From Simple to Complex: Extending the Generative Capabilities of Attribute-Based Latent Space Regularization through AR-VAE-Diffusion

Stephen James Krol[1,*], Abhinav Sood[2] and Maria Teresa Llano[3]

[1]SensiLab, Monash University, Caulfield East, Victoria 3145, Australia

Abstract

Progress in deep learning has driven the development of diverse creativity support tools (CST) capable of producing a range of creative artefacts. However, deep generative models are not inherently controllable, posing challenges in their guidance and prompting research focused on incorporating control mechanisms into models. One such method, Attribute-Based Latent Space Regularisation (ALSR), has demonstrated notable controllability when implemented within an Attribute-Regularised Variational Autoencoder (AR-VAE) for music and simple image generation. However, ALSR's effectiveness is constrained by the generative capabilities of the AR-VAE and is unable to control generations for high-fidelity images. In this work, we add a Denoising Diffusion Probabilistic Model (DDPM) to the AR-VAE and demonstrate that the resulting AR-VAE-Diffusion model is capable of generating and controlling high fidelity images, thus broadening the applicability of ALSR and providing a new pathway for introducing controllability into future deep learning CSTs.

Keywords

Creativity Support Tool, Latent Space Regularisation, Diffusion Model

1. Introduction

Deep generative modelling has seen significant improvements over the past 5 years, with systems now producing realistic images from text [1], music with long-term structure [2] and poetry [3]. Although some of these systems have been designed to create independent of human input [2], many have been designed as tools to aid in the creative process [4]. These tools are referred to as AI-based Creativity Support Tools (AI-CST) and have demonstrated capabilities in various creative fields and in different stages of the creative process [5]. However, while deep learning has allowed machines to produce more complex generations, many early deep generative models are limited by their lack of controllability [6, 7]. Although one could argue that controllability is not an essential component for an effective Creativity Support Tool (CST) (take for example, Brian Enos Oblique Strategy [8]), incorporating controllability into a system can contribute to more tailored creative outcomes. This has led to research focused on incorporating control mechanisms into deep learning models, enabling control over various artifacts like images and music [9, 10, 11].

*Corresponding author.

stephen.krol@monash.edu (S. J. Krol)

0000–0002–9474–3838 (S. J. Krol); 0009-0005-5891-0335 (A. Sood); 0000–0002–4898–1755 (M. T. Llano)

One such technique is Attribute-Based Latent Space Regularization (ALSR) [12] which is a method for latent space disentanglement utilised in the training of a Variational Autoencoder (VAE), a deep generative model capable of generating outputs across diverse domains [13]. Compared to other similar techniques [10, 9], ALSR provides a simple formulation that introduces controllability into a VAE by associating individual dimensions of the latent vector with specific features of the generation without adversarial training. Additionally, ALSR provides more flexibility in the type of attribute functions that can be encoded into the latent space. VAEs that utilize this regularization during training are referred to as Attribute-Regularized Variational Autoencoders (AR-VAE) and have demonstrated notable results in controlling the generations of music [11] and simple images [12]. However, while the AR-VAE demonstrates impressive controllability, it is limited by its generative capabilities for complex images, struggling to produce detailed outputs and often generating blurry artifacts - a common limitation of VAE based models [14]. This restricts the type of images that can be controlled using ALSR and thus the scope of its application.

In this work, we address this gap and demonstrate that ALSR can be used to control the generations of complex images by incorporating a Denoising Diffusion Probabilistic Model (DDPM) [15] into the AR-VAE architecture. We build upon work on VAE-Diffusion models [16] and demonstrate that controllability is maintained even with the incorporation of the diffusion model. To showcase this, we utilise two different datasets. The first is the Curl Noise dataset, which was created for this project and contains abstract flow-field images generated by an agent-based line drawing system, the second is the Kaggle abstract art dataset [17] which contains images of abstract paintings. Both datasets were selected to showcase how ALSR can be utilised in a more artistic setting. This work widens the applicable scope of ALSR, making it a plausible method for incorporating controllability of high fidelity images in AI-CSTs.

2. Background

2.1. AI-based Creativity Support Tools

With the rise of novel AI algorithms, a new generation of AI-based Creativity Support Tools (AI-CST) has been introduced to aid their use. Although the power of these tools have opened up many possibilities for artistic creation, users often struggle with different aspects of the interaction. Of relevance to this work, is the difficulty of handling unpredictable outputs [18, 19] and the lack of capabilities to explore the design space [5].

AI-CSTs can vary in the type of tasks they perform and the domain of application, with tools used for automating difficult or time consuming tasks [20, 21], for aiding ideation [22, 23, 24] and for content editing (for instance the in-painting or out-painting capabilities of Text-To-Image (TTI) systems). Although these tools enable interactions with AI models, usually through high-level representations of the output, the models behind them remain black boxes [11, 25]. This results in limited exploration capabilities, restricting the user's ability to further develop ideas and express their artistic intentions [26, 5].

Although generative AI technologies exhibit creativity largely due to their unpredictable nature, users often struggle to build upon the models' outputs, failing to acknowledge creative practice as a reflective process [27]. While some creatives have found ways to work around

these limitations [28], users still face difficulties using these systems [29]. A path to enhance the use and interaction with state of the art AI generative models is the development of mechanisms to allow the exploration of the design space. We show in this work that the use of ALSR on the VAE-Diffusion model provides a controllable mechanism at a more granular level (i.e. focusing on specific dimensions of the latent space) while maintaining the quality of complex images.

2.2. Deep Generative Modelling For Images

Denoising Diffusion Probabilistic Models (DDPM) [15] are generative models that have recently demonstrated impressive results in generating various artefacts [1, 30, 31]. The training of a diffusion model involves a diffusion process, which successively adds noise to an input, and a reverse-diffusion process, which trains a model that tries to predict how much noise should be removed at each step. During inference, noise is sampled and passed through the reverse-diffusion process to produce an output.

Recent advances in language modelling [32, 1] have demonstrated that diffusion controllability can be added through textual interfacing, commonly referred to as 'prompting'. While this method proves popular and effective for the general guidance of the generative process, arguments have been made regarding the constraints of language as an interface and how it limits creation, particularly in the realms of abstract art or technical designs [33].

Variational Autoencoders (VAE) [34] are generative models that are trained to compress and reconstruct data to a probability distribution. This paired with the Kullback-Leibler (KL) Divergence [35] allows one to sample latent vectors from a probability distribution and generate outputs using the VAEs decoder. A VAE is trained by passing an input image into the network and using the reconstructed image to calculate a reconstruction loss, i.e., how well the model could compress then recreate an artefact. Additionally, a KL-Divergence penalty is added to the loss function to ensure that the latent vectors of the model, i.e., the compressed representations, are aligned with a specified probability distribution. This allows users to sample from this distribution and use the model to generate new artefacts. Compared to other generative models, VAEs are easier to train and have a regularised latent space that allows for interpolations between generations and the addition of control vectors. This approach has proven useful in various creative domains, as seen with MusicVAE [36], which enabled users to navigate its latent space to explore different melodies or drum beats. However, when compared to diffusion models or generative adversarial networks (GANs) [6], VAEs often lack the generative capability to produce highly detailed complex imagery. To address this, work has been done on combining the VAE with a DDPM to take advantage of the VAE's low-dimensional, interpretable latent space, while still maintaining high-quality generations. This was first done in [16], where the authors built a DiffuseVAE and showcased its generative performance on the CelebA-HQ 256 dataset [37]. In this work, we enhance the DiffuseVAE model by integrating it with an AR-VAE, obtaining a more controllable model while maintaining the quality of the generated images.

2.3. Disentangling the Latent Space for Controllability

FaderNetworks [9] disentangle the latent space of an encoder-decoder model from specific attributes of the images to produce controllability. Attributes are then applied using a conditional

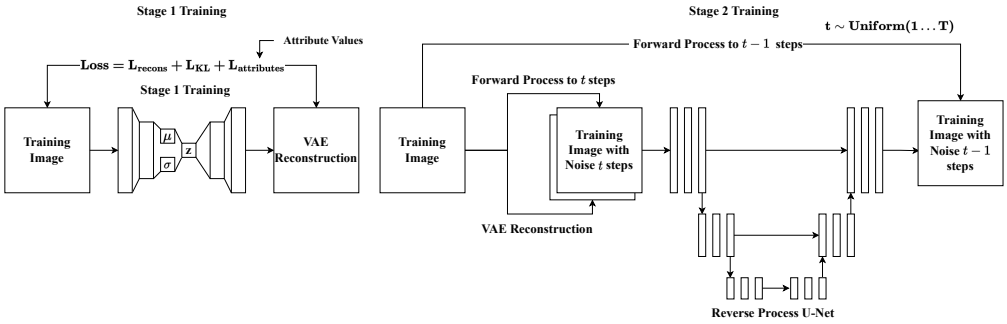


Figure 1: Two-Stage training process of our DiffuseVAE

vector of binary categorical attributes ranging from 0 to 1. Despite demonstrating controllability, FaderNetworks were limited to categorical attributes that had clear upper and lower bounds - making it difficult to apply this regularization to strictly continuous variables. Additionally, FaderNetworks incorporated adversarial learning with a discriminator to disentangle the latent space, adding an extra layer of complexity to training.

Another way to add controllability to deep generative networks is to apply constraints to the latent space. One method to do this is Attribute-Based Latent Space Regularization (ALSR) presented in [12]. Here, the authors add an extra regularization term to the loss function of a VAE so that specific dimensions of the model's latent vector correlate with attributes of the generations. Increasing or decreasing the values of these dimensions would result in the generations having more or less of these respective attributes. The authors refer to this model as an Attribute-Regularized Variational Autoencoder (AR-VAE) and demonstrate that it adds significant controllability to generations of the MNIST dataset [38] as well as generations of monophonic measures of music [11]. Additionally, compared to Geodesic Latent Space Regularization (GSLR) [10], ALSR works with non-differentiable attribute functions, expanding the possibilities of regularized attributes. However, despite the impressive controllability, these models are restricted by the generative limitations of the VAE where outputted images are often blurry and lack the fine resolution of other generative models [6, 15], justifying our approach for an AR-VAE-Diffusion model.

3. AR-VAE-Diffusion Model

Our AR-VAE-Diffusion model is an extension of the DiffuseVAE [16], wherein the VAE now incorporates a regularisation term that corresponds to the AR-VAE [12]. This simple modification allows for the introduction of attributes into a system with high image quality. We refer to the introduced term L_{arr} as the attribute loss which is computed as per Algorithm 1.

Thus the modified training objective of the VAE in the VAE-Diffusion model is of the form:

$$L_{AR-VAE} = L_{recons} + \beta * L_{KL} + \gamma * L_{arr}$$

Here, L_{recons} is the reconstruction loss for which we use the mean squared error, L_{KL} is the KL-Divergence between the VAE's latent distributions and a standard normal distribution, and

Algorithm 1 Computation of attribute loss for one mini-batch with m training examples

Input: **Training Examples**: $x_1 \ldots x_m$
Attribute Values for given training examples:
$a_1 = \begin{bmatrix} a_{11} & \cdots & a_{1m} \end{bmatrix} \ldots a_n = \begin{bmatrix} a_{n1} & \cdots & a_{nm} \end{bmatrix}$
Latent Values corresponding to the dimension of the latent space we want to regularise our attributes against:
$z_1 = \begin{bmatrix} z_{11} & \cdots & z_{1m} \end{bmatrix} \ldots z_n = \begin{bmatrix} z_{n1} & \cdots & z_{nm} \end{bmatrix}$
where m is the size of our mini-batch, n is the number of attributes.
Output: Attribute Loss for mini-batch
forEach $arr \in attributes$ **do**
$a_{mm} \leftarrow \begin{bmatrix} a_{arr}^T & \cdots & a_{arr}^T \end{bmatrix}$ ▷ stack a_{arr}^T m times
$z_{mm} \leftarrow \begin{bmatrix} z_{arr}^T & \cdots & z_{arr}^T \end{bmatrix}$ ▷ stack z_{arr}^T m times
$Da_{arr} \leftarrow a_{mm} - a_{mm}^T$
$Dz_{arr} \leftarrow z_{mm} - z_{mm}^T$
end forEach
$L_{arr} \leftarrow \sum_{attributes} \text{MAE}(tanh(\delta Dz_{arr}) - sgn(Da_{arr}))$ where sgn is the sign function and MAE is the mean absolute error. δ is a tunable hyperparameter to control the spread of the posterior.

L_{arr} is the attribute-based regularization loss. More details regarding ALSR are described in the AR-VAE paper [12]. The use of ALSR introduces two new hyper-parameters during training, γ and δ (in **Algorithm 1**). γ can be used to specify the strength of the regularisation, and δ is a tunable hyperparameter to control the spread of the posterior.

Figure 1 provides insight into the training and high-level structure of our AR-VAE-Diffusion Model. In our model, only the reverse process of the DDPM is conditioned on the VAE reconstructions. This is consistent with the first formulation presented in the original DiffuseVAE paper [16].

During inference, the process to control a generated output can be described as follows:

Let $z_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I_n})$ be a vector sampled from a normal Gaussian distribution with n dimensions. For each dimension i corresponding to some attribute arr, the value in the vector z_n is modified as follows:

$$z_n[i] = z_n[i] + \Delta$$

where $\Delta \in \mathbb{R}$ represents the change desired for the specific attribute. In practice we found constraining the domain of Δ yielded the best results, the domains used for each dataset was determined through trial and error and involved identifying the values of Δ which resulted in a generated images that were drastically different to the original input. These values and other hyperparameters are recorded in the code base and are based off suggestions from the original DiffuseVAE paper [16] as well as trial and error. The modified vector $\hat{z_n}$ is then fed into the VAE decoder to generate an image. This can be defined as:

$$\hat{x} = VAE_{decoder}(\hat{z_n})$$

Figure 2: Example images from our datasets compared to datasets used in previous studies. Columns 1 and 2 are MNIST and 2d-sprites, used in previous studies. Columns 3 and 4 are Curl Noise (new) and Abstract Art.

Subsequently, the reverse process of the DDPM is applied to generate the final output. This involves conditioning on the VAE-generated image. More precisely we get the generation y as:

$$y = DDPM_{reverse}(z_{latent}|\hat{x}) \text{ where } z_{latent} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{(\mathbf{h},\mathbf{w})})$$

Here, h and w represent the size of the image dataset the DDPM is trained on. The code from this project is available here[1]

4. Experiments

To evaluate the AR-VAE-Diffusion model, two separate models were trained on two different datasets to assess its capability in generating and controlling complex images. This section details the datasets and image attributes used for training and provides an overview of the metrics employed to evaluate disentanglement.

4.1. Datasets

The previous datasets used to evaluate ALSR were MNIST [38] and 2d-sprites [39], both of which contain simple images and can be seen in the first two columns of Figure 2. Since the AR-VAE has already proven its capability to control these basic images, we utilised two different datasets to evaluate the performance of the AR-VAE-Diffusion model. Both of the datasets were selected due to their complexity, which we define as the degree of intricacy and richness present within an image, and encompasses a range of factors such as details, patterns, textures and colour variations. Both datasets are also abstract in order to simulate a more artistic model and to test the model's capability on attributes that are challenging to define. Additionally, the attributes embedded within the latent space were chosen by the authors for their perceived potential to produce interesting variations on generated outputs. A comparison of our datasets vs the previous datasets can be seen in Figure 2.

[1]https://github.com/SensiLab/AR-VAE-Diffusion

4.1.1. Curl Noise

The curl noise dataset is a novel image dataset generated from an agent-based line drawing system, named Curl Noise [40]. Curl Noise, utilises 14 parameters that are used to produce abstract complex images based on flow fields. We used the Curl Noise system to generate a dataset of approximately 90000 designs of resolution 512x512 [2]. After removing blank and highly faded generations, we were left with approximately 68000 images for training and testing. This dataset is available here [41].

The image attributes used to test controllability of generations from the Curl Noise dataset are described below:

Pixel density: defined as the quantity and intensity of pixels present in the image, and was calculated as follows:

$$pixel\,density = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where x_i represents each pixel value, and n is the number of pixels in the image.

Size: defined as the minimum enclosing circle of a threshed, dilated and eroded image, and was calculated as follows:

$$min\ r$$
$$s.t.\,(x_i^2 + y_i^2) \leq r\,\forall x_i, y_i \in P$$

Where r is the radius of the circle and P is the set of (x_i,y_i) points in the design. OpenCV was utilised to both preprocess the image and calculate the size attribute. To ensure both attributes have equal importance in the Regularization of our variational auto-encoder's latent space, we standardise both attribute values.

4.1.2. Kaggle Abstract Art

The Kaggle Abstract Art [17] dataset contains 28820 512x512 RGB images of abstract art. The image attributes used to control generations from the abstract art dataset are described below:

Colour Diversity: defined as the approximate number of perceived colors in the image. We created this metric specifically for analyzing the variety of colors in an abstract art piece. This attribute is calculated as:

$$f(x) = \text{round}\left(\frac{x}{x_{\text{tolerance}}}\right) \times x_{\text{tolerance}}$$
$$\forall (r_i, g_i, b_i) \in \text{image add } (f(r_i), f(g_i), f(b_i)) \text{ to a set } A$$
$$\text{color diversity}_{\text{image}} = |A|$$
$$\text{where} \quad r_{\text{tolerance}} := 0.299 \times \text{tolerance}$$
$$g_{\text{tolerance}} := 0.587 \times \text{tolerance}$$

[2] This was done with the consent of the creator of the Curl Noise system who has allowed distribution of this dataset for non-commercial uses.

$$b_{\text{tolerance}} := 0.114 \times \text{tolerance}$$

Tolerance is a hyper-parameter that determines how close colors have to be for them to be grouped as one. We set a tolerance of 35 to effectively group colors, determined through trial and error and validated by visual inspection. The specific tolerance constants (0.299, 0.587, and 0.114) are borrowed from constants used in luminance calculation, following ITU-R Recommendation BT.601 [42]. These are used to account for the perceived brightness of different colors to the human eye. While alternative constants (0.2126, 0.7152, and 0.0722) as per ITU-R Recommendation BT.709 were tested [43], they failed to produce as visually accurate outcomes.

Structural Complexity: defined similarly to as in [44], structural complexity, attempts to measure how structurally complex an image is, modified to act as a proxy for the aesthetic complexity of abstract art images [45]. Intuitively, this is measured through image compression, an image that compresses to a smaller file is perceived as less structurally complex compared to one compressing into a larger file size. More concretely, for a given image:

1. Divide the image into patches.
2. Bin the patches into 4 values based on mean intensity.
3. Compute a compression ratio between the binned form and the original grayscale form of the image.

A larger compression ratio means that the image is more difficult to compress, and thus is more visually complex and vice versa. The specific implementation of this attribute, based off [40], is available in our code.

4.2. Disentanglement Metrics

To investigate whether the AR-VAE on its own offers better controllability and disentanglement of the latent space, along with a visual examination of our results, we utilize a variety of disentanglement metrics as used in the AR-VAE paper [12] which have demonstrated practical value in both the image and music domains. We summarize these metrics below. The implementation of these metrics has been borrowed from [46]. For all the metrics, except interpretability, we compute the mean across the attributes. We hold-out 20% of the dataset to compute the disentanglement metrics for the Curl Noise dataset. For the Abstract Art dataset, as the number of training images is already very limited, we compute the disentanglement metrics on the entire dataset.

Interpretability: Interpretability measures the existence of a simple linear probabilistic relationship between a specified attribute and the latent space [47].

Mutual Information Gap (MIG): Ideally, each attribute should only depend on one latent dimension. The Mutual Information Gap (MIG) helps us assess this property by computing the difference between the top two latent dimensions that have maximal mutual information with respect to a given attribute [48].

Modularity: Modularity measures if each latent dimension encodes information on only a single attribute. This is done by calculating the deviation from an idealized scenario where each latent dimension has high mutual information with one attribute and zero mutual information

Disentanglement Metrics	Curl Noise		Abstract Art	
	Beta-VAE	*AR-VAE*	*Beta-VAE*	*AR-VAE*
Interpretability - Pixel Density	0.6854	**0.8657**	—	—
Interpretability - Size	0.4098	**0.4875**	—	—
Interpretability - Structural Complexity	—	—	2.59717e-07	**3.52932e-05**
Interpretability - Color Diversity	—	—	6.08712e-07	**1.25237e-05**
Modularity	0.6822	**0.8129**	0.68047	**0.68367**
Mutual Information Gain	0.0827	**0.0925**	9.95232e-05	**1.111e-04**
Seperated Attribute Predictability	0.3634	**0.4758**	**5.07138e-05**	4.47036e-05
Spearman Correlation Coefficient	0.8129	**0.8779**	**0.020022**	0.020021

Table 1
Disentanglement measurements: The beta-VAE is a standard VAE without regularisation. The AR-VAE is the VAE used in our AR-VAE-Diffusion model.

with respect to all other attributes [49]. High deviations imply that the latent space is not very modular.

Separated Attribute Predictability (SAP): Much like MIG, SAP computes the difference between the top two latent dimensions that have a maximal R^2 Score (for continuous attributes) with respect to a given attribute [50].

Spearman Correlation Coefficient (SCC) Score: The Spearman Correlation Coefficient represents the degree to which the relationship between two variables can be explained by a monotonic function. The maximum value of the Spearman Correlation Coefficient between an attribute and each of the latent dimensions is the SCC score [50].

5. Results

5.1. Disentanglement Metrics

The disentanglement metrics for both a standard beta VAE and our AR-VAE can be seen in Table 1. For the Curl Noise dataset, our AR-VAE outperforms the standard Beta VAE in all disentanglement metrics. There is a noticeable difference between the controllability of pixel density and size as seen when comparing the two interpretability metrics. This can also be seen visually when comparing the controllability of Figures 4a and 4b.

For the Abstract Art Dataset, the interpretability metric for the AR-VAE is higher than that of the Beta-VAE. Between the two attributes, structural complexity displays a higher interpretability value. Among the other metrics, the results between the two models are very similar, this is discussed further in section 6.3. Whereas for Separated Attribute Predictability and the Spearman Correlation Coefficient, the Beta-Vae displays slightly larger disentanglement scores, the AR-VAE outperforms the Beta-VAE on Modularity and Mutual Information.

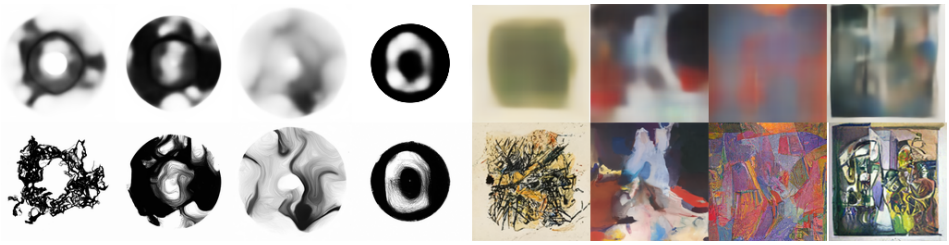
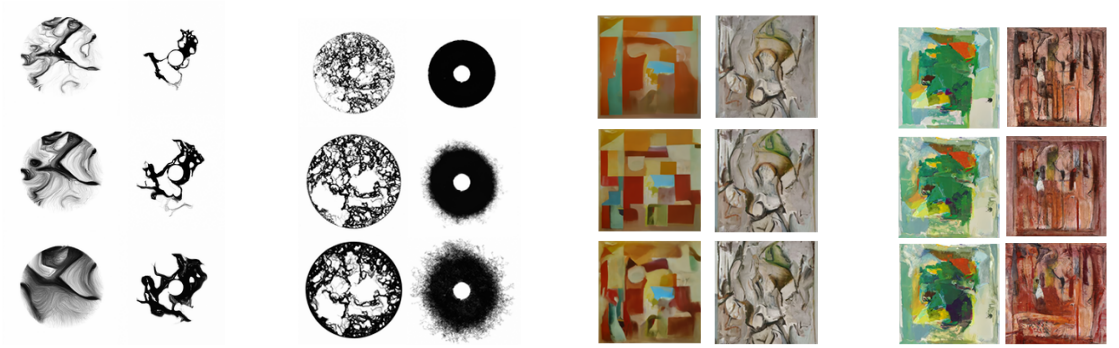


Figure 3: Comparison of AR-VAE generations (top) and AR-VAE-Diffusion (bottom) generations



(a) Curl Noise: Pixel Density

(b) Curl Noise: Generation Size

(c) Abstract Art: Structural Complexity

(d) Abstract Art: Colour Diversity

Figure 4: Examples of controllability over the four attributes for the two datasets. Each attribute has two examples with the first row representing generations from latent vectors where the attribute is lowest and the last row representing generations from latent vectors where the attribute is highest.

5.2. Controllability of Generations

Generations from the VAE-Diffusion model can be seen in Figure 3. As expected, when comparing the generations of the AR-VAE (top) to the generations of the VAE-Diffusion model (bottom), the generations of the VAE-Diffusion model are of higher quality, with images containing significantly more detail. Moreover, Figure 4 illustrates that even with the incorporation of the diffusion model, image attributes still remain controllable, with changes in attributes still maintaining the original essence of the image. More examples can be found here.[3]. The level of controllability varies noticeably across different attributes. For instance, the controllability of Pixel Density is visually more apparent than that of color diversity. Increasing attributes can also have an unexpected affect on the resulting design as shown in figure 4b where increasing size resulted in a hazy exterior being added to the second design.

[3]See Github

6. Discussion

In this work, we demonstrated that ALSR can be applied to complex data using an AR-VAE-Diffusion model, overcoming the original limitation of the AR-VAE model, which was unable to generate detailed outputs, as seen in the top row of Figure 3, where generations are blurry and undefined. We showcased the proficiency of the model to not only generate high fidelity images from two distinct datasets, but also the ability to control the generations using four different attributes (as shown in Figures 3 and 4). This broadens the scope of ALSR, unlocking the potential for novel AI-CSTs that empower users to manipulate more complex images. Additionally, compared to text-based interfaces, this method allows for fine-grained controllability of specific attributes that can be specified by the developer.

The AR-VAE-Diffusion model also offers a more straightforward training approach compared to FaderNetworks, which employ adversarial training to disentangle attributes. This mitigates potential problems associated with adversarial convergence and mode collapse [51], making the development of powerful controllable models easier.

6.1. Implications for AI-CSTs

An important aspect of AI-CSTs is to leverage unpredictable behaviours that can surprise and inspire the user [20, 5]. Although the objective of the AR-VAE-Diffusion model is to add controllability to the process, the model still retains some agency in the creative process when applying the attribute manipulations. This is illustrated by the examples in Figure 4, where it can be seen how the new generations introduce new elements to the images while still keeping the essence of previous generations. How much the new generations deviate from the previous ones depends highly on the formulation of the attributes.

Additionally, the flexibility of ALSR provides developers with the freedom to explore and encode different attribute functions. For instance, the formulation for structural complexity followed here is correlated with a measure of aesthetic complexity identified in [45]; however, this attribute could take other forms. To illustrate, we could define structural complexity based on the amount of white space an image has, or the diversity of shapes within the image, or the presence of repeated patterns. The formulation of attributes is flexible allowing developers to explore different creative possibilities. This work also has promise beyond high-fidelity image generation in fields such as controllable music generation [52] and 3D-modelling [53], as both these fields that have shown promise with diffusion models.

Finally, the nature of the technique allows for formulations of attributes that may not have been possible with other methods [10]; providing more flexibility in the type of attribute functions that can be encoded.

6.2. Considerations for Attribute Selection

Our experiments revealed that some attributes perform better than others. For example, in figure 4, changes in pixel density are more obvious than changes in colour diversity, likely due to the complexity of the colour diversity function. If an AR-VAE-Diffusion model is being designed to produce visually predictable controllability, image attribute functions must be carefully defined.

For example, in figure 4a as the size dimension is increased in the second design, the generations begin to grow a fuzzy exterior. While this may not have been predictable, it is still consistent with how size was defined in section 4.1, which is as the minimum enclosing circle around the design. Therefore, for predictable results, attribute functions must be carefully defined.

The importance of attribute selection is illustrated by the difference in the disentanglement metrics (see Table 1) between our two datasets. Compared to the attributes used in the Curl Noise dataset, the attributes used in the Abstract Art dataset exhibit significantly higher complexity. Consequently, an analysis of the disentanglement metrics (see Table 1) reveals that the latent space is not as disentangled. Nevertheless, the interpretability metrics indicate that, in contrast to a Beta-VAE, the AR-VAE still maintains a stronger linear probabilistic relationship between the attributes of interest and the latent space. Thus, by manipulating the attribute respective latent dimensions of the AR-VAE by largely increasing/decreasing their values, we can control the attributes even though the latent space is not as disentangled. Despite the complexity of the attributes, training still yields relatively controllable output, as evident from visual generations.

6.3. Limitations of the Approach

Although the AR-VAE-Diffusion model significantly improves the generative capabilities of the AR-VAE, there is an added computational cost in both training and inference, potentially limiting the applicability of the model in real-word applications for users with limited computational resources. However, as demonstrated in Figure 3, without the addition of the Diffusion model, the AR-VAE is unable to generate high fidelity generations of complex images, justifying the increased computational cost of the AR-VAE-Diffusion model. Additionally, many modern creative tools such as photoshop or ableton require decent compute resources for their operations, making the computational demands of the AR-VAE-Diffusion model less prohibitive within certain professional contexts.

7. Conclusion

In this work we demonstrate that ALSR can be applied to more complex images through the use of an AR-VAE-Diffusion model. This extends the applicable scope of ALSR making it now a plausible method for controlling the generations of high-fidelity images. Additionally, the flexibility of ALSR provides new opportunities for developers to build deep learning based AI-CSTs that provide controllability of a wide range of attributes. Future work will involve testing how different forms of attribute regularisation on the AR-VAE-Diffusion model improve the level of controllability in the model.

References

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv.org (2022).

[2] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, C. Hawthorne, A. M. Dai, M. D.

Hoffman, D. Eck, Music transformer: Generating music with long-term structure, arXiv preprint arXiv:1809.04281 (2018).
[3] N. Rajcic, J. McCormack, Mirror ritual: An affective interface for emotional self-reflection, in: Conference on Human Factors in Computing Systems - Proceedings, CHI '20, ACM, Ithaca, 2020, pp. 1–13.
[4] J. McCormack, T. Gifford, P. Hutchings, M. T. Llano, M. Yee-King, M. d'Inverno, In a silent way: Communication between ai and improvising musicians beyond sound, ACM, New York, NY, 2019. doi:https://doi.org/10.1145/3290605.3300268, paper No. 38.
[5] A. H.-C. Hwang, Too late to be creative? ai-empowered tools in creative processes, in: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, CHI EA '22, Association for Computing Machinery, 2022. doi:10.1145/3491101.3503549.
[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (2020) 139–144.
[7] D. P. Kingma, M. Welling, Auto-encoding variational bayes, 2013. URL: https://arxiv.org/abs/1312.6114. doi:10.48550/ARXIV.1312.6114.
[8] B. Eno, P. Schmidt, Oblique strategies, Opal.(Limited edition, boxed set of cards.)[rMAB] (1975).
[9] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, M. Ranzato, Fader networks: Manipulating images by sliding attributes, 2017. URL: https://arxiv.org/abs/1706.00409. doi:10.48550/ARXIV.1706.00409.
[10] G. Hadjeres, F. Nielsen, F. Pachet, Glsr-vae: Geodesic latent space regularization for variational autoencoder architectures, 2017. URL: https://arxiv.org/abs/1707.04588. doi:10.48550/ARXIV.1707.04588.
[11] N. Bryan-Kinns, B. Banar, C. Ford, C. N. Reed, Y. Zhang, S. Colton, J. Armitage, Exploring XAI for the arts: Explaining latent space in generative music, in: eXplainable AI approaches for debugging and diagnosis., 2021. URL: https://openreview.net/forum?id=GLhY_0xMLZr.
[12] A. Pati, A. Lerch, Attribute-based Regularization of Latent Spaces for Variational Auto-Encoders, Neural Computing and Applications (2020). URL: https://doi.org/10.1007/s00521-020-05270-2. doi:10.1007/s00521-020-05270-2.
[13] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).
[14] M. Sami, I. Mobin, A comparative study on variational autoencoders and generative adversarial networks, in: 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), IEEE, 2019, pp. 1–5.
[15] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, 2020. URL: https://arxiv.org/abs/2006.11239. doi:10.48550/ARXIV.2006.11239.
[16] K. Pandey, A. Mukherjee, P. Rai, A. Kumar, Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents, 2022. URL: https://arxiv.org/abs/2201.00308. doi:10.48550/ARXIV.2201.00308.
[17] G. Ogden, Abstract art, 2022. URL: https://www.kaggle.com/datasets/goprogram/abstract-art/data.
[18] G. Dove, K. Halskov, J. Forlizzi, J. Zimmerman, Ux design innovation: Challenges for working with machine learning as a design material, in: Proceedings of the 2017 chi

conference on human factors in computing systems, 2017, pp. 278–288.
[19] L. E. Holmquist, Intelligence on tap: artificial intelligence as a new design material, interactions 24 (2017) 28–33.
[20] J. J. Y. Chung, Artistic user expressions in ai-powered creativity support tools, in: Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, UIST '22 Adjunct, Association for Computing Machinery, 2022. URL: https://doi.org/10.1145/3526114.3558531. doi:10.1145/3526114.3558531.
[21] C. Yan, J. J. Y. Chung, Y. Kiheon, Y. Gingold, E. Adar, S. R. Hong, Flatmagic: Improving flat colorization through ai-driven design for digital comic professionals, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–17.
[22] D. P. Jaiswal, S. Kumar, Y. Badr, Towards an artificial intelligence aided design approach: application to anime faces with generative adversarial networks, Procedia Computer Science 168 (2020) 57–64.
[23] F. Ibarrola, O. Bown, K. Grace, Towards co-creative drawing based on contrastive language-image models, in: The 13th International Conference on Computational Creativity (ICCC'22), volume 10, 2022, p. 2.
[24] M. Zammit, A. Liapis, G. N. Yannakakis, Seeding diversity into ai art, Proceedings of the Thirteen International Conference on Computational Creativity, ICCC'22 (2022).
[25] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature machine intelligence 1 (2019) 206–215.
[26] Q. Yang, A. Steinfeld, C. Rosé, J. Zimmerman, Re-examining whether, why, and how human-ai interaction is uniquely difficult to design, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for Computing Machinery, 2020, p. 1–13. URL: https://doi.org/10.1145/3313831.3376301. doi:10.1145/3313831.3376301.
[27] D. A. Schon, The reflective practitioner: How professionals think in action, volume 5126, Basic books, 1984.
[28] N. Collins, V. Ruzicka, M. Grierson, Remixing ais: mind swaps, hybrainity, and splicing musical models, in: Proc. The Joint Conference on AI Music Creativity, 2020.
[29] R. Louie, A. Coenen, C. Z. Huang, M. Terry, C. J. Cai, Novice-ai music co-creation via ai-steering tools for deep generative models, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1–13. URL: https://doi.org/10.1145/3313831.3376739. doi:10.1145/3313831.3376739.
[30] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al., Make-a-video: Text-to-video generation without text-video data, arXiv preprint arXiv:2209.14792 (2022).
[31] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, M. Chen, Point-e: A system for generating 3d point clouds from complex prompts, 2022. URL: https://arxiv.org/abs/2212.08751. doi:10.48550/ARXIV.2212.08751.
[32] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
[33] J. McCormack, C. Cruz Gambardella, N. Rajcic, S. J. Krol, M. T. Llano, M. Yang, Is writing

prompts really making art?, in: Artificial Intelligence in Music, Sound, Art and Design: 12th International Conference, EvoMUSART 2023, Held as Part of EvoStar 2023, Brno, Czech Republic, April 12–14, 2023, Proceedings, Springer, 2023, pp. 196–211.
[34] D. P. Kingma, M. Welling, Auto-encoding variational bayes, 2022. arXiv:1312.6114.
[35] S. Kullback, R. A. Leibler, On information and sufficiency, The annals of mathematical statistics 22 (1951) 79–86.
[36] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, D. Eck, A hierarchical latent vector model for learning long-term structure in music, in: International Conference on Machine Learning (ICML), 2018. URL: http://proceedings.mlr.press/v80/roberts18a.html.
[37] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, arXiv preprint arXiv:1710.10196 (2017).
[38] L. Deng, The mnist database of handwritten digit images for machine learning research, IEEE Signal Processing Magazine 29 (2012) 141–142.
[39] L. Matthey, I. Higgins, D. Hassabis, A. Lerchner, dsprites: Disentanglement testing sprites dataset, https://github.com/deepmind/dsprites-dataset/, 2017.
[40] J. Mccormack, C. Cruz Gambardella, S. Krol, Creative discovery using quality-diversity search, in: Proceedings of the Companion Conference on Genetic and Evolutionary Computation, 2023, pp. 747–750.
[41] S. Krol, J. McCormack, A. Sood, Curl noise 90000 dataset, 2024. URL: https://bridges.monash.edu/articles/dataset/CURL_NOISE_90000_DATASET/26868943. doi:10.26180/26868943.
[42] ITU-R, Recommendation itu-r bt.601-7, 2011. https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.601-7-201103-I!!PDF-E.pdf.
[43] ITU-R, Recommendation itu-r bt.709-6, 2011. https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.709-6-201506-I!!PDF-E.pdf.
[44] S. Lakhal, A. Darmon, J.-P. Bouchaud, M. Benzaquen, Beauty and structural complexity, Phys. Rev. Res. 2 (2020) 022058. URL: https://link.aps.org/doi/10.1103/PhysRevResearch.2.022058. doi:10.1103/PhysRevResearch.2.022058.
[45] J. McCormack, C. Cruz Gambardella, Quality-diversity for aesthetic evolution, in: International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar), Springer, 2022, pp. 369–384.
[46] A. Pati, A. Lerch, Is disentanglement enough? on latent representations for controllable music generation, in: 22nd International Society for Music Information Retrieval Conference (ISMIR), Online, 2021.
[47] T. Adel, Z. Ghahramani, A. Weller, Discovering interpretable representations for both deep generative and discriminative models, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 50–59. URL: https://proceedings.mlr.press/v80/adel18a.html.
[48] R. T. Q. Chen, X. Li, R. B. Grosse, D. K. Duvenaud, Isolating sources of disentanglement in variational autoencoders, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/1ee3dfcd8a0645a25a35977997223d22-Paper.pdf.
[49] K. Ridgeway, M. C. Mozer, Learning deep disentangled embeddings with the f-

statistic loss, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/2b24d495052a8ce66358eb576b8912c8-Paper.pdf.

[50] A. Kumar, P. Sattigeri, A. Balakrishnan, Variational inference of disentangled latent concepts from unlabeled observations, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018. URL: https://openreview.net/forum?id=H1kG7GZAW.

[51] D. Saxena, J. Cao, Generative adversarial networks (gans) challenges, solutions, and future directions, ACM Computing Surveys (CSUR) 54 (2021) 1–42.

[52] G. Mittal, J. Engel, C. G.-M. Hawthorne, I. Simon, Symbolic music generation with diffusion models, 2021. URL: https://arxiv.org/abs/2103.16091.

[53] S. Luo, W. Hu, Diffusion probabilistic models for 3d point cloud generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2837–2845.