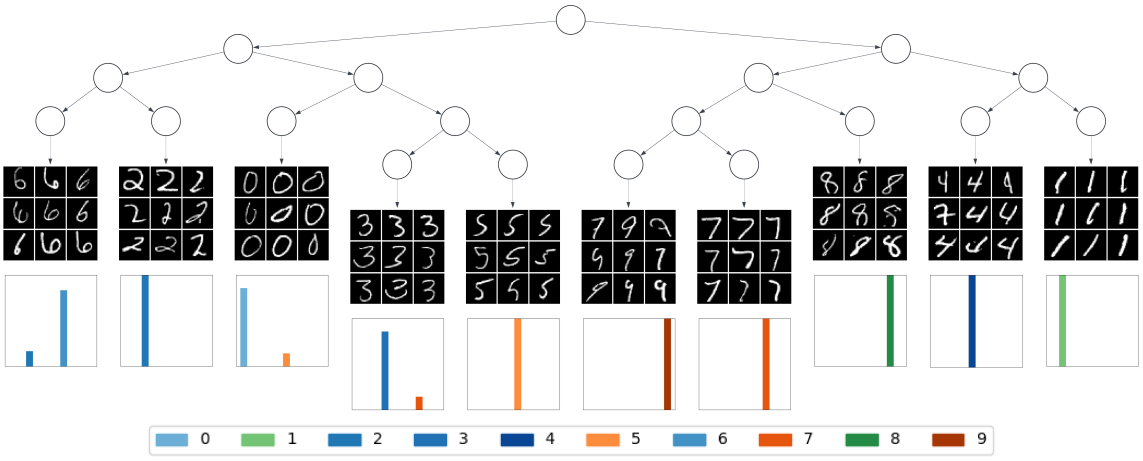

Figure 6. Diffuse-TreeVAE model trained on MNIST. For each cluster, random newly generated images are displayed. Below each set of images, a normalized histogram (ranging from 0 to 1) shows the distribution of predicted classes from an independent, pre-trained classifier on MNIST for all newly generated images in each leaf with a significant probability of reaching that leaf.

A.2. Additional Generation Examples for Conditional vs. Unconditional Diffuse-TreeVAE

Figure 7 presents three additional examples similar to Figure 5, comparing the leaf-specific generations of the conditional and unconditional Diffuse-TreeVAE models, alongside the underlying CNN-TreeVAE generations.

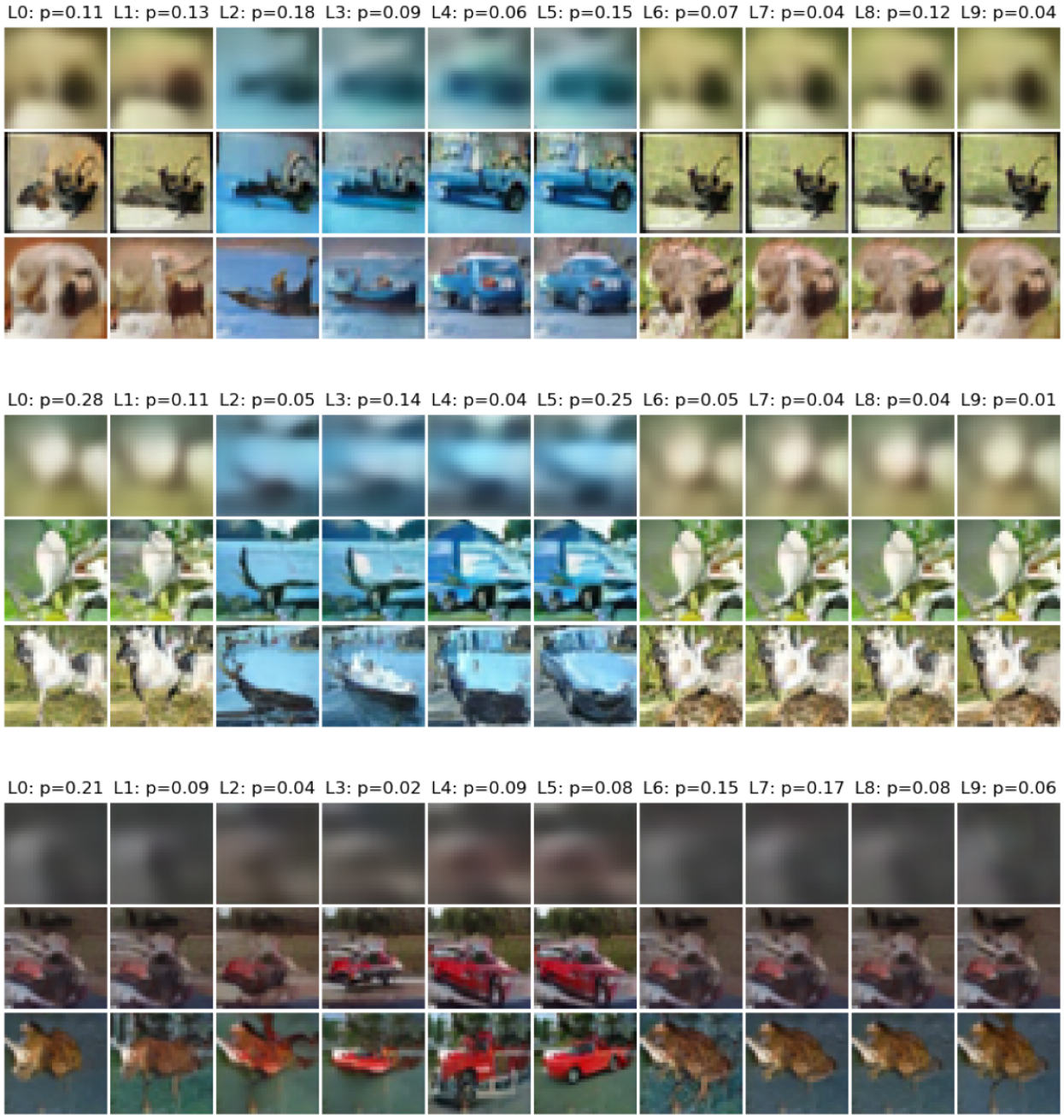

Figure 7. For each example, we show image generations from each leaf of (top) a CNN-TreeVAE, (middle) a cluster-unconditional Diffuse-TreeVAE, and (bottom) a cluster-conditional Diffuse-TreeVAE, all trained on CIFAR-10. Each row displays the generated images from all leaves of the specified model, starting with the same sample from the root. The corresponding leaf probabilities are shown at the top of the image and are by design the same for all models.