TreeDiffusion: Hierarchical Generative Clustering for Conditional Diffusion

Jorge da Silva Gonçalves (✉), Laura Manduchi, Moritz Vandenhirtz, and Julia E. Vogt

ETH Zurich, Rämistrasse 101, 8092 Zurich, Switzerland
jorge.dasilvagoncalves@inf.ethz.ch

Abstract. Generative modeling and clustering are conventionally distinct tasks in machine learning. Variational Autoencoders (VAEs) have been widely explored for their ability to integrate both, providing a framework for generative clustering. However, while VAEs can learn meaningful cluster representations in latent space, they often struggle to generate high-quality samples. This paper addresses this problem by introducing TreeDiffusion, a deep generative model that conditions diffusion models on learned latent hierarchical cluster representations from a VAE to obtain high-quality, cluster-specific generations. Our approach consists of two steps: first, a VAE-based clustering model learns a hierarchical latent representation of the data. Second, a cluster-aware diffusion model generates realistic images conditioned on the learned hierarchical structure. We systematically compare the generative capabilities of our approach with those of alternative conditioning strategies. Empirically, we demonstrate that conditioning diffusion models on hierarchical cluster representations improves the generative performance on real-world datasets compared to other approaches. Moreover, a key strength of our method lies in its ability to generate images that are both representative and specific to each cluster, enabling more detailed visualization of the learned latent structure. Our approach addresses the generative limitations of VAE-based clustering approaches by leveraging their learned structure, thereby advancing the field of generative clustering.

Keywords: Generative Modeling · Hierarchical Clustering · Conditional Diffusion.

1 Introduction

Generative modeling and clustering are two fundamental yet different tasks in machine learning. Generative modeling focuses on approximating the underlying data distribution, enabling the generation of new samples [15,8]. Clustering, on the other hand, seeks to uncover meaningful and interpretable structures within data through the unsupervised detection of intrinsic relationships and dependencies [38,7], facilitating better visualization and interpretation of the data. By integrating hierarchical dependencies into a deep latent variable model, Tree-VAE [20] was recently proposed to bridge these two research directions. While

TreeVAE is effective at hierarchical clustering, it falls short in generating high-quality images. Like other VAE-based models, it faces common issues such as producing blurry outputs [4]. In contrast, diffusion models [31,12] have recently gained prominence for their superior image generation capabilities, progressively refining noisy inputs to produce sharp, realistic images.

Our work bridges this gap by introducing a second-stage diffusion model that is conditioned on the hierarchical cluster representations learned by TreeVAE. The proposed framework, **TreeDiffusion**, combines the strengths of both models to generate high-quality, cluster-specific images, achieving improved performance in image generation. The generative process begins by sampling the root embedding of a latent tree, which is learned during training. From there, the sample is propagated from the root to one leaf by (a) sampling a path through the tree and (b) applying a sequence of stochastic transformations to the root embedding along the chosen hierarchical path. Subsequently, the diffusion model harnesses the hierarchical information by conditioning its reverse diffusion process on the sampled path representations of the latent tree through a path encoder. A key strength of TreeDiffusion is its ability to generate images tailored to each cluster, providing enhanced visualization of the learned representations, as demonstrated by our qualitative results. For the same sample, our method can produce leaf-specific images that share common general properties but differ by features encoded in the latent hierarchy. Moreover, this approach overcomes the generative limitations of VAE-based hierarchical clustering models like TreeVAE while preserving their clustering performance.

Generative clustering finds application in domains with abundant unlabeled data, where both group discovery and synthetic data generation are valuable. In the medical domain, for example, our model could aid in identifying subgroups within image data, while simultaneously providing visualizations that enhance the interpretation of the discovered groupings. Moreover, once meaningful clusters have been identified, the ability to generate representative samples enables data augmentation for downstream tasks.

1.1 Main Contributions

Our main contributions include (i) a unified framework that integrates hierarchical clustering into diffusion models, and (ii) a novel mechanism for controlling image synthesis based on learned clusters. We demonstrate that our approach (a) surpasses the generative limitations of VAE-based clustering models, and (b) produces samples that are both more representative of their respective clusters and closer to the true data distribution than models without hierarchical clustering integration.

2 Related Work

Variational Approaches for Hierarchical Clustering. Since their introduction, Variational Autoencoders (VAEs) [15] have been widely used for clustering

tasks, due to their ability to learn structured latent representations [14]. One line of work integrates hierarchical Bayesian non-parametric priors into the latent space of VAEs by applying nested Chinese Restaurant Processes to cluster the data based on infinitely deep and branching trees [9]. Another approach, TreeVAE [20], models the data distribution by learning an optimal tree structure of latent stochastic variables. This results in latent embeddings that are automatically organized into a hierarchy, mimicking the hierarchical clustering process. Single-cell TreeVAE [34] extends this framework to single-cell RNA sequencing data by incorporating batch correction, which facilitates biologically plausible hierarchical structures. Despite their strong clustering performance, these models often exhibit limited generative quality, with few providing quantitative or qualitative evaluations of their sample generation capabilities. Aside from generative approaches, discriminative deep hierarchical clustering methods include DeepECT [21] and CoHiClust [42]. Although not directly designed for clustering, ClusterNet [5] is a 3D object classification model that leverages hierarchical clustering to improve the quality of its learned representations. In this work, however, we focus on generative hierarchical clustering and its use as a conditioning signal for diffusion models.

Diffusion Models. Diffusion models have become state-of-the-art for image generation tasks over the past few years [31,12,32,22,6,33,28,26]. One drawback of diffusion models is that their latent variables lack interpretability compared to the latent spaces of VAEs. To take advantage of the strengths of both approaches, researchers have explored architectures that combine the more interpretable latent spaces of VAEs with the advanced generative capabilities of diffusion models. Notable examples include DiffuseVAE [24], Diffusion Autoencoders [25], and InfoDiffusion [37]. Furthermore, representation-conditioned image generation [18] illustrates how self-supervised learning can improve generative diffusion frameworks in unsupervised settings, reducing the gap between class-conditional and unconditional image generation.

Connecting Diffusion with Clustering. The research most closely related to our work focuses on using clustering as conditioning signals for diffusion models to improve generative quality. One approach [1] utilizes cluster assignments from k-means or TEMI clustering [2]. Similarly, another one [13] introduces a framework that employs the k-means clustering algorithm as an annotation function, generating self-annotated image-level, box-level, and pixel-level guidance signals. Both studies demonstrate the benefits of conditioning on clustering information to improve generative performance without going into the specifics of clustering performance itself. In contrast, our work further investigates which types of clustering information are most beneficial for the model, employing learned latent cluster representations alongside cluster assignments for conditioning. Related to conditioning on clusters, both kNN-Diffusion [29] and Retrieval-Augmented Diffusion Models [3] utilize nearest neighbor retrieval to condition generative models on similar embeddings, minimizing the need for large parametric models

and paired datasets in tasks like text-to-image synthesis. Diffusion models have also been applied in incomplete multiview clustering to generate missing views to improve clustering performance [39,40]. On a different note, recent research shows that training diffusion models is equivalent to solving a subspace clustering problem, explaining their ability to learn image distributions with few samples [36]. Finally, diffusion models have also been applied as a post-hoc method to enhance the generation quality of multimodal clustering models [23]. However, to the best of our knowledge, no existing diffusion model explicitly uses hierarchical clustering to enhance the interpretability and generative performance of generative clustering models.

3 Method

We propose **TreeDiffusion** [1], a two-stage framework consisting of a VAE-based generative hierarchical clustering model, followed by a hierarchy-conditional diffusion model. In the first stage, TreeVAE [20] serves as the clustering model, encoding hierarchical clusters within its latent tree structure, where the leaf nodes represent clusters. We select TreeVAE as it provides structured hierarchical latent representations from root to leaf, which are then processed by a path encoder to create the conditioning signal. In the second stage, a denoising diffusion implicit model (DDIM) [24], uses this conditioning signal to generate cluster-conditional samples. Hence, our model enables cluster-guided diffusion in unsupervised settings, analogously to classifier-guided diffusion [6] in supervised settings. Figure 1 illustrates the workflow of TreeDiffusion. The following sections provide a detailed description of each stage of the model.

3.1 Hierarchical Clustering with TreeVAE

The first part of TreeDiffusion involves a Tree Variational Autoencoder (TreeVAE) [20]. TreeVAE is a generative model that learns to hierarchically separate data into clusters through a latent tree structure. During training, the model dynamically grows a binary tree structure of stochastic variables. The process begins with a simple tree composed of a root and two child nodes, and it optimizes the corresponding ELBO over a fixed number of epochs. Afterward, the tree expands by adding two child nodes to an existing leaf node, prioritizing nodes with the highest assigned sample count to promote balanced leaves. This expansion continues iteratively, training only the subtree formed by the new leaves while freezing the rest of the model. This process repeats until the tree reaches a predefined depth or leaf count, alternating between optimizing model parameters and expanding the tree structure.

To formalize the latent tree, we adopt the original notation from TreeVAE and refer the reader to the original paper [20] for a more comprehensive introduction. Let the set $\mathbb{V}$ represent the nodes of the tree. Each node corresponds to

[1] The code and supplementary material are publicly available at `https://github.com/JoGo175/TreeDiffusion`.

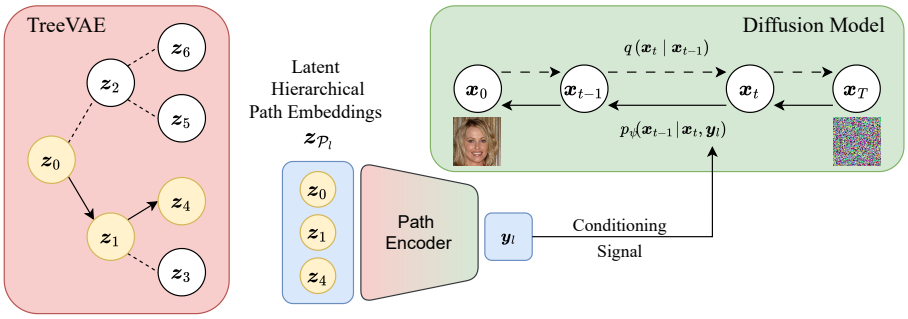

Fig. 1. Schematic overview of the TreeDiffusion framework: TreeVAE encodes data into hierarchical latent variables, where a path is sampled from the root to a leaf node. An encoder network creates a conditioning signal using the sampled hierarchical path embeddings. The diffusion model leverages this information to condition its reverse process and generate a cluster-specific image.

a stochastic latent variable, denoted as $\mathbf{z}_0, \ldots, \mathbf{z}_V$. The generative process starts at the root node, where $\mathbf{z}_0$ is sampled from a standard Gaussian distribution, i.e., $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The remaining latent variables follow a Gaussian distribution, whose parameters depend on their parent node through neural network layers. The set of leaves $\mathbb{L}$, with $\mathbb{L} \subset \mathbb{V}$, 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 of moving to the left or right child at each internal node are determined by neural networks termed *routers*. The decisions of moving to either child node are denoted by c_i for each non-leaf node i and follow a Bernoulli distribution, where $\mathrm{c}_i = 0$ indicates the selection of the left child. The path $\mathcal{P}_l$ refers to the sequence of nodes from the root to one leaf l. Moreover, let $\boldsymbol{z}_{\mathcal{P}_l} = \{\boldsymbol{z}_i \mid i \in \mathcal{P}_l\}$ denote the set of latent embeddings for each node in the path $\mathcal{P}_l$. The latent tree encodes a sample-specific probability distribution of paths. Each leaf embedding, $\mathbf{z}_l$ for $l \in \mathbb{L}$, represents the learned latent representation for that cluster. In TreeVAE, leaf-specific decoders use these embeddings to reconstruct or generate new cluster-specific images, i.e., given a dataset $\boldsymbol{X}$, TreeVAE reconstructs $\hat{\boldsymbol{X}} = \{\hat{\boldsymbol{X}}^{(l)} \mid l \in \mathbb{L}\}$. In summary, the generative model (1) and inference model (2) of TreeVAE are defined as follows:

$$p_\theta\left(\boldsymbol{z}_{\mathcal{P}_l}, \mathcal{P}_l\right) = p\left(\boldsymbol{z}_0\right) \prod_{i \in \mathcal{P}_l \setminus \{0\}} \underbrace{p\left(\mathrm{c}_{pa(i) \to i} \mid \boldsymbol{z}_{pa(i)}\right)}_{\text{decision probability}} \underbrace{p\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{pa(i)}\right)}_{\text{sample probability}} \tag{1}$$

$$q\left(\boldsymbol{z}_{\mathcal{P}_l}, \mathcal{P}_l \mid \boldsymbol{x}\right) = q\left(\boldsymbol{z}_0 \mid \boldsymbol{x}\right) \prod_{i \in \mathcal{P}_l \setminus \{0\}} q\left(\mathrm{c}_{pa(i) \to i} \mid \boldsymbol{x}\right) q\left(\boldsymbol{z}_i \mid \boldsymbol{z}_{pa(i)}\right) \tag{2}$$

3.2 Diffusion conditioned on Hierarchical Clusters

The second part of TreeDiffusion incorporates a conditional diffusion model. We assume the same forward process as in standard Denoising Diffusion Probabilistic Models (DDPM) [12], which gradually introduces noise to the data $\boldsymbol{x}_0$ over T steps. The intermediate states, $\boldsymbol{x}_t$ for $t = 1, \ldots, T$, follow a trajectory determined by a noise schedule $\beta_1, \ldots, \beta_T$ that controls the rate of data degradation, whereby $\alpha_t = (1 - \beta_t)$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Hence, the forward process can be summarized as follows:

$$q\left(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0\right) = \prod_{t=1}^{T} q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) \tag{3}$$

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) = \mathcal{N}\left(\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right) \tag{4}$$

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right) = \mathcal{N}\left(\sqrt{\overline{\alpha}_t}\boldsymbol{x}_0, \left(1-\overline{\alpha}_t\right)\boldsymbol{I}\right) \tag{5}$$

For the reverse process, TreeDiffusion starts with random noise, similar to standard diffusion models. Our model relies exclusively on the latent hierarchical information provided by TreeVAE, which is based on the tree structure learned during the first stage. The generative process begins by sampling the root embedding of the latent tree. A path is then sampled from the root to a leaf node l, and a sequence of stochastic transformations is applied to the root embedding along this path. Specifically, the tree leaf l corresponds to the selected cluster and represents the unique path through the hierarchical structure. The hierarchical conditioning information is derived from $\boldsymbol{z}_{\mathcal{P}_l}$, the set of latent embeddings along the path from the root node to the chosen leaf. These embeddings are further processed by a dedicated path encoder, which aggregates the information to produce the conditioning signal $\boldsymbol{y}_l$:

$$\boldsymbol{y}_l = \sum_{i \in \mathcal{P}_l} \left(f_{\text{embed}}(\boldsymbol{z}_i) + f_{\text{node}}(i)\right).$$

Here, f_{embed} and f_{node} are each implemented as projection blocks consisting of two MLP layers with an activation in-between, and jointly trained with the diffusion model. For each node in the path, its embedding and corresponding node index are projected independently into the time embedding dimension of the U-Net decoder [27,22]. The resulting projections are aggregated into the unified conditioning signal $\boldsymbol{y}_l$, which is then combined with the time-step embeddings to guide the U-Net during the denoising process. Consequently, this conditioning mechanism directly influences the reverse process. Let ψ denote the parameters of the denoising model, and let $p(l|\boldsymbol{x}_0)$ be the probability that the sample $\boldsymbol{x}_0$ is assigned to leaf l in the latent tree. The reverse process can then be summarized as follows:

$$\begin{aligned} &l \sim p(l|\boldsymbol{x}_0), \\ &p_\psi(\boldsymbol{x}_{0:T} \,|\, \boldsymbol{y}_l) = p\left(\boldsymbol{x}_T\right) \prod_{t=1}^{T} p_\psi(\boldsymbol{x}_{t-1} \,|\, \boldsymbol{x}_t, \boldsymbol{y}_l), \end{aligned} \tag{6}$$

The path sampling ensures that different leaves are considered, prompting the diffusion model to perform effectively across all leaves. Consequently, our approach addresses the distinct clusters inherent to TreeVAE, allowing the model to adapt and encouraging cluster-aware refinements in the images. This guidance in the image generation process assists the denoising model in learning cluster-specific image reconstructions. Currently, sampling is limited to paths originating from the root. We leave partial sampling using subtrees as well as the exploration of alternative methods for constructing the path embeddings for future work.

The following design considerations are implemented in TreeDiffusion to achieve computational efficiency without compromising effectiveness. Due to the large number of denoising steps required, DDPM sampling can be computationally expensive. To address this issue, we opt for the DDIM sampling procedure [32] instead of the standard DDPM [12]. DDIMs significantly accelerate inference by using only a subset of denoising steps, making the process more efficient while maintaining high-quality results. Finally, by employing a two-stage training strategy, where the conditional diffusion model is trained using a pre-trained TreeVAE model, TreeDiffusion preserves the hierarchical clustering performance of TreeVAE. Hence, we can combine the effective clustering of TreeVAE with the superior image generation capabilities of diffusion models.

4 Experiments

We present a series of experiments to evaluate the performance of TreeDiffusion across various datasets. The experiments are carried out on MNIST [17], FashionMNIST [41], CIFAR-10 [16], CelebA [19], and CUBICC [23]. The CUBICC dataset is a variant of the CUB Image-Captions dataset [35,30], consisting of bird images grouped into eight distinct species. In Section 4.1, we compare the generative performance of TreeDiffusion against baseline methods. In Section 4.2, we evaluate how specific the generated images are to their source clusters and analyze how distinct images from different clusters are from one another. Finally, in Section 4.3, we perform an ablation study on conditioning signals, examining various model configurations to identify which signals most effectively improve generative performance. Specifically, we compare conditioning strategies based on hierarchical clustering, flat clustering, and no clustering information.

4.1 Generative Performance

The following analysis compares the proposed TreeDiffusion model with TreeVAE [20], and a naive hybrid baseline, referred to as "TreeVAE + Diffusion". In this hybrid approach, inspired by DiffuseVAE [24], the diffusion model refines the output image generated or reconstructed by TreeVAE, but it is not conditioned on any latent information from the hierarchical structure. In contrast, TreeDiffusion introduces a key novelty by conditioning the diffusion model on the latent hierarchical path information extracted from TreeVAE. Moreover, TreeDiffusion

initiates the denoising process from random noise rather than using the TreeVAE image outputs as the denoising starting point. This enables TreeDiffusion to leverage the structure of the latent tree for cluster-specific generation. Hence, both TreeVAE + Diffusion and TreeDiffusion use TreeVAE as the first-stage model, with the main differences lying in the conditioning mechanism and the starting point of the denoising process.

The evaluation considers both reconstruction and generative performance, measured using the Fréchet Inception Distance (FID) [11]. Reconstruction performance is assessed by computing the FID score between reconstructed images and their corresponding test set images. Generative performance is evaluated by calculating the FID score for 10,000 newly generated images. The results of this analysis are summarized in Table 1.

The naive approach (TreeVAE + Diffusion) and TreeDiffusion both achieve substantial improvements over the baseline TreeVAE, their first-stage model, reducing FID scores by approximately an order of magnitude across all datasets. The naive approach performs better on simpler grayscale datasets, primarily excelling at image reconstruction rather than generation. This highlights a tendency toward overfitting. Conversely, TreeDiffusion consistently outperforms on the more complex, real-world color datasets at generating new images. Most likely, the difference in performance stems from how the denoising process is initialized. The naive model begins denoising from TreeVAE reconstructions, thereby making it highly dependent on the reconstruction quality provided by TreeVAE. Given that TreeVAE struggles more with generating new images than with reconstruction, this limitation is propagated into the naive approach. TreeDiffusion circumvents this issue by initializing the denoising directly from noise, using only latent representations from TreeVAE. As a result, TreeDiffusion achieves a better balance between reconstruction and generation quality, leading to better FID scores on newly generated images. Figure 2 compares image reconstructions on CIFAR-10, demonstrating that both diffusion-based models significantly improve upon the image quality produced by TreeVAE.

4.2 Cluster-specific Representations

Higher quality cluster-specific generations. In Figure 3, we present randomly generated images for the CUBICC dataset for both TreeVAE and TreeDiffusion, where each column corresponds to an independently generated sample. For each generation, we first sample the root embedding; then, we sample the path in the tree and the refined representations along the selected path iteratively until a leaf is reached. The hierarchical representation is then used to condition the inference in TreeDiffusion. As can be seen, the TreeDiffusion generations show substantially higher generative quality. In the following, we examine the first generated sample from Figure 3 in more detail. For this one sample, we present the generations of all leaves in Figure 4 by propagating the corresponding root representation across all paths in the tree. Note that the selected sample shown in Figure 3 ended up stemming from leaf 3 in Figure 4. When comparing the generated images across the leaves for both models, it is evident that

Table 1. Test set generative performances of the compared models. FID scores for 10,000 samples (lower is better) computed across 10 random model initializations.

Dataset	Method	FID (rec) ↓	FID (gen) ↓
MNIST	TreeVAE	24.0 ± 0.9	21.8 ± 0.7
	TreeVAE + Diffusion	$\mathbf{1.4} \pm 0.0$	$\mathbf{1.8} \pm 0.1$
	TreeDiffusion	1.5 ± 0.0	$\mathbf{1.8} \pm 0.1$
Fashion	TreeVAE	40.7 ± 2.1	41.9 ± 2.1
	TreeVAE + Diffusion	$\mathbf{4.8} \pm 0.2$	$\mathbf{4.8} \pm 0.2$
	TreeDiffusion	5.5 ± 0.6	5.4 ± 0.4
CIFAR-10	TreeVAE	175.8 ± 1.4	188.0 ± 2.0
	TreeVAE + Diffusion	$\mathbf{12.3} \pm 0.1$	19.7 ± 0.2
	TreeDiffusion	12.5 ± 0.4	$\mathbf{17.8} \pm 0.4$
CUBICC	TreeVAE	232.5 ± 7.1	255.3 ± 8.8
	TreeVAE + Diffusion	$\mathbf{12.7} \pm 6.5$	96.0 ± 2.1
	TreeDiffusion	13.4 ± 0.9	$\mathbf{29.0} \pm 5.4$
CelebA	TreeVAE	75.2 ± 15.0	77.9 ± 5.6
	TreeVAE + Diffusion	15.4 ± 3.2	30.1 ± 7.5
	TreeDiffusion	$\mathbf{14.1} \pm 6.0$	$\mathbf{18.4} \pm 7.2$

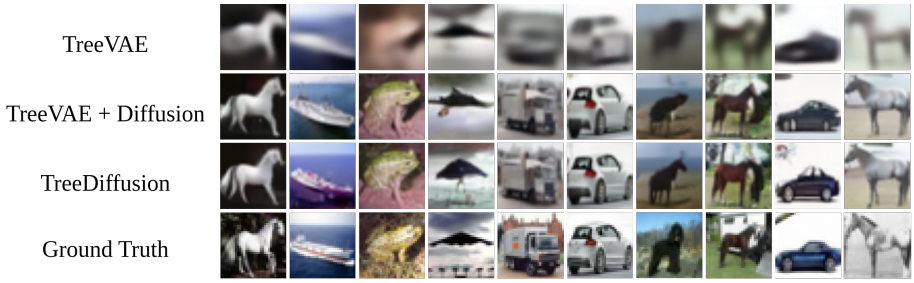

Fig. 2. Ten different CIFAR-10 reconstructions generated by the TreeVAE model, each obtained by sampling a single path in the tree. Corresponding reconstructions from TreeVAE + Diffusion, which begins denoising with the TreeVAE reconstructions, are shown alongside those from TreeDiffusion, which conditions on the same selected path and embeddings but starts denoising from noise.

TreeDiffusion not only produces sharper images for all clusters but also generates a greater diversity of images. Note that both models utilize the same latent information for image generation. While TreeVAE and TreeDiffusion maintain similar overall color distribution and structural characteristics, TreeDiffusion significantly enhances cluster specificity, resulting in images with greater clarity and distinctiveness for each cluster. Further examples of leaf-specific image generations are available in the supplementary material.

Fig. 3. Ten different samples generated by the TreeVAE model, each generated by sampling one path in the tree, and corresponding samples from the TreeDiffusion model, conditioned on the same selected path and embeddings from TreeVAE.

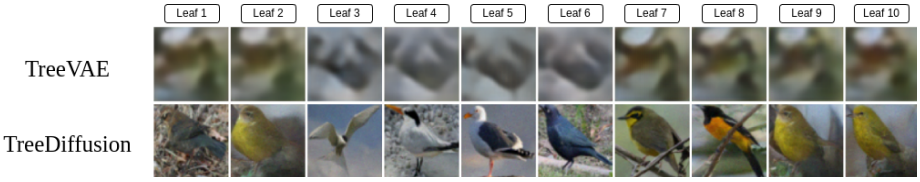

Fig. 4. Image generations from every leaf of the TreeVAE and TreeDiffusion model, both trained on the CUBICC dataset. Each row shows the generated images from all leaves of the respective model, starting with the same root sample.

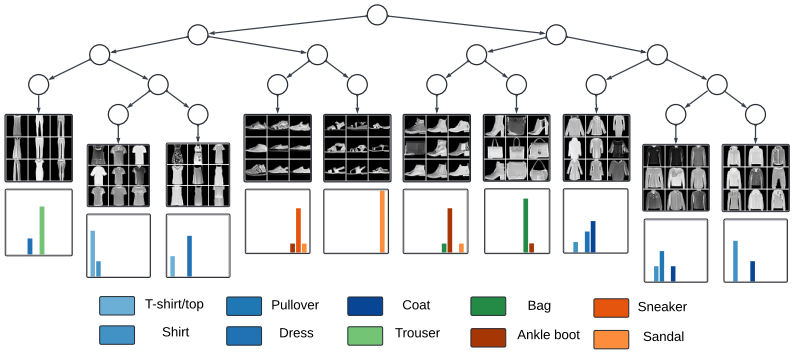

Fig. 5. TreeDiffusion model trained on FashionMNIST. For each cluster, random newly generated images are displayed. Below each set of images, a normalized histogram (ranging from 0 to 1) shows the distribution of predicted classes from an independent, pre-trained classifier on FashionMNIST for all newly generated images in each leaf with a significant probability of reaching that leaf.

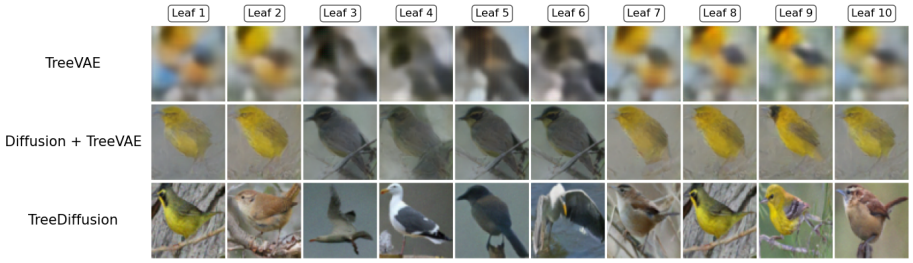

Fig. 6. Image generations from each leaf of (top) TreeVAE, (middle) TreeVAE + Diffusion which starts denoising with the TreeVAE images, and (bottom) TreeDiffusion model conditioned on the hierarchical path embeddings, all trained on CUBICC. Each row displays the generated images from all leaves of the 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.

Cluster information is retained across generations. To quantitively assess whether the newly generated images retain their cluster information, we train a classifier on the original labeled training data and then use it to classify generated images of TreeDiffusion. Specifically, we classify the generations for each cluster separately. The idea is that "pure" leaves should create samples that are classified into one or very few classes. For this classification task, we use a ResNet-50 model [10] trained on each dataset. In Figure 5, we present randomly generated images from a TreeDiffusion model trained on FashionMNIST, together with normalized histograms depicting the distribution of the predicted classes for each leaf. 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 are characterized by a mixture of classes, indicating that they are grouped together. Overall, we observe that the leaf-specific generations retain the hierarchical clustering structure found by TreeVAE, thereby enhancing the interpretability in diffusion models.

On the benefits of hierarchical conditioning. We hereby assess whether the conditioning on hierarchical representations improves cluster-specific generative quality. To this end, we compare the generations of TreeDiffusion, which is conditioned on the hierarchical representation, to the baseline TreeVAE + Diffusion from earlier, which is not conditioned on the latent cluster information. For this experiment, we use the previously introduced independent classifier to create the normalized histograms for each leaf to evaluate how cluster-specific the newly generated images are. As mentioned above, ideally, the majority of generated images from one leaf should be classified into one or very few classes from the original dataset. To quantify this, we compute the average entropy for all cluster-specific histograms. Lower entropy indicates less variation in the his-

tograms and, thus, more cluster-specific generations. Table 2 presents the results for all labeled datasets.

Table 2. Cluster-specificity of TreeDiffusion generations comparing cluster-unconditional and cluster-conditional reverse models, measured by mean entropy. Lower entropy indicates more cluster-specific generations. The best result for each dataset is marked in **bold**.

Dataset	Method	Cluster Conditioning	Mean Entropy
MNIST	Diffusion + TreeVAE	×	1.24
	TreeDiffusion	✓	**0.33**
Fashion	Diffusion + TreeVAE	×	0.66
	TreeDiffusion	✓	**0.65**
CIFAR10	Diffusion + TreeVAE	×	1.12
	TreeDiffusion	✓	**0.93**
CUBICC	Diffusion + TreeVAE	×	**0.07**
	TreeDiffusion	✓	0.20

For most datasets, the conditional model exhibits lower mean entropy, indicating that cluster conditioning indeed helps guide the model to generate more distinct and representative images for each leaf. However, for the CUBICC dataset, we observe that the mean entropy is lower for the cluster-unconditional model. This is because the classifier tends to predict all images into a single class, a result of model degeneration, where it primarily generates images for only a few classes. Figure 6 visually presents the leaf generations for one sample of these models alongside the underlying TreeVAE generations. It can be observed that both the cluster-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. Notably, across all models, the leaf-specific images share common properties, such as background color and overall shape, sampled at the root while varying in cluster-specific features from leaf to leaf within each model.

4.3 Ablation study on conditioning information

Finally, we conduct an ablation study to evaluate the impact of different conditioning signals on the generative performance. Specifically, we compare three types of conditioning: (i) hierarchical clustering signals derived from the latent embeddings of the selected cluster path $\boldsymbol{z}_{\mathcal{P}_l}$, (ii) flat clustering signals, including leaf assignment l, leaf embedding $\boldsymbol{z}_l$, or both, and (iii) an unconditioned setting where the diffusion model does not utilize any latent cluster representations from TreeVAE. Additionally, we examine the effect of using the TreeVAE leaf reconstruction $\hat{\boldsymbol{x}}_0^{(l)}$ as the starting point for the denoising process in the second-stage diffusion model. The results, outlined in Table 3, show the FID score calculated from $10,000$ samples generated using 100 DDIM steps, averaged over 10 seeds.

Note that the first row in the table represents the TreeVAE + Diffusion model from the previous experiments, whereas the last row corresponds to the proposed TreeDiffusion method.

Table 3. Effect of conditioning signals on generative performance for CIFAR-10. FID scores for 10,000 samples (lower is better) computed across 10 random model initializations.

Conditioning Type	$\hat{\boldsymbol{x}}_0^{(l)}$	l	$\boldsymbol{z}_l$	$\boldsymbol{z}_{\mathcal{P}_l}$	**FID** ↓
No Cluster Conditioning	✓				19.7 ± 0.2
Flat Clustering-Based Conditioning	✓	✓			19.1 ± 0.3
	✓		✓		18.9 ± 0.3
	✓	✓	✓		19.2 ± 0.2
		✓	✓		19.1 ± 0.5
Hierarchical Clustering-Based Conditioning	✓			✓	18.2 ± 0.3
				✓	$\mathbf{17.8} \pm 0.4$

The findings suggest that incorporating latent leaf information — whether through leaf assignment, leaf embedding, or both — significantly improves generative performance compared to relying solely on leaf reconstructions. This highlights the added benefit of conditioning on flat clustering information. Furthermore, conditioning on the full path $\boldsymbol{z}_{\mathcal{P}_l}$, which integrates all embeddings and intermediate node assignments from the root to the leaf, leads to an even greater performance boost. This underscores the effectiveness of hierarchical clustering information beyond flat clustering. As a result, harnessing $\boldsymbol{z}_{\mathcal{P}_l}$ from the hierarchical structure not only produces more structured generations, as illustrated in Figure 4, but also enhances the generative performance of generative clustering models. Notably, when conditioning on the full path, the model performs better without relying on TreeVAE reconstructions. Instead, the conditional diffusion model generates new images from scratch, guided solely by the latent information.

5 Conclusion

In this work, we present TreeDiffusion, a novel approach to integrate hierarchical clustering into diffusion models. By enhancing TreeVAE with a Denoising Diffusion Implicit Model conditioned on latent hierarchical representations, we propose 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 the generated images but also facilitates cluster visualization. TreeDiffusion offers a robust framework that bridges the gap between clustering and generative performance, thereby expanding the potential appli-

cations of generative models in areas requiring detailed and more interpretable visual data interpretation.

Acknowledgments. Jorge da Silva Gonçalves is supported by the grant #2021-911 of the Strategic Focal Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology). 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.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adaloglou, N., Kaiser, T., Michels, F., Kollmann, M.: Rethinking cluster-conditioned diffusion models. arXiv preprint arXiv:2403.00570 (2024)
2. Adaloglou, N., Michels, F., Kalisch, H., Kollmann, M.: Exploring the limits of deep image clustering using pretrained models. In: 34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023. pp. 297–299. BMVA Press (2023)
3. Blattmann, A., Rombach, R., Oktay, K., Müller, J., Ommer, B.: Retrieval-augmented diffusion models. Advances in Neural Information Processing Systems **35**, 15309–15324 (2022)
4. Bredell, G., Flouris, K., Chaitanya, K., Erdil, E., Konukoglu, E.: Explicitly Minimizing the Blur Error of Variational Autoencoders. In: The Eleventh International Conference on Learning Representations (2023)
5. Chen, C., Li, G., Xu, R., Chen, T., Wang, M., Lin, L.: Clusternet: Deep hierarchical cluster network with rigorously rotation-invariant representation for point cloud analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4994–5002 (2019)
6. Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. In: Advances in Neural Information Processing Systems. vol. 34, pp. 8780–8794. Curran Associates, Inc. (2021)
7. Ezugwu, A.E., Ikotun, A.M., Oyelade, O.O., Abualigah, L., Agushaka, J.O., Eke, C.I., 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)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014)
9. Goyal, P., Hu, Z., Liang, X., Wang, C., Xing, E.P.: Nonparametric variational auto-encoders for hierarchical representation learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5094–5102 (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016), iSSN: 1063-6919

11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
12. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
13. Hu, V.T., Zhang, D.W., Asano, Y.M., Burghouts, G.J., Snoek, C.G.: Self-guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18413–18422 (2023)
14. Jiang, Z., Zheng, Y., Tan, H., Tang, B., Zhou, H.: Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. pp. 1965–1972. Melbourne, Australia (2017)
15. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014)
16. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images (2009)
17. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
18. Li, T., Katabi, D., He, K.: Self-conditioned image generation via generating representations. arXiv preprint arXiv:2312.03701 (2023)
19. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (2015)
20. Manduchi, L., Vandenhirtz, M., Ryser, A., Vogt, J.: Tree Variational Autoencoders. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
21. Mautz, D., Plant, C., Böhm, C.: DeepECT: The Deep Embedded Cluster Tree. Data Science and Engineering **5**(4), 419–432 (2020)
22. Nichol, A.Q., Dhariwal, P.: Improved Denoising Diffusion Probabilistic Models. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8162–8171. PMLR (2021), iSSN: 2640-3498
23. Palumbo, E., Manduchi, L., Laguna, S., Chopard, D., Vogt, J.E.: Deep Generative Clustering with Multimodal Diffusion Variational Autoencoders. In: The Twelfth International Conference on Learning Representations (2023)
24. Pandey, K., Mukherjee, A., Rai, P., Kumar, A.: DiffuseVAE: Efficient, Controllable and High-Fidelity Generation from Low-Dimensional Latents. Transactions on Machine Learning Research (2022)
25. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10619–10629 (2022)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis With Latent Diffusion Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
27. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Lecture Notes in Computer Science, Springer International Publishing, Cham (2015)

28. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022)
29. Sheynin, S., Ashual, O., Polyak, A., Singer, U., Gafni, O., Nachmani, E., Taigman, Y.: knn-diffusion: Image generation via large-scale retrieval. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023)
30. Shi, Y., N, S., Paige, B., Torr, P.: Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
31. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
32. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. In: International Conference on Learning Representations (2020)
33. Vahdat, A., Kreis, K., Kautz, J.: Score-based Generative Modeling in Latent Space. In: Advances in Neural Information Processing Systems. vol. 34, pp. 11287–11302. Curran Associates, Inc. (2021)
34. Vandenhirtz, M., Barkmann, F., Manduchi, L., Vogt, J.E., Boeva, V.: sctree: Discovering cellular hierarchies in the presence of batch effects in scrna-seq data. arXiv preprint arXiv:2406.19300 (2024)
35. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset (2011)
36. Wang, P., Zhang, H., Zhang, Z., Chen, S., Ma, Y., Qu, Q.: Diffusion models learn low-dimensional distributions via subspace clustering. arXiv preprint arXiv:2409.02426 (2024)
37. Wang, Y., Schiff, Y., Gokaslan, A., Pan, W., Wang, F., Sa, C.D., Kuleshov, V.: InfoDiffusion: Representation Learning Using Information Maximizing Diffusion Models. In: Proceedings of the 40th International Conference on Machine Learning. pp. 36336–36354. PMLR (2023), iSSN: 2640-3498
38. Ward Jr, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American statistical association **58**(301), 236–244 (1963)
39. Wen, J., Deng, S., Wong, W., Chao, G., Huang, C., Fei, L., Xu, Y.: Diffusion-based missing-view generation with the application on incomplete multi-view clustering. In: Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net (2024)
40. Wen, J., Zhang, Z., Zhang, Z., Fei, L., Wang, M.: Generalized incomplete multiview clustering with flexible locality structure diffusion. IEEE transactions on cybernetics **51**(1), 101–114 (2020)
41. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms (2017), arXiv:1708.07747 [cs, stat]
42. Znalezniak, M., Rola, P., Kaszuba, P., Tabor, J., Śmieja, M.: Contrastive hierarchical clustering. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 627–643. Springer (2023)