



## OPEN ACCESS

## EDITED BY

Xintao Wu,  
University of Arkansas, United States

## REVIEWED BY

Aziz Darouichi,  
Cadi Ayyad University, Morocco  
Abhiroop Chatterjee,  
Jadavpur University, India

## \*CORRESPONDENCE

Fasih Haider  
✉ Fasih.Haider@ed.ac.uk

RECEIVED 01 September 2025

ACCEPTED 28 October 2025

PUBLISHED 19 November 2025

CORRECTED 21 November 2025

## CITATION

Haider F, Moroshko E, Xue Y and Tsafaris SA (2025) Do generative models learn rare generative factors? *Front. Artif. Intell.* 8:1697139. doi: 10.3389/frai.2025.1697139

## COPYRIGHT

© 2025 Haider, Moroshko, Xue and Tsafaris. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Do generative models learn rare generative factors?

Fasih Haider\*, Edward Moroshko, Yuyang Xue and Sotirios A. Tsafaris

School of Engineering, The University of Edinburgh, Edinburgh, United Kingdom

Generative models are becoming a promising tool in AI alongside discriminative learning. Several models have been proposed to learn in an unsupervised fashion the corresponding generative factors, namely the latent variables critical for capturing the full spectrum of data variability. Diffusion Models (DMs), Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are of particular interest due to their impressive ability to generate highly realistic data. Through a systematic empirical study, this paper delves into the intricate challenge of how DMs, GANs and VAEs internalize and replicate *rare* generative factors. Our findings reveal a pronounced tendency toward memorization of these factors. We study the reasons for this memorization and demonstrate that strategies such as spectral decoupling can mitigate this issue to a certain extent.<sup>1</sup>

## KEYWORDS

**generative factors, latent variables, diffusion models (DMs), generative adversarial networks (GANs), variational autoencoders (VAEs), rare generative factors, rare generative factors (RGFs)**

## 1 Introduction

In recent years, the machine learning field has witnessed a significant increase in the popularity and advancement of generative models (Scao et al., 2022; OpenAI, 2022; Taylor et al., 2022; Zhang S. et al., 2022; Iyer et al., 2022; Touvron et al., 2023). These models have significantly advanced approaches to e.g., image generation and natural language processing, demonstrating the ability to create outputs that closely resemble real-world data (e.g., Karras et al., 2020; Zhang B. et al., 2022). The ongoing development and increasing adoption of these technologies, particularly large language models, have garnered substantial attention from academia and industry, while also becoming a topic of public interest (De Angelis et al., 2023; Mohamadi et al., 2023).

At the heart of these generative models lies the concept of *generative factors* (also known as factors of variation or latent variables), which fundamentally affect the characteristics of the generated outputs (Liu et al., 2023; Bengio et al., 2013; Higgins et al., 2018; Träuble et al., 2021). These factors encompass many elements, from simple attributes such as color or size in images to more complex features such as sentence structure or thematic elements in text. Understanding and manipulating these generative factors is a key to harnessing the full potential of generative models (Fard et al., 2023; Yang et al., 2021; Shao et al., 2018).

Despite extensive research surrounding generative models (Bond-Taylor et al., 2022), one aspect remains notably under-explored: their ability to learn and replicate *rare generative factors*. Rare generative factors (RGFs) are latent variables that are highly skewed in their frequency of appearance in the real world (and hence in datasets) but play a

<sup>1</sup> The link to the code: [https://gitlab.com/haider84/frontierai\\_rgf.git](https://gitlab.com/haider84/frontierai_rgf.git).

critical role in the underlying data-generating process. RGFs appear in a wide range of applications, including medical imaging (Liu et al., 2022), natural language generation (Mercatali and Freitas, 2021), and others.

**A motivating example** Consider a dataset composed of electrocardiogram (ECG) recordings with RGF that shows the presence of Brugada syndrome, a rare disorder that can lead to sudden cardiac arrest. This syndrome is more prevalent in people in their 30s or 40s (Speranzon et al., 2024) but can also occur in childhood (Peltenburg et al., 2022). Therefore, a dataset collected of patients with the disease is more likely to have people aged 30 to 50 years with the disease. Generative models could generate new data to enrich the diversity of the dataset, improve AI-based diagnostic tools, or facilitate the early detection of this syndrome in a wider patient population, ultimately leading to timely interventions and more precise medical prognoses. This goal requires generative models not only to replicate the distinct ECG patterns associated with the syndrome within the subset of recordings where it is predominantly found, but also to introduce these patterns into ECG recordings at other ages not commonly associated with the syndrome.

Focussing on Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models (DMs), in this paper, we take a step forward by exploring their ability to capture these rare generative factors. We introduce a framework specifically designed to examine the effect of rarity in generative factors on the learning process of generative models. Focussing on simple canonical models [i.e., the original (plain) GAN architecture (Goodfellow et al., 2014), the standard VAE, a simple diffusion probabilistic model (Ho et al., 2020)] allows us to distill insights without the confounding effects of additional complexities introduced in variant models. It helps in maintaining focus on core learning dynamics across all three model types.

By taking rarity to the extreme, considering datasets where the skew in the distribution of generative factors is pronounced, we pose a fundamental question: *When faced with a dataset that is heavily skewed in terms of the coverage of generative factors, will a generative model successfully learn rare generative factors?* Addressing this question is crucial to understanding the limits of current generative models and developing new methodologies that can better capture and represent the diversity of generative factors, especially those that are rare. This exploration not only aims to enhance the fidelity and diversity of model-generated output, but also seeks to contribute to the broader discourse on model robustness and fairness when dealing with skewed data distributions. Understanding how generative models behave under extreme data imbalance has direct implications for fairness, robustness, and trustworthy AI, where under-represented factors often correspond to minority groups, rare pathologies, or uncommon real-world conditions that are crucial for equitable model performance.

We show that plain GAN, VAE, and DM generally struggle to learn RGFs, tending instead to *memorize* them. This memorization is distinct from the memorization of individual training examples, as highlighted by recent studies. For instance, de Wynter et al. (2023) demonstrated how large language models exhibit example memorization, while Carlini et al. (2023) found that diffusion models tend to reproduce training examples during test time. Maini et al. (2023) showed that example memorization can be distributed

across various neurons and layers, and Akbar et al. (2023) demonstrated memorization in diffusion models for synthetic brain tumor images. However, to the best of our knowledge, the memorization of generative factors remains significantly under-explored in the literature (Jegorova et al., 2023).

Our work provides valuable insights into the limitations of current generative models in learning robust, transferable representations from imbalanced datasets, opening new avenues for improving their generalization capabilities.

To summarize, we make five main **contributions**:

- A framework designed to systematically study the learning of RGFs in generative models.
- A statistical testing pipeline using z-scores and p-values to quantify the extent of memorization and assess factor-wise generalization at a class-specific level, rather than relying on global distribution metrics.
- A baseline comparison using matched datasets (balanced vs. skewed) to control for confounding variables and isolate the impact of data skew on generative learning performance.
- Through an extensive empirical study, we evaluate the capability of GANs, VAEs, and DMs to learn and replicate RGFs, providing valuable insights into the dynamics of generative learning in the presence of data rarity.
- We identify and discuss the limitations in the context of RGF learning, explore the underlying reasons for these limitations, and evaluate a potential mitigation strategy specifically for GANs.

## 2 Related work

Generative models can replicate the data distribution they are trained on but this is *not* what we aim for. We focus on a crucial aspect of unsupervised feature extraction: the ability to disentangle and generalize RGF. We deliberately create skewed datasets where specific generative factors are present only in one class, not to test if models can mimic this distribution, but to examine if they can abstract these factors. Hence we focus not on how well models reproduce training data statistics, but on their capacity to learn generalizable latent representations from biased inputs. The tendency of models to memorize rare factor-class associations, rather than extending them to other classes, reveals a limitation in their ability to discover the underlying data generating process (Liu et al., 2022). This memorization of generative factors, highlights a significant challenge in unsupervised representation learning. It underscores the difficulty these models face in separating class-specific features from generalizable attributes when presented with skewed data. We also differentiate our focus on RGFs from the causal disentanglement approaches highlighted by Zhang et al. (2024). While Zhang et al. (2024) provide identifiability guarantees for disentangling causal variables using soft interventions, their emphasis lies in leveraging interventions to establish robust causal structures. Our study takes a different path, examining how generative models manage Rare generative factors (RGFs) under extreme data imbalance. Unlike causal disentanglement, we make no assumptions about causality or intervention-based data. Instead, we investigate the mechanisms behind the memorization

and generalization of RGFs, shedding light on the strengths and limitations of generative models in representing underrepresented factors (i.e., RGFs). This perspective offers a complementary angle to the causal disentanglement literature, enriching the broader discourse on disentanglement in generative modeling.

Garrido et al. (2020) primarily emphasize the prediction of rare feature combinations in population synthesis (i.e., zero-cell problem), particularly for unique categorical features, using Variational Autoencoders (VAE) and Wasserstein Generative Adversarial Networks (WGAN). While this approach provides valuable insights into modeling rare occurrences (addressing zero-cell problem), it operates at a narrow granularity that often overlooks broader dependencies and emergent patterns spanning the full image (or in raw data). This limitation risks underestimating the inherent complexity of rare variations, particularly when these involve intricate feature inter-dependencies.

Garrido et al. (2020) evaluate the learning of unique feature combinations by focusing on sampling zeros (i.e., logically possible combinations absent from the training data but present in test data) and structured zeros (i.e., logically impossible absent from training and test data). However, this methodology confines rarity to isolated feature combinations, limiting the model's ability to capture the interplay of features across the full image context. Consequently, the proposed evaluation framework lacks a controlled experimental baseline (i.e., comparison of generated sampling zeros while trained on balanced and highly skewed data). Our problem differs from the zero-cell problem because it focuses on image generation, where rare-generated factors (RGFs) dynamically affect the dataset. Some RGFs, like color, thickening, or thinning, influence the entire image globally, while others, like fractures or swelling, have random, localized effects on parts of the images. Unlike the zero-cell problem, which deals with fixed absences in specific feature combinations, our challenge involves ensuring models can learn and generalize these rare and dynamically applied factors across samples. Moreover, evaluating the zero-cell problem is straightforward when dealing with tabular data, as it involves directly comparing whether the generated samples match values present in the test data. However, this approach cannot be directly applied to image generation tasks, where the outputs are in raw form, high-dimensional, and lack discrete categories for direct comparison.

Our study addresses these gaps by implementing a controlled experimental setup and constructing datasets where rarity is defined at the level of the entire image. Additionally, an evaluation framework has also been proposed to address this issue systematically. This approach enables a more comprehensive assessment of how generative models, including VAEs, GANs, and Diffusion Models (DMs), handle rare factors (RGFs) present in raw data (i.e., images).

### 3 Preliminaries

Consider a dataset  $\{(\mathbf{x}_i, f_i, y_i)\}_{i=1}^n$ , where  $\mathbf{x}_i \in \mathcal{X}$  is a data instance,  $f_i \in \{0, 1\}$  is a binary<sup>2</sup> generative factor and  $y_i \in \{1, \dots, C\}$

2 Our work can be extended to non-binary generative factors.

is a class label. For example,  $\mathbf{x}_i$  is an image of a digit,  $f_i$  indicates the color (green for 0, red for 1), and  $y_i$  is the value of the digit.

Central to our work are the generative factors, informally defined as: [Generative Factors, informal] The generative factors are the underlying latent variables that fully characterize the variation of the data in the domain  $\mathcal{X}$ .

Our work focuses on the case of rare generative factors, formally defined as follows: [Rare Generative Factor, RGF] For  $c \in \{1, \dots, C\}$ , let  $S_{c,0} = \{i | y_i = c \text{ and } f_i = 0\}$  and  $S_{c,1} = \{i | y_i = c \text{ and } f_i = 1\}$ . A generative factor  $f$  is rare if there exists a class  $k \in \{1, \dots, C\}$  such that  $|S_{k,0}| \ll |S_{k,1}|$  and for all  $c \neq k$ ,  $|S_{c,0}| \gg |S_{c,1}|$ .

Intuitively, a dataset with a RGF is skewed. In this paper, we take the skewness to the extreme<sup>3</sup> and consider the case where  $|S_{k,0}| = 0$  for a particular class  $k$  and  $|S_{c,1}| = 0$  for all other classes  $c \neq k$ .

Definition 2 characterizes a rare RGF as one whose distribution is highly skewed with respect to the class label. Specifically, the factor  $f$  is considered rare if, for some class  $k \in \{1, \dots, C\}$ , it appears exclusively (or overwhelmingly) in class  $k$ , and is absent (or nearly absent) in all other classes. The sets  $S_{c,0}$  and  $S_{c,1}$  represent the indices of samples in class  $c$  where the generative factor  $f$  takes value 0 or 1, respectively. The condition  $|S_{k,0}| \ll |S_{k,1}|$  and  $|S_{c,0}| \gg |S_{c,1}|$  for all  $c \neq k$  implies that  $f = 1$  is strongly concentrated in class  $k$ . This definition implicitly captures a significant variation in the conditional distribution  $\mathbb{P}(f | y)$  across classes. In the extreme case we study, this variation is taken to its limit:  $f = 1$  occurs only in class  $k$ , and  $f = 0$  in all other classes. This setup enables a controlled analysis of whether generative models generalize the factor  $f$  beyond class  $k$  or memorize its co-occurrence, thereby disentangling generalization from class-conditional memorization.

Although we restrict our current analysis to binary generative factors for interpretability and statistical tractability (i.e., z-tests on proportions), the proposed framework naturally generalizes to continuous or multi-valued factors. In practice, continuous generative factors can be converted into multiple discrete classes by defining thresholds or quantile-based bins (e.g., dividing a continuous attribute such as brightness or texture smoothness into low, medium, and high levels).

Note that we *only* use the data instances  $\mathbf{x}_i$  for the training of generative models. Generative factors  $f_i$  and class labels  $y_i$  serve *exclusively* to evaluate (after training) the model's ability to learn the generative factors. This setting reflects real-world scenarios where explicit labels or factors might not be readily available, challenging the model to capture the generative factors accurately.

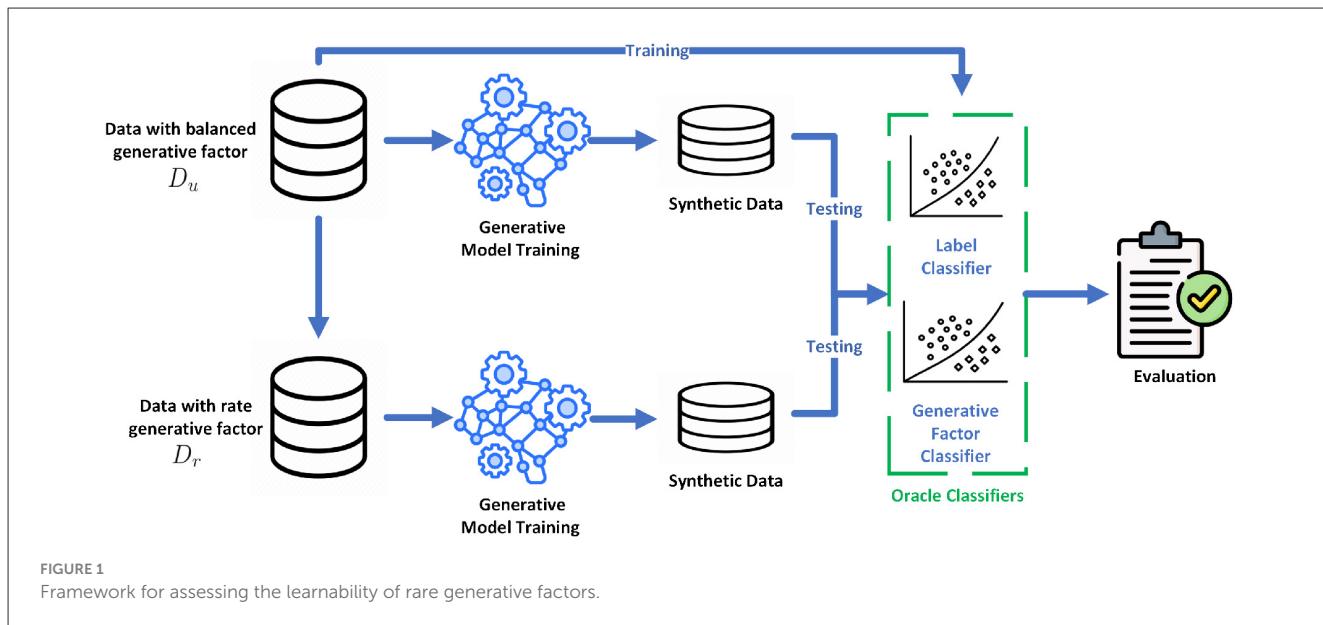
### 3.1 Examples

We now briefly discuss motivating real-world examples of rare generative factors. For each example, we provide a detailed description of the role of  $\mathbf{x}_i$ ,  $f_i$  and  $y_i$ .

**Example 1: Medical imaging for brain health across different ages**

- $\mathbf{x}_i$  - MRI scan of the brain.

3 We relax it in Appendix Section 5.



- $f_i$  - A binary generative factor indicating the age group of the patient, either young (under 60) or old (60+).
- $y_i$  - The health condition identified by the scan, such as normal aging, mild cognitive impairment, or Alzheimer's disease.

In this example, the distribution of age is skewed because Alzheimer's disease mostly affects older people. Consequently, learning to understand the concept of age in relation to Alzheimer's and generating MRI images that accurately depict Alzheimer's in younger individuals, which is still possible with early-onset Alzheimer's (Mendez, 2019), poses a significant challenge. This difficulty arises from the rarity of early-onset Alzheimer's cases in younger populations, making it difficult for models to capture and replicate this condition accurately in generated images.

#### Example 2: Text style in literary genres

- $x_i$  - A passage of text.
- $f_i$  - A binary generative factor indicating the text style, e.g. whether the text includes archaic English words or not (a modern style).
- $y_i$  - The literary genre of the text, such as modern fiction, contemporary poetry, or historical fiction.

In this example, text style might be a rare generative factor, since archaic English is uncommon in modern fiction and contemporary poetry but frequently found in historical fiction. The challenge for generative models is to learn the concept of text style from such skewed data.

#### Example 3: Car images in urban and rural environments

- $x_i$  - Image of a car.
- $f_i$  - The environment in which the car is captured, urban or rural.
- $y_i$  - The brand of the car.

In this example, the rarity of the generative factor arises because luxury car brands, such as BMW, are frequently observed in urban landscapes but are considerably less common in rural environments. This discrepancy presents a challenge in learning the generative factor of the environment effectively.

## 4 Framework for assessing the learnability of RGFs

We now present our framework for studying the learnability of Rare generative factors (RGFs), illustrated in Figure 1.

**Setup:** We start our investigation with a dataset  $D_u = \{(x_i^{(u)}, f_i^{(u)}, y_i^{(u)})\}$  characterized by a *uniform* distribution of the generative factor; that is, within each class, the number of samples with  $f_i = 1$  equals those with  $f_i = 0$ . This balanced dataset serves as a baseline for understanding how generative models perform under standard conditions, where no generative factor is particularly rare.

To understand the impact of an RGF, we construct a new dataset,  $D_r = \{(x_i^{(r)}, f_i^{(r)}, y_i^{(r)})\}$ , derived from the original data instances in  $D_u$ . In this tailored dataset, we introduce a *deliberate skew*: for some selected class  $k$ , all examples have  $f_i = 1$ , which signifies the presence of the RGF. In contrast, for all other classes  $c \neq k$ , all examples have  $f_i = 0$ , indicating the absence of this factor. These two datasets ( $D_u$  and  $D_r$ ) allow us to closely examine how the presence of a rare generative factor influences the learning and generative capabilities of generative models.

To this end, we train two separate generative models (of the same type) for  $\{x_i^{(u)}\}$  and  $\{x_i^{(r)}\}$ , respectively. From each trained model, we then generate  $M$  samples for evaluation. To evaluate these generated samples, we employ two oracle classifiers. These classifiers are trained on the balanced dataset  $D_u$ , serving two functions:

1. **Label classifier:** This classifier is trained using data pairs  $\{(x_i^{(u)}, y_i^{(u)})\}$ , which consist of the data instances and their corresponding class labels. Its role is to categorize the generated

samples into the correct classes, assessing the model's ability to maintain class-specific characteristics in the generated data.

2. **Generative factor classifier:** This binary classifier, trained on  $\{(\mathbf{x}_i^{(u)}, f_i^{(u)})\}$  pairs, focuses on identifying the presence or absence of the generative factor within each sample.

We ensure that both classifiers achieve high accuracy (on a separate test set).

Next, we use the classifiers to determine both the class label and the binary generative factor for each of the  $M$  samples produced by the respective generative model, and then calculate the distribution of the generative factor for each class  $c$ . We denote by  $P_c^{(u)}$  the proportion of instances with  $f = 1$  within class  $c$ , generated by the generative model trained on the uniformly distributed dataset  $D_u$ . Similarly,  $P_c^{(r)}$  represents the proportion of instances with  $f = 1$  from class  $c$ , generated by the generative model that is trained on the skewed dataset  $D_r$ .

## 4.1 Our hypothesis

We hypothesize that for each class  $c$ , the proportion of generated instances by both trained models will be comparable. This hypothesis is grounded in the notion that effective learning by generative models should allow them to extract the generative factors, regardless of their rarity in the training data, with a high degree of fidelity. Essentially, this suggests that the models' ability to discern and generate generative factors is *not* significantly hindered by the skewed distribution of these factors in the training dataset.

## 4.2 Assessing the learning of RGF

We perform a statistical test of the hypothesis to compare the proportions  $P_c^{(u)}$  and  $P_c^{(r)}$ . We employ a one-sample  $z$ -test, which allows us to determine whether the observed differences in proportions between the two groups are statistically significant. We denote by  $z_c$  the  $z$ -score<sup>4</sup> corresponding to class  $c$ ,

$$z_c = \left( P_c^{(r)} - P_c^{(u)} \right) \sqrt{\frac{P_c^{(u)} (1 - P_c^{(u)})}{M}}. \quad (1)$$

To evaluate the capability of generative models to learn RGFs, we calculate the p-value associated with each computed  $z$ -score  $z_c$  for class  $c$ . When p-value  $> 0.05$ , we uphold the null hypothesis, which implies that the model has effectively *learned* the generative factor. This outcome suggests that there is no significant difference between the expected and observed frequencies of the RGF among the generated instances, indicating successful learning by the generative model.

Conversely, a p-value less than 0.05 leads to the rejection of the null hypothesis. Specifically, for the class  $k$  where the rare generative factor has been introduced, and where  $z_k > 0$ , this outcome signifies that the generative model has not learned but rather *memorized* the generative factor for this class. Similarly, if we observe a p-value below 0.05 for a class  $c \neq k$  accompanied by

<sup>4</sup> The  $z$  notation should not be confused with a latent space.

$z_c < 0$ , this also indicates memorization of the generative factor by the generative model for classes other than  $k$ . It is noteworthy to mention that deviations from these specified conditions are rare in practice, underscoring the models' tendency to either learn or memorize generative factors. The subsequent section details the datasets and the specific generative factors employed in our study.

## 4.3 Justification for the chosen test statistic

We employ a one-sample  $z$ -test to compare the class-wise proportions  $P_c^{(u)}$  and  $P_c^{(r)}$  of the generative factor in the synthetic data generated from balanced and skewed training datasets, respectively. The  $z$ -test is appropriate in our setting because we are comparing an observed proportion  $P_c^{(r)}$  against a reference population proportion  $P_c^{(u)}$  obtained from the balanced dataset. Given the sufficiently large number of generated samples ( $M = 1,000$ ), to ensure that the sampling distribution of the proportion approximates a normal distribution, satisfying the assumptions of the test. This test offers an interpretable and computationally efficient means of quantifying deviations from the expected behavior under the null hypothesis that the model has learned the generative factor in a generalizable way. Moreover, the  $z$ -score provides not just significance testing but also directionality (i.e., whether the factor appears more or less frequently than expected), which is critical for distinguishing between learning and memorization. We selected the one-sample  $z$ -test due to its simplicity, suitability for proportion data in large samples.

## 5 Dataset and generative factors

In this work we primarily utilized the colored-MNIST dataset (Arjovsky et al., 2019) and the Morpho-MNIST dataset (Castro et al., 2019), both are stylish versions of the classical greyscale handwritten digits classification MNIST dataset (LeCun et al., 1998). The colored-MNIST dataset enhances the original digit images by incorporating a color scheme of green and red. The Morpho-MNIST dataset modifies the digits with morphological modifications, such as variations in thickness, swelling, and the introduction of fractures. To extend our analysis beyond handwritten digits, we also employed a subset of the Comprehensive Cars (CompCars) Surveillance dataset (Yang et al., 2015). From this dataset, we selected images of two car makes (Volkswagen and Toyota) in two colors (black and white), allowing us to explore our hypotheses in a different domain. Supplementary Table S2 details the sample distribution of our CompCars subset. While the datasets employed (colored-MNIST, Morpho-MNIST, and CompCars) are relatively simple, their controlled structure is deliberate. They allow us to manipulate factor rarity in a quantifiable and interpretable way, ensuring that observed effects are due to data imbalance rather than uncontrolled complexity. Such canonical datasets have historically provided valuable insight into generalization behavior before scaling to complex, real-world scenarios.

We designed our VAE, GAN, and DM to work with RGB (3 channels) images. Consequently, to accommodate the greyscale

images from the Morpho-MNIST dataset, we transformed them into color images. This is achieved by randomly assigning either a red or a green color to each image, ensuring an equal probability distribution between the two colors for the images with morphological modifications.

As detailed in Section 4, for each generative factor under consideration we created two datasets:

1. A balanced dataset  $D_u$ , where the generative factor is uniformly distributed across all classes. For MNIST-based experiments, this dataset comprises 60,000 images with an equal representation of each digit. In the case of the CompCars subset, we utilized 1,448 images, ensuring an even distribution between Volkswagen and Toyota cars.
2. A dataset  $D_r$  with rare generative factor. For MNIST-derived datasets, we introduce the rare generative factor to a single digit class. We specifically chose digits “1” and “2” as representative cases, conducting separate experiments where the rare factor is exclusively associated with each of these digits. This approach allows us to examine how the shape of the digit might influence the model’s ability to learn or memorize the rare factor. For the CompCars subset, we assign the rare generative factor to car make.

We trained VAE, GAN, and DM separately on each dataset. The full training details and model architectures are described in [Supplementary Section 1](#).

After training the models for each generative factor, we generated  $M = 1,000$  synthetic images. The oracle classifiers are used to detect the class (digit for MNIST, car make for CompCars) and the presence of the generative factor in the synthetic images.

## 5.1 Generative factors

Variations in color and morphology are naturally used in our work as generating factors, as they are important in determining the visual appearance of the digits. Specifically, we defined the following 5 generative factors for digits: Color, Fracture, Thinning, Thickening, and Swelling. Note that *only* one generative factor is introduced at a time. [Supplementary Figure S2](#) demonstrates the case of **rare** generative factors where digit “1” is selected as the class in which the generative factor is introduced (for example, for the Thickening factor all images of digit “1” are thick while other digits retain a standard thickness). For the color factor, the presence of green is designated as the rare generative factor. For CompCars, color is the generative factor, where all Volkswagen cars are white and Toyota cars are black.

For digits, the generative factors are introduced in the images using the Morpho-MNIST python library.<sup>5</sup> For Thinning and Thickening the value of the *amount* parameters is 0.7 and 1, respectively. For Swelling the value of the *strength* parameter is 3 and the *radius* is 7. For Fracture the value of *num\_frac* is 3. For cars, the generative factor is introduced by selecting the corresponding subset of the CompCars dataset.

<sup>5</sup> <https://github.com/dccastro/Morpho-MNIST>

## 5.2 Oracle classifiers

As mentioned in Section 4, we rely on oracle classifiers to categorize images generated by VAEs, GANs, and DMs. We employed Convolutional Neural Networks (CNN) as our oracle classifiers. The details of the architectures appear in [Supplementary Section 1](#). For each generative factor we trained two oracle classifiers on the balanced dataset. For the MNIST-derived datasets, we trained one classifier for digit classification and another for factor classification, resulting in a total of 10 classifiers. Some images from the dataset used to train the digit classifier (10-class problem) and color classifier (2-class problem) appear in [Supplementary Figure S2](#). For cars, we trained one classifier for car make classification and another for color classification, using the data shown in [Supplementary Table S2](#).

The MNIST oracle classifiers are trained using SGD for 8 epochs employing the cross entropy loss, batch size of 64, learning rate of 0.01, and momentum of 0.5. For car make classification, we used 100 epochs. To evaluate the performance of these classifiers, we used a test-set of 20,000 samples for digits and 185 samples for cars. The classification accuracies, as detailed in [Supplementary Table S1](#), show that all classifiers achieved a test-set accuracy exceeding 92%, underscoring their high efficacy in accurately identifying both digits, car make and generative factors.

We further validated the robustness of these oracle classifiers by evaluating their performance on generated images. Although detection accuracy decreased slightly (by approximately 5%) compared to the test set, it remained consistently high across all generative models. This indicates that the classifiers retained reliable discrimination capability even when applied to synthetic data. Therefore, the observed memorization-vs.-learning effects are unlikely to be artifacts of classifier noise and instead reflect genuine generative behavior.

## 5.3 Statistical testing

For each class  $c$ , we compare the observed proportion  $P_c^{(r)}$  (model trained on  $D_r$ ) to the reference proportion  $P_c^{(u)}$  (model trained on  $D_u$ ) using a two-sided one-sample  $z$ -test with  $M = 1,000$  generated samples per model/factor/dataset.

## 5.4 Frame work pipeline

We have summarized the pipeline as follow:

1. Start from a balanced dataset  $D_u$  where, for each class  $y=c$ , the factor  $f$  is uniform (50/50).
2. Create a skewed dataset  $D_r$  by selecting a target class  $k$  and setting  $f=1$  for all  $y=k$  and  $f=0$  for all  $y \neq k$ .
3. Train identical model architectures separately on  $\{x_i^{(u)}\}$  and  $\{x_i^{(r)}\}$  without access to  $f$  or  $y$ .
4. Generate  $M=1,000$  samples per trained model.
5. Use high-accuracy oracle classifiers (trained only on  $D_u$ ) to label  $y$  and detect  $f$  in generated samples.
6. Compute class-wise proportions  $P_c^{(u)}$  and  $P_c^{(r)}$ , then test for learning vs. memorization as in [Equation 1](#).

TABLE 1 Z-scores for all models (VAE, GAN, GAN-SD, DM) where all images of digit “1” have the rare generative factor (RGF).

Digit	VAE	GAN	GAN-SD	DM
<b>(a) Color</b>				
0	–	–	–	–
1	–6.14	17.05	–5.49	14.77
2	–	–40.92	–2.34	–
3	–24.94	–37.48	–82.23	–
4	–	–	–	–
5	–	–	–	–
6	–	–34.87	–4.93	–
7	–	–40.20	–7.77	–
8	–10.29	–65.37	–2.97	–
9	–	–11.09	–6.50	–
<b>Total</b>	–75.30	–39.18	–44.87	–42.67
<b>(b) Fracture</b>				
0	<b>–1.80</b>	<b>0.01</b>	<b>1.36</b>	–28.56
1	3.92	<b>–0.97</b>	<b>1.44</b>	32.75
2	<b>–1.71</b>	–15.49	–2.34	–4.42
3	–2.30	–10.30	–5.54	–14.90
4	<b>0.03</b>	–15.20	–5.08	–37.92
5	<b>0.59</b>	–4.26	–2.48	–11.92
6	<b>–1.65</b>	–34.87	–4.93	–16.97
7	<b>–0.79</b>	–16.46	–7.77	–14.11
8	–2.25	<b>–0.87</b>	–2.97	–14.22
9	–5.48	–11.09	–6.50	–14.44
<b>Total</b>	–2.21	–21.28	–9.57	–18.87
<b>(c) Swell</b>				
0	–5.28	–4.77	<b>0.89</b>	–6.51
1	<b>–0.94</b>	4.23	–5.68	9.57
2	–8.48	–4.87	<b>–0.29</b>	–9.43
3	–2.19	–5.89	–11.27	–6.85
4	–7.23	–14.59	–9.91	–7.60
5	–3.55	–9.86	–16.00	–9.21
6	–3.07	–13.66	–8.55	–5.63
7	–10.78	–7.93	–9.53	–13.31
8	–5.66	–8.03	<b>–0.64</b>	–7.26
9	–8.57	–12.33	–3.48	–7.04
<b>Total</b>	–14.60	–21.13	–15.49	–17.49
<b>(d) Thickening</b>				
0	–4.66	–9.09	–	–9.28
1	2.39	2.96	–4.97	26.33
2	–7.11	–4.36	–7.89	–12.10
3	–12.21	–8.36	–4.25	–14.93

(Continued)

TABLE 1 (Continued)

Digit	VAE	GAN	GAN-SD	DM
4	–5.97	–56.40	–16.55	–8.45
5	–22.98	–19.24	–15.89	–20.45
6	–12.03	–42.80	–14.31	–14.76
7	–2.38	–8.80	–6.09	–22.88
8	<b>–1.34</b>	–14.22	–3.38	–23.75
9	<b>–1.62</b>	–23.56	–11.47	–15.23
<b>Total</b>	–14.01	–33.41	–24.97	–20.64
<b>(e) Thinning</b>				
0	–3.70	–8.65	<b>0.82</b>	–40.68
1	7.15	22.90	<b>0.16</b>	14.54
2	–2.36	–5.82	–7.17	–16.81
3	–3.62	–19.58	–7.74	–50.81
4	<b>–1.23</b>	–8.71	–4.45	–15.66
5	–3.60	–12.31	–4.39	–12.13
6	–5.57	–11.97	–11.03	–66.40
7	<b>–0.78</b>	–7.90	<b>0.47</b>	–7.90
8	–5.59	–11.85	–11.35	–13.32
9	<b>–1.25</b>	–11.60	–7.83	–6.49
<b>Total</b>	–7.86	–21.09	–13.27	–35.08

Bold values ( $p > 0.05$ ) indicate RGF learning. –, not estimated or not applicable.

*Interpretation.* For each digit class  $c$ ,  $z$  compares the proportion of  $f=1$  from the model trained on  $D_r$  (skewed) against the reference proportion from  $D_u$  (balanced). Bold denotes  $p > 0.05$  (no significant difference), interpreted as learning rather than memorization.

*Legend.* VAE, Variational Autoencoder; GAN, Generative Adversarial Network; GAN-SD, GAN with Spectral Decoupling; DM, Diffusion Model.

–, Not estimated or not applicable for that cell (e.g., insufficient support).

## 6 Results and discussion

Utilizing the framework of Section 4 and the datasets (Section 5), we now present our findings. Due to space constraints, we have placed the majority of tables and figures in the [Appendix](#).

Initially, we used the balanced datasets  $D_u$  for each Rare generative factor (RGF), trained the models, and then generated  $M = 1,000$  synthetic images. As expected,  $P_c^{(u)}$  approximates 0.5 in the majority of cases, indicating a balanced representation of the generative factors within the synthetic images (for details see [Supplementary Tables S3, S4](#))

Subsequently, for each RGF, we trained the models using the skewed dataset  $D_r$  and determined the proportions  $P_c^{(r)}$  for each digit (for MNIST dataset) and car (for CompCars dataset). We then used [Equation 1](#) to calculate the z-scores and report the results in [Tables 1–3](#).

We then used [Equation 1](#) to calculate the class-wise z-scores, which quantify the difference between the proportions of samples containing the rare generative factor ( $f = 1$ ) generated by models trained on the skewed dataset  $D_r$  and the balanced dataset  $D_u$ . [Tables 1–3](#) summarize these results for all datasets and model types.

TABLE 2 Z-scores for all models (VAE, GAN, GAN-SD, DM) where all images of digit “2” have the rare generative factor (RGF).

Digit	VAE	GAN	GAN-SD	DM
<b>(a) Color</b>				
0	−20.84	−	−78.05	−
1	−23.41	−12.64	−22.99	−89.10
2	17.24	13.64	3.12	42.09
3	−26.85	−25.03	−30.88	−
4	−43.88	−	−	−
5	−	−	−	−
6	−	−49.63	−16.42	−
7	−17.70	−35.28	−	−70.75
8	−55.44	−45.78	−8.21	−69.90
9	−	−	−	−
<b>Total</b>	−39.94	−37.66	−42.60	−47.28
<b>(b) Fracture</b>				
0	<b>−1.16</b>	<b>0.59</b>	2.73	−22.58
1	<b>−0.38</b>	−42.82	−38.40	−26.36
2	<b>1.88</b>	<b>0.42</b>	−2.38	<b>1.83</b>
3	−4.10	<b>−0.65</b>	−2.84	−6.92
4	<b>−0.27</b>	−29.01	−6.32	−9.89
5	−4.36	<b>−0.07</b>	−4.39	−3.67
6	<b>−0.76</b>	−19.33	−16.32	−10.92
7	−2.25	−16.87	−7.84	−7.93
8	<b>−0.30</b>	−2.86	−2.12	−7.11
9	−3.49	−23.71	−11.12	−9.14
<b>Total</b>	−4.27	−21.74	−19.12	−20.83
<b>(c) Swell</b>				
0	<b>−1.63</b>	−4.54	3.36	−24.44
1	−8.76	−11.04	−14.67	−14.38
2	3.27	<b>−0.40</b>	<b>−1.17</b>	8.23
3	−4.10	−4.31	−11.34	−6.57
4	−6.16	−2.21	−10.69	−7.06
5	−2.00	−4.89	−11.96	−21.07
6	−2.17	−6.03	−5.50	−7.05
7	−17.03	−4.31	−5.56	−10.39
8	−7.87	−8.50	−4.17	−7.70
9	−7.85	−7.26	−10.43	−8.23
<b>Total</b>	−14.81	−15.71	−19.33	−20.83
<b>(d) Thickening</b>				
0	−9.86	−8.20	−4.92	−21.55
1	−7.24	−28.73	−45.43	−15.81
2	6.16	9.99	<b>0.06</b>	11.74
3	−13.58	−15.00	−10.94	−36.83

(Continued)

TABLE 2 (Continued)

Digit	VAE	GAN	GAN-SD	DM
4	−5.12	−	−62.51	−8.78
5	−22.69	−43.46	−22.24	−
6	−9.70	−27.05	−21.60	−11.03
7	−7.93	−12.31	−22.25	−13.44
8	<b>−1.91</b>	<b>−1.93</b>	−9.05	−22.24
9	−2.80	−32.17	−29.76	−10.42
<b>Total</b>	−17.05	−34.28	−43.87	−23.01
<b>(e) Thinning</b>				
0	−4.11	−10.53	<b>1.74</b>	−42.08
1	−6.84	−14.10	−5.68	−20.25
2	5.04	7.25	−3.14	15.93
3	−2.26	−32.70	−11.01	−18.24
4	−3.65	−12.04	−12.73	−10.14
5	−2.87	−16.09	−9.24	−12.53
6	−5.34	−17.38	−6.17	−30.48
7	<b>−1.28</b>	−17.33	<b>0.17</b>	−20.80
8	−5.03	−17.35	−6.72	−18.66
9	−4.98	−14.81	−5.90	−10.60
<b>Total</b>	−9.95	−35.36	−15.23	−32.41

Bold values ( $p > 0.05$ ) indicate RGF learning. NA/−, not estimated or not applicable.

*Interpretation.* For each digit class  $c$ ,  $z$  compares the proportion of  $f=1$  from the model trained on  $D_r$  (skewed) against the reference proportion from  $D_u$  (balanced). Bold denotes  $p > 0.05$  (no significant difference), interpreted as learning rather than memorization.

*Legend.* VAE, Variational Autoencoder; GAN, Generative Adversarial Network; GAN-SD, GAN with Spectral Decoupling; DM, Diffusion Model.

“−”. Not estimated or not applicable for that cell (e.g., insufficient support).

- Table 1 presents  $z$ -scores for all models (VAE, GAN, GAN-SD, and Diffusion Model) when the rare generative factor (RGF) is introduced in digit “1.” Each row corresponds to a digit class, and each column to a generative model. Bold values indicate non-significant differences ( $p > 0.05$ ), interpreted as *learning* rather than memorization.
- Table 2 reports the analogous results when the RGF is introduced in digit “2.”
- Table 3 shows the  $z$ -scores for the CompCars dataset, where the RGF corresponds to color (white Volkswagen vs. black Toyota). Here again, positive  $z$ -scores denote over-representation of the rare factor, indicating memorization, whereas values near zero imply generalization.

Together, these tables provide a quantitative summary of the models’ ability to learn or memorize rare generative factors across both digit- and object-based datasets.

While standard machine learning theory predicts that underrepresented features are difficult to learn due to their low empirical frequency, our focus on rare generative factors (RGFs) highlights an important distinction: RGFs actively shape the data generation process, meaning their absence or distortion can affect not only classification but also the model’s ability to

TABLE 3 CompCars, z-scores for all models (VAE, GAN without SD, GAN with SD, DM), with color RGF: white Volkswagen, black Toyota.

Make	VAE			GAN			GAN-SD			Diffusion models		
	Black	White	z	Black	White	z	Black	White	z	Black	White	z
Volkswagen	161	397	13.11	153	425	12.28	132	350	10.64	153	204	9.98
Toyota	336	106	-11.33	334	88	-8.16	454	64	-17.05	605	38	-19.45
Total	497	503	2.09	487	513	4.62	586	414	-1.67	758	242	-2.07

Bold: similar proportions ( $p > 0.05$ ), indicating RGF learning.

generate coherent, semantically consistent samples. This makes their impact more profound than that of merely rare labels or attributes. Theoretically, RGFs define a low-probability region in the generative manifold, and standard likelihood-based training may fail to sufficiently penalize errors in such regions. This underlines the need for tailored inductive biases or priors that preserve generative completeness, especially in applications where coverage of rare modes is critical.

## 6.1 Memorization of RGF

Comparing the proportions  $P_c^{(u)}$  and  $P_c^{(r)}$  via the z-scores in Tables 1–3 underscores the propensity of generative models to memorize RGFs. For instance, GAN exhibits a notable bias toward associating the green color with digits “1” and “2,” in contrast to the red color, which is more frequently linked with the remaining digits. Specifically, when the green color is assigned to digit “1,” an overwhelming 87% of generated images display this characteristic, a stark contrast to the 35% for the balanced data. Conversely, the presence of green in images of other digits is minimal, hovering around 1%, indicating a clear memorization of the green color for digit “1” without extending this rare factor to other digits. A similar trend is evident when the color factor is applied to digit “2” (see Supplementary Section 4 for detailed results).

### 6.1.1 Theoretical perspective on RGF memorization

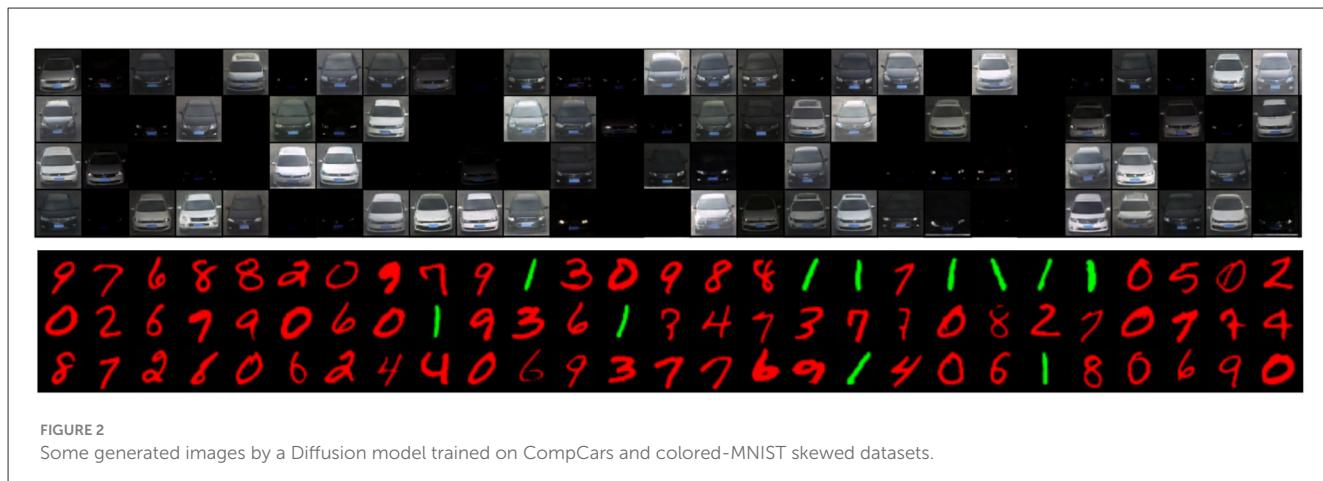
From a theoretical standpoint, rare generative factors (RGFs) occupy low-probability regions in the data manifold. In likelihood-based training objectives such as the VAE’s evidence lower bound or the diffusion model’s denoising score matching, gradients are dominated by high-density regions, while low-density (rare) regions receive vanishing gradient updates. This imbalance implicitly biases the model toward reconstructing frequent factors and memorizing rare ones, rather than forming disentangled, transferable representations. Such behavior aligns with the “gradient starvation” phenomenon (Pezeshki et al., 2021), where dominant correlations absorb most gradient flow, leaving underrepresented modes under-trained. From an information-theoretic perspective, RGFs correspond to directions of high Fisher Information curvature but limited support, making them unstable under empirical risk minimization. Therefore, memorization emerges naturally as an energy-efficient solution

in overparameterized networks that minimize loss without guaranteeing uniform coverage of the generative manifold.

The large z-scores highlight the significant differences in proportions between  $P_c^{(u)}$  and  $P_c^{(r)}$ , confirming the memorization effect. This memorization phenomenon is *not* limited to color in digit datasets. It extends, yet to varying degrees, across other generative factors we studied. In the case of car images, we observe a similar trend where the models tend to strongly associate color with a car make. The observed pattern suggests a broader trend: *GANs and DMs exhibit a stronger tendency toward memorization of RGFs compared to VAEs*, both in digit recognition and car classification tasks. Visual inspection suggests that DM provides the highest image quality, as shown in Figure 2, but at the cost of increased memorization (the images generated using VAE and GAN are shown in Supplementary Section 4). This different behavior across model types and datasets highlights the nuanced ways in which various generative architectures approach the challenge of learning from skewed data distributions.

### 6.1.2 Distinguishing memorization from semantic correlation

A crucial distinction in our study is between (a) *memorization* of rare generative factors (RGFs) and (b) genuine *learning* of generative factors that are strongly correlated with semantic class features. Memorization, in our context, refers to the model reproducing RGFs only within the class where they were seen during training (e.g., generating green digits exclusively for class “1” if green was only present in that class). This indicates that the model has not abstracted the RGF as a transferable concept, but instead has tightly coupled it with the class identity. In contrast, learning is evidenced when the model applies the RGF to other classes not seen with that factor in training, thereby indicating that it has captured the generative factor independently of class label. To empirically distinguish the two, we rely on the distributional comparison between  $P_c^{(r)}$  and  $P_c^{(u)}$  using the z-test, where  $P_c^{(u)}$  acts as a reference distribution under balanced conditions. A non-significant z-score ( $p > 0.05$ ) suggests that the model has generalized the RGF across classes, whereas a significant z-score in the direction of the skew (i.e., high  $P_c^{(r)}$  for the class with the RGF and low for all others) indicates memorization. Thus, semantic correlation alone is not sufficient to explain this behavior unless it holds under balanced data, in which case  $P_c^{(u)}$  would already show asymmetry. Our framework explicitly controls for such effects by comparing against the balanced baseline.



## 6.2 How RGF memorization originates in GANs?

We are interested in understanding how memorization of RGFs happens. We picked GANs for two main reasons: first, because they exhibited a stronger tendency to memorize RGFs in our experiments compared to VAEs, and second, because their architecture includes a discriminator that allows us to explore the role of adversarial training in potentially encouraging this memorization behavior. Indeed, we analyzed the discriminator loss during GAN training with respect to the “real label” using a separate balanced validation set of 2,000 images of digits and 185 images of cars.

To do this, we computed the loss only for images where RGFs are applied (“1” and “2” for MNIST and Volkswagen for CompCars). We differentiate between images featuring RGFs and those without.

Figure 3 illustrates the discriminator loss for the color factor in MNIST data, with RGF present in digit “1” (Supplementary Section 4) presents results for other RGFs and digits). In this plot, solid lines depict the loss associated with images containing RGFs (i.e., green images), while dashed lines indicate the loss for images lacking RGFs (i.e., red images). A green horizontal dashed line represents the threshold loss at the discriminator’s decision boundary between identifying images as real or fake, corresponding to a loss of  $\log(2)$  when the discriminator output logit is 0.

When training the GAN with the balanced dataset  $D_u$ , there appears to be no significant discrepancy between the loss for images with RGF and those without, suggesting that the discriminator does not differentiate based on the presence of RGF. In other words, the discriminator is invariant to RGF. However, training on the skewed dataset  $D_r$ , we observe a gap between the losses for images with and without RGF. This indicates that despite all images being “real,” the discriminator classifies images with and without RGFs differently, losing its invariance to RGFs. This differentiation likely stems from the spurious correlation between the digit and the RGF, reminiscent of the “gradient starvation” phenomenon identified by Pezeshki et al. (2021) in the context of discriminative learning. This phenomenon, where the model excessively focuses on dominant features at the expense of others, may explain the

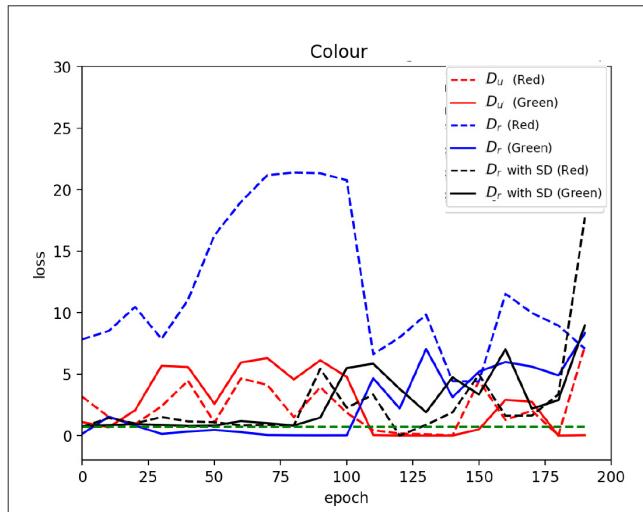


FIGURE 3  
Discriminator loss by presence of the rare factor (color). Solid lines represent batches containing the RGF (green), while dashed lines correspond to batches without the RGF (red). The horizontal dashed line indicates the discriminator’s decision boundary [ $\log(2)$ ].

discriminator’s skewed learning, underlining the complexity of addressing memorization of RGFs in GANs.

## 6.3 Mitigating memorization in GANs by spectral decoupling

Our next focus is to evaluate if the Spectral Decoupling (SD) technique, previously proposed by Pezeshki et al. (2021) to address the issue of gradient starvation, can also help in reducing the memorization of RGFs by GANs.

In the context of discriminative learning, SD augments the loss function with a regularization term  $\frac{\lambda}{2} \|\hat{y}\|^2$ , where  $\lambda$  is a regularization strength hyperparameter, and  $\hat{y}$  is the logits vector output by the model for a given input batch. This regulariser aims to restrain the magnitudes of logits, thereby preventing any single (and potentially spurious) feature from overpowering the model’s output.

We incorporated this regularization method into the GAN training process for the initial 80 epochs by adding the SD regulariser to the discriminator's loss computation for real image batches, with  $\lambda = 0.8$  (Supplementary Figures S19, S20 presents results for different  $\lambda$  values). After 80 epochs we removed the regulariser for further training until 200 epochs, allowing the GAN image quality to improve. We use SD as a regulariser: strong enough to mitigate early gradient starvation, but removed later to allow the generator to recover full visual quality.

The effect of SD is evident in Figure 3, where the discriminator loss dynamics (illustrated by solid and dashed black lines) converge more closely during the SD application phase (up to epoch 80), suggesting increased discriminator invariance to RGF and thus mitigating the memorization problem. In addition, Tables 1, 2 demonstrate that applying SD generally results in smaller z-scores, suggesting reduced memorization.

Finally, in Table 4 we used the p-values corresponding to the z-scores in Tables 1, 2 (for MNIST data) to deduce whether the RGF is learned (L) or memorized (M). Note that all DM values are M, indicating a strong tendency of diffusion models to memorize RGFs. We observe that SD helps in mitigating memorization to some extent for GAN. For CompCars data, GAN with SD achieved learning in one case only (Table 3). We report results using two additional random seeds in the Supplementary material, further validating these findings.

Beyond Spectral Decoupling, several complementary strategies may further mitigate RGF memorization. Causal disentanglement frameworks Zhang et al. (2024) could encourage the model to separate causal generative mechanisms from observational correlations. Similarly,  $\beta$  – VAEs and InfoGANs introduce inductive biases that promote factorized latent spaces, which may improve transfer of rare factors. Contrastive or self-supervised regularisers (e.g., VICReg, SimCLR) could also enhance invariance by encouraging feature alignment across samples differing only in rare factors. Investigating these directions remains an important next step for improving generalization under imbalance.

## 7 Conclusion

We are interested in examining how generative models like VAEs, GANs, and DMs learn rare generative factors (without explicit supervision). Through a systematic empirical study involving several generative factors and two datasets, we showed that generative models exhibit a propensity toward memorizing rare generative factors. We demonstrated that regularization techniques such as spectral decoupling can mitigate this memorization tendency to a certain degree.

Our current study deliberately focused on GAN, VAE, and Diffusion architectures to isolate fundamental learning dynamics without confounding factors introduced by architectural enhancements. Future work will extend this analysis to state-of-the-art variants such as StyleGAN2, VQ-VAE, and Latent Diffusion Models to assess whether large-scale pretraining or richer priors alleviate the memorization of RGFs. We also plan to scale to multimodal datasets and explore causal or contrastive regularization strategies. Such extensions will clarify whether memorization of rare generative factors is a fundamental

TABLE 4 Summary of RGF learning (L) vs. memorization (M) for digits "1" and "2."

Digit	Color	Frac	Swell	Thick	Thin
<b>(a) RGF in digit "1"</b>					
0	M/M/M/M	L/L/L/M	M/M/L/M	M/M/M/M	M/M/L/M
1	M/M/M/M	M/L/L/M	L/M/M/M	M/M/M/M	M/M/L/M
2	M/M/M/M	L/M/M/M	M/M/L/M	M/M/M/M	M/M/M/M
3	M/M/M/M	M/M/M/M	M/M/M/M	M/M/M/M	M/M/M/M
4	M/M/M/M	L/M/M/M	M/M/M/M	M/M/M/M	L/M/M/M
5	M/M/M/M	L/M/M/M	M/M/M/M	M/M/M/M	M/M/M/M
6	M/M/M/M	L/M/M/M	M/M/M/M	M/M/M/M	M/M/M/M
7	M/M/M/M	L/M/M/M	M/M/M/M	M/M/M/M	L/M/L/M
8	M/M/M/M	M/L/M/M	M/M/L/M	L/M/M/M	M/M/M/M
9	M/M/M/M	M/M/M/M	M/M/M/M	L/M/M/M	L/M/M/M
all	M/M/M/M	M/M/M/M	M/M/M/M	M/M/M/M	M/M/M/M
Count	0/0/0/0	6/3/2/0	1/0/3/0	2/0/0/0	3/0/3/0
<b>(b) RGF in digit "2"</b>					
0	M/M/M/M	L/L/M/M	L/M/M/M	M/M/M/M	M/M/L/M
1	M/M/M/M	L/M/M/M	M/M/M/M	M/M/M/M	M/M/M/M
2	M/M/M/M	L/L/M/L	M/L/L/M	M/M/L/M	M/M/M/M
3	M/M/M/M	M/L/M/M	M/M/M/M	M/M/M/M	M/M/M/M
4	M/M/M/M	L/M/M/M	M/M/M/M	M/M/M/M	M/M/M/M
5	M/M/M/M	M/L/M/M	M/M/M/M	M/M/M/M	M/M/M/M
6	M/M/M/M	L/M/M/M	M/M/M/M	M/M/M/M	M/M/M/M
7	M/M/M/M	M/M/M/M	M/M/M/M	M/M/M/M	M/M/L/M
8	M/M/M/M	L/M/M/M	M/M/M/M	M/L/M/M	L/M/M/M
9	M/M/M/M	M/M/M/M	M/M/M/M	L/M/M/M	M/M/M/M
all	M/M/M/M	M/M/M/M	M/M/M/M	M/M/M/M	M/M/M/M
Count	0/0/0/0	6/4/0/1	1/1/1/0	1/1/1/0	1/0/2/0

Notation: VAE/GAN/GAN-SD/DM. Bold values indicate learning ( $p > 0.05$ ). A total of 43 cases were learned out of 440.

limitation of generative learning or a challenge mitigable through architectural and theoretical innovations.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

FH: Writing – review & editing, Conceptualization, Methodology, Software, Investigation, Writing – original draft, Formal analysis, Data curation. EM: Writing – review & editing, Conceptualization, Investigation, Methodology, Writing – original draft. YX: Software, Writing – review & editing. ST: Writing

– review & editing, Conceptualization, Funding acquisition, Methodology, Writing – original draft, Supervision, Resources, Project administration.

## Funding

The author(s) declare that financial support was received for the research and/or publication of this article. We would like to acknowledge UK's Engineering and Physical Sciences Research Council (EPSRC) (grant EP/X017680/1). We also would like to acknowledge the UKRI AI programme and ESPRC, for CHAI-Causality in Healthcare AI Hub (grant number EP/Y028856/1).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Correction note

This article has been corrected with minor changes. These changes do not impact the scientific content of the article.

## References

Akbar, M. U., Wang, W., and Eklund, A. (2023). Beware of diffusion models for synthesizing medical images-a comparison with gans in terms of memorizing brain tumor images. *arXiv preprint arXiv:2305.07644*. doi: 10.2139/ssrn.4611613

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

Bond-Taylor, S., Leach, A., Long, Y., and Willcocks, C. G. (2022). Deep generative modelling: a comparative review of vae's, gans, normalizing flows, energy-based and autoregressive models. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 7327–7347. doi: 10.1109/TPAMI.2021.3116668

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., et al. (2023). Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*.

Castro, D. C., Tan, J., Kainz, B., Konukoglu, E., and Glocker, B. (2019). Morpho-MNIST: quantitative assessment and diagnostics for representation learning. *J. Mach. Learn. Res.* 20, 1–29. doi: 10.3929/ethx-b-000391770

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., et al. (2023). Chatgpt and the rise of large language models: the new AI-driven infodemic threat in public health. *Front. Public Health* 11:1166120. doi: 10.3389/fpubh.2023.1166120

de Wynter, A., Wang, X., Sokolov, A., Gu, Q., and Chen, S.-Q. (2023). An evaluation on large language model outputs: discourse and memorization. *arXiv preprint arXiv:2304.08637*. doi: 10.1016/j.jnp.2023.100024

Fard, A. P., Mahoor, M. H., Lamer, S. A., and Sweeny, T. (2023). Ganalyzer: analysis and manipulation of gans latent space for controllable face synthesis. *arXiv preprint arXiv:2302.00908*.

Garrido, S., Borysov, S. S., Pereira, F. C., and Rich, J. (2020). Prediction of rare feature combinations in population synthesis: application of deep generative modelling. *Transport. Res. Part C* 120:102787. doi: 10.1016/j.trc.2020.102787

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, eds. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Curran Associates, Inc.).

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., et al. (2018). Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*.

Ho, J., Jain, A., and Abbeel, P. (2020). “Denoising diffusion probabilistic models,” in *Advances in Neural Information Processing Systems*, 6840–6851.

Iyer, S., Lin, X. V., Pasunuru, R., Mihaylov, T., Simig, D., Yu, P., et al. (2022). OPT-IML: scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Jegorova, M., Kaul, C., Mayor, C., O’Neil, A. Q., Weir, A., Murray-Smith, R., et al. (2023). Survey: leakage and privacy at inference time. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 9090–9108. doi: 10.1109/TPAMI.2022.3229593

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). “Analyzing and improving the image quality of stylegan,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8107–8116. doi: 10.1109/CVPR42600.2020.00813

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Liu, X., Sanchez, P., Thermos, S., O’Neil, A. Q., and Tsaftaris, S. A. (2022). Learning disentangled representations in the imaging domain. *Med. Image Anal.* 80:102516. doi: 10.1016/j.media.2022.102516

Liu, X., Yuan, J., An, B., Xu, Y., Yang, Y., and Huang, F. (2023). “C-disentanglement: discovering causally-independent generative factors under an inductive bias of confounder,” in *ICML 2023 Workshop on Structured Probabilistic Inference Generative Modeling*.

Maini, P., Mozer, M. C., Sedghi, H., Lipton, Z. C., Kolter, J. Z., and Zhang, C. (2023). Can neural network memorization be localized? *arXiv preprint arXiv:2307.09542*.

Mendez, M. F. (2019). Early-onset alzheimer disease and its variants. *Continuum* 25, 34–51. doi: 10.1212/CON.000000000000687

Mercatalli, G., and Freitas, A. (2021). “Disentangling generative factors in natural language with discrete variational autoencoders,” in *Conference on Empirical Methods in Natural Language Processing*. doi: 10.18653/v1/2021.findings-emnlp.301

Mohamadi, S., Mujtaba, G., Le, N., Doretto, G., and Adjero, D. A. (2023). Chatgpt in the age of generative AI and large language models: a concise survey. *arXiv preprint arXiv:2307.04251*.

## Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2025.1697139/full#supplementary-material>

Peltenburg, P., Hoedemaekers, Y., Clur, S.-A., Blom, N., Blank, A., Boesaard, E., et al. (2022). Screening, diagnosis and follow-up of brugada syndrome in children: a dutch expert consensus statement. *Neth. Heart J.* 31, 133–137. doi: 10.1007/s12471-022-01723-6

Pezeshki, M., Kaba, O., Bengio, Y., Courville, A. C., Precup, D., and Lajoie, G. (2021). “Gradient starvation: a learning proclivity in neural networks,” in *Advances in Neural Information Processing Systems*, 1256–1272.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Shao, H., Kumar, A., and Thomas Fletcher, P. (2018). “The riemannian geometry of deep generative models,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 315–323. doi: 10.1109/CVPRW.2018.00071

Speranzon, A., Chicco, D., Bonazza, P., D’Alfonso, R., Bobbo, M., D’Agata Mottolese, B., et al. (2024). Brugada syndrome: focus for the general pediatrician. *Children* 11:281. doi: 10.3390/children11030281

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., et al. (2022). Galactica: a large language model for science. *arXiv preprint arXiv:2211.09085*.

Team OpenAI (2022). *Chatgpt: Optimizing Language Models for Dialogue*. OpenAI.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). Llama: open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Träuble, F., Creager, E., Kilbertus, N., Locatello, F., Dittadi, A., Goyal, A., et al. (2021). “On disentangled representations learned from correlated data,” in *Proceedings of the 38th International Conference on Machine Learning*, eds. M. Meila, and T. Zhang (PMLR), 10401–10412.

Yang, G., Fei, N., Ding, M., Liu, G., Lu, Z., and Xiang, T. (2021). “L2m-gan: learning to manipulate latent space semantics for facial attribute editing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2951–2960. doi: 10.1109/CVPR46437.2021.00297

Yang, L., Luo, P., Change Loy, C., and Tang, X. (2015). “A large-scale car dataset for fine-grained categorization and verification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi: 10.1109/CVPR.2015.72 99023

Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., et al. (2022). “Styleswin: Transformer-based gan for high-resolution image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11304–11314. doi: 10.1109/CVPR52688.2022.01102

Zhang, J., Greenwald, K., Squires, C., Srivastava, A., Shanmugam, K., and Uhler, C. (2024). “Identifiability guarantees for causal disentanglement from soft interventions,” in *Advances in Neural Information Processing Systems*, 36.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., et al. (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.