

# Diffusion Models for Counterfactual Explanations

Guillaume Jeanneret, Loïc Simon and Frédéric Jurie

Normandy University, ENSICAEN, UNICAEN, CNRS, GREYC, France  
guillaume.jeanneret-sanmiguel@unicaen.fr

**Abstract.** Counterfactual explanations have shown promising results as a post-hoc framework to make image classifiers more explainable. In this paper, we propose DiME, a method allowing the generation of counterfactual images using the recent diffusion models. By leveraging the guided generative diffusion process, our proposed methodology shows how to use the gradients of the target classifier to generate counterfactual explanations of input instances. Further, we analyze current approaches to evaluate spurious correlations and extend the evaluation measurements by proposing a new metric: Correlation Difference. Our experimental validations show that the proposed algorithm surpasses previous State-of-the-Art results on 5 out of 6 metrics on CelebA.

**Keywords:** Counterfactual Explanations, Diffusion Models, Spurious Correlation Detection

## 1 Introduction

Convolutional neural networks (CNNs) reached performances unimaginable a few decades ago, thanks to the adoption of very large and deep models (e.g. with hundreds of layers and nearly billions of trainable parameters). Yet, it is difficult to explain their decisions because they are highly non-linear and over-parametrized. Moreover, for real-life applications, if a model exploits spurious correlations of data to forecast a prediction, the end-user will doubt the validity of the decision. Particularly, in high-stake scenarios like medicine or critical systems, ML must guarantee the usage of correct features to compute a prediction and prevent counterfeit associations. For this reason, the Explainable Artificial Intelligence (XAI) research field has been growing in recent years to progress towards understanding the decision-making mechanisms in black-box models.

In this paper, we focus on *post-hoc* explanation methods. Notably, we concentrate on the growing branch of Counterfactual Explanations (CE) [62]. CEs aim to create minimal but meaningful perturbations of an input sample to change the original decision given by a black-box model. Although the objective between CE and adversarial examples share some similarities [44], the CEs' perturbations must be understandable. In contrast, adversarial examples [37] contain high-frequency noise *indistinguishable* for the human eye. On overall, CEs target three goals: (*i*) create proximal images with sparse modifications, *i.e.* instances

with the smallest perturbation, *(ii)* the explanations must be realistic and understandable by a human, and *(iii)* the counterfactual generation method must create diverse instances. In general, counterfactual explanations seek to reveal the learned correlations related to the model’s decisions.

Multiple works on CE use generative models to create tangible changes in the image [28, 48, 51]. Further, these architectures recognize the factors to generate images near the image-manifold [4]. Given the recent advances within image synthesis community, we propose DiME: Diffusion Models for counterfactual Explanations. DiME harnesses the denoising diffusion probabilistic models [19] to produce CEs. For simplicity, we will refer to these models as diffusion models or DDPMs. To the best of our knowledge, we are the first to exploit these new synthesis methods in the context of CE.

Diffusion models offer several advantages compared to alternate generative models, such as GANs. First of all, DDPMs have several latent spaces; each one controls coarse and fine-grained details. We take advantage of low-level noise latent spaces to generate semantically-meaningfully changes in the input image. These spaces only have been recently studied by [38] for inpainting. Secondly, due to their probabilistic nature, they produce a diverse set of images. Stochasticity is ideal for CEs because multiple explanations may explain a classifier’s error modes. Third, Nichol and Dhariwal [42] results suggest that DDPMs cover a broader range of the target image distribution. Indeed, they noticed that for similar FID, the recall is much higher on the improved precision-recall metrics [32]. Finally, DDPMs’ training is more stable than the State-of-the-Art synthesis models, notably GANs. Due to their relatively new development, DDPMs are under-studied, and multiple aspects are yet to be deciphered.

We contribute a small step into the XAI community by studying the low-level noised latent spaces of DDPMs in the context of counterfactual explanations. We summarize our contributions as follows:

- DiME uses the recent diffusion models to generate counterfactual examples. Unlike other generative models, our CE algorithm does not require training the diffusion model in a conditioned way or retraining it using gradients, *i.e.* we rely on a single trained unconditional DDPM to achieve our objective.
- We derive a new way to leverage an existing (target) classifier to guide the generation process instead of ones trained on noisy instances.
- We set a new State-of-the-Art result on CelebA, surpassing the previous works on counterfactual explanations on the FID, FVA, and MNAC metrics for the *Smile* attribute and the FID and MNAC for the *Young* feature.
- We show that the MNAC provides a false sense of evaluating counterfactuals correctly. So we introduce a new metric, dubbed Correlation Difference, to evaluate subtle spurious correlations on a CE setting.

## 2 Related Work

Our work contributes to the field of XAI, within which two families can be distinguished: interpretable-by-design and *post-hoc* approaches. The former includes,

at the design stage, human interpretable mechanisms [2, 3, 6, 9, 22, 40, 69]. The latter aims at understanding the behavior of existing ML models without modifying their internal structure. Our method belongs within this second family. The two have different objectives and advantages; one benefit of *post-hoc* methods is that they rely on existing models that are known to have good performance, whereas XAI by design often leads to a performance trade-off.

**Post-hoc methods:** In the field of *post-hoc* methods, there are several explored directions. Model Distillation strategies [13, 58] approach explainability through fitting an interpretable model on the black-box models’ predictions. In a different vein, some methods generate explanation in textual form [17, 43, 66]. When it comes to explaining visual information, feature importance is arguably the most common approach, often implemented in the form of saliency maps computed either using the gradients within the network [8, 27, 33, 53, 63, 72] or using the perturbations on the image [45, 46, 61, 68]. Concept attribution methods seek the most recurrent traits that describe a particular class or instance. Intuitively, concept attribution algorithms use [29] or search [13, 14, 67, 73] for human-interpretable notions such as textures or shapes.

**Counterfactual Explanations (CEs):** CEs is a branch of post-hoc explanations. They are relevant to legally justify decisions made automatically by algorithms [62]. In a nutshell, a CE is the smallest meaningful change to an input sample to obtain a desirable outcome of the algorithm. Some recent methods [15, 64] exploit the query image’s regions and a different classified picture to interchange semantic appearances, creating counterfactual examples. Other works [52, 62] leverage the input image’s gradients with respect to the target label to create meaningful perturbations. Conversely, [1] find patterns via prototypes that the image must contain to alter its prediction. Similarly, [36, 47] follow a prototype-based algorithm to generate the explanations. Even Deep Image Priors [59] and Invertible CNNs [23] have shown the capacity to produce counterfactual examples. Furthermore, theoretical analyses [25] found similarities between counterfactual explanations and adversarial attacks.

Due to the nature of the problem, the generation technique used is the key element to produce data near the image manifold. For instance, [12] optimizes the residual of the image directly using an autoencoder as a regularizer. Other works propose to use generative networks to create the CEs, either unconditional [41, 48, 54, 71] or conditional [34, 55, 60]. In this paper, we adopt more recent generation approaches, namely *diffusion models*; an attempt never considered in the past for counterfactual generation.

**Diffusion Models:** Diffusion models have recently gained popularity in the image generation research field [19, 56]. For instance, DDPMs approached inpainting [49], conditional and unconditional image synthesis [10, 19, 42], super-resolution [50], even fundamental tasks such as segmentation [5], providing performance similar or even better than State-of-the-Art generative models. Further, studies like [20, 57] show score-based approaches and diffusion are alternative formulations to denoise the reverse sampling for data generation. Due to the

recursive generation process, DDPMs sampling is expensive. Many works have studied alternative approaches to accelerate the generation process [31, 65].

The recent method of [11] targets conditional image generation with diffusion models, which they do by training a specific classifier on noisy instances to bias the generation process. Our work bears some similarities to this method, but, in our case, explaining an existing classifier trained uniquely in clean instances poses an additional challenge. In addition, unlike past diffusion methods, we perform the image editing process from an intermediate step rather than the final one. To the best of our knowledge, no former study has considered diffusion models to explain a neural network counterfactually.

### 3 Methodology

#### 3.1 Preliminaries

We begin by introducing the generation process of diffusion models. They rely on two Markov chain sampling schemes that are inverse of one another. In the forward direction, the sampling starts from a natural image  $x$  and iteratively sample  $z_1, \dots, z_T$  by replacing part of the signal with white Gaussian noise. More precisely, letting  $\beta_t$  be a prescribed variance, the forward process follows the recursive expression:

$$z_t \sim \mathcal{N}(\sqrt{1 - \beta_t} z_{t-1}, \beta_t I), \quad (1)$$

where  $\mathcal{N}$  is the normal distribution,  $I$  the identity matrix, and  $z_0 = x$ . In fact, this process can be simulated directly from the original sample with

$$z_t \sim \mathcal{N}(\sqrt{\alpha_t} x, (1 - \alpha_t) I), \quad (2)$$

where  $\alpha_t := \prod_{k=1}^t (1 - \beta_k)$ . For clarification, through the rest of the paper, we will refer to clean images with an  $x$ , while noisy ones with a  $z$ .

In the reverse process, a neural network recurrently denoises  $z_T$  to recover the previous samples  $z_{T-1}, \dots, z_0$ . This network takes the current time step  $t$  and a noisy sample  $z_t$  as inputs, and produces an average sample  $\mu(t, z_t)$  and a covariance matrix  $\Sigma(t, z_t)$ , shorthanded as  $\mu(z_t)$  and  $\Sigma(z_t)$ , respectively. Then  $z_{t-1}$  is sampled with

$$z_{t-1} \sim \mathcal{N}(\mu(z_t), \Sigma(z_t)). \quad (3)$$

So, the DDPM algorithm iteratively employs Eq. 3 to generate an image  $z_0$  with zero variance, *i.e.* a clean image. Some diffusion models use external information, such as labels, to condition the denoising process. However, in this paper, we employ an unconditional DDPM.

In practice, the series of variances  $\{\beta_t\}$  are chosen such that  $z_T \sim \mathcal{N}(0, I)$ . Further, the DDPM's trainable parameters are fitted so that the reverse and forward processes share the same distribution. For details on training schemes, we recommend the studies of Ho *et al.* [19] and Nichol and Dhariwal [42] to the reader. Once the network is trained, one can rely on the reverse Markov chain

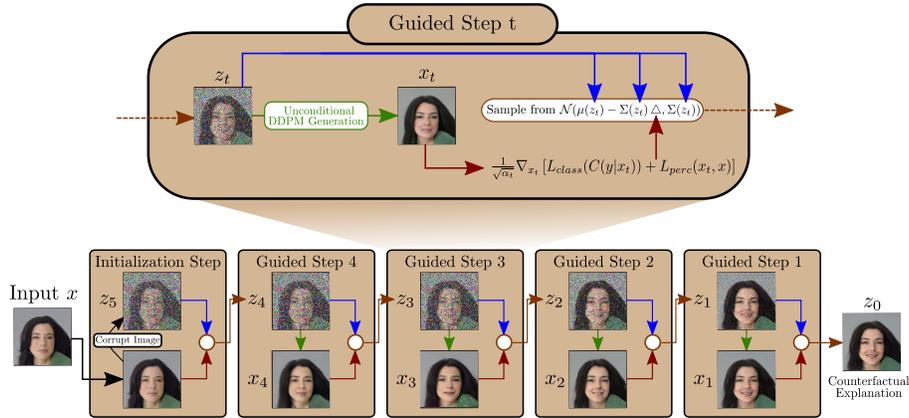


Fig. 1: **DiME: Diffusion Models for Counterfactual Explanations**. Given an input instance  $x$ , we perturb it following Eq. 2 to get  $z_\tau$  (here  $\tau = 5$ ). At time step  $t$ , we use the DDPM model to generate a clean image  $x_t$  to obtain the clean gradient  $L_{class}$  and  $L_{perc}$  with respect to  $x_t$ . Finally, we sample  $z_{t-1}$  using the guiding optimization process on Eq. 4, using the previously extracted clean gradients.

process to generate a clean image from a random noise image  $z_T$ . Besides, the sampling procedure can be adapted to optimize some properties following the so-called *guided diffusion* scheme proposed in [11]<sup>1</sup>:

$$z_{t-1} \sim \mathcal{N}(\mu(z_t) - \Sigma(z_t) \nabla_{z_t} L(z_t; y), \Sigma(z_t)), \quad (4)$$

where  $L$  is a loss function using  $z_t$  to specify the wanted property of the generated image, for example, to condition the generation on a prescribed label  $y$ .

### 3.2 DiME: Diffusion Models for Counterfactual Explanations

We take an image editing standpoint on CE generation, as illustrated Fig. 1. We start from a query image  $x$ . Initially, we rely on the forward process starting from  $x_\tau = x$  to compute a noisy version  $z_\tau$ , with  $1 \leq \tau \leq T$ . Then we go back in the reverse Markov chain using the guided diffusion (Eq 4) to recover a counterfactual (hence altered) version of the query sample. Building upon previous approaches for CEs based on other generative models [26,55,62], we rely on a loss function composed of two components to steer the diffusion process: a classification loss  $L_{class}$ , and a perceptual loss  $L_{perc}$ . The former guides the image edition into imposing the target label, and the latter drives the optimization in terms of proximity.

<sup>1</sup> In [11], the guided diffusion is restricted to a specific classification loss. Still, for the sake of generality and conciseness, we provide its extension to an arbitrary loss

In the original implementation of the guided diffusion [11], the loss function uses a classifier applied directly to the current noisy image  $z_t$ . This approach is appropriate since the considered classifier can make robust predictions under noisy observations, *i.e.* it was trained on noisy images. Regardless, such an assumption on the classifier under scrutiny would imply a substantial limitation in the context of counterfactual examples. We circumvent this obstacle by adapting the guided diffusion mechanism. To simplify the notations, let  $x_t$  be the clean image produced by the iterative unconditional generation on Eq 3 using as the initial condition  $z_t$ . In fact, this makes  $x_t$  a *function* of  $z_t$  because we denoise  $z_t$  recursively with the diffusion model  $t$  times to obtain  $x_t$ . Luckily, we can safely apply the classifier to  $x_t$  since it is not noisy. So, we express our loss as:

$$L(z_t; y, x) = \underbrace{\mathbb{E}[\lambda_c L_{class}(C(y|x_t)) + \lambda_p L_{perc}(x_t, x)]}_{\tilde{L}(x_t; y, x)}, \quad (5)$$

where  $C(y|x_t)$  is the posterior probability of the category  $y$  given  $x_t$ , and  $\lambda_c$  and  $\lambda_p$  are constants. Note that an expectation is present due to the stochastic nature of  $x_t$ . In practice, computing the loss gradient would require sampling several realizations of  $x_t$  and taking an empirical average. We restrict ourselves to a single realization per step  $t$  for computational reasons and argue that this is not an issue. Indeed, we can partly count on an averaging effect along the time steps to cope with the lack of individual empirical averaging. Besides, the stochastic nature of our implementation is, in fact, an advantage because it introduces more diversity in the produced CEs, a desirable feature as advocated by [48].

Using this strategy, the dependence of the loss on  $x_t$ , rather than directly from  $z_t$ , renders the gradient computation more challenging. Indeed, formally it would require to apply back-propagation from  $x_t$  back to  $z_t$ :

$$\nabla_{z_t} L(z_t; y, x) = \left( \frac{Dx_t}{Dz_t} \right)^T \cdot \nabla_{x_t} \tilde{L}(x_t; y, x). \quad (6)$$

Unfortunately, this computation requires retaining Jacobian information throughout the entire computation graph, which is very deep when  $t$  is close to  $\tau$ . As a result, backpropagation is too memory intensive to be considered an option. To bypass this pitfall, we shall rely on the forward sampling process, which operates in a single stage (Eq. 2). Using the reparametrization trick [30], one obtains

$$z_t = \sqrt{\alpha_t} x_t + \sqrt{1 - \alpha_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \quad (7)$$

Thus, by solving  $x_t$  from  $z_t$ , we can leverage the gradients of the loss function with respect to the noisy input, a consequence of the chain rule. Henceforth, the gradients of  $L$  with respect to the noisy image become

$$\nabla_{z_t} L(z_t; y, x) = \frac{1}{\sqrt{\alpha_t}} \nabla_{x_t} \tilde{L}(x_t; y, x). \quad (8)$$

This approximation is possible since the DDPM estimates the reverse Markov chain to fit the forward corruption process. Thereby, both processes are similar.

To sum up, Fig. 1 depicts the generation of a counterfactual explanation with our algorithm: DiME. We start by corrupting the input instance  $x = x_\tau$  following Eq. 2 up to the noise level  $t = \tau$ . Then, we iterate the following two stages until  $t = 0$ : (i) First, using the gradients of the previous clean instance  $x_{t-1}$ , we guide the diffusion process to obtain  $z_{t-1}$  using Eq. 4 with the gradients computed in Eq. 8. (ii) Next, we estimate the clean image  $x_t$  for the current time step  $z_{t-1}$  with the unconditional generation pipeline of DDPMs. The final instance is the counterfactual explanation. If we do not find an explanation that fools the classifier under observation, we increase the constant  $\lambda_c$  and repeat the process.

**Implementation Details.** To train the unconditional DDPM model, we used the publicly available code of [11]. We include all training and architectural details in the supplemental material. In practice, we incorporate additionally an  $\ell_1$  loss,  $\eta \|z_t - x\|_1$ , between the noisy image  $z_t$  and the input  $x$  to improve the  $\ell_1$  metric on the pixel space. We empirically set  $\eta$  small to avoid any significant impact on the quality of the explanations. Our diffusion model generates faces using 500 diffusion steps from the normal distribution. We re-spaced the sampling process to boost inference speed to generate images with 200 time-steps at test time. We use the following hyper-parameters settings:  $\lambda_p = 30$ ,  $\eta = 0.05$ , and  $\tau = 60$ . Finally, we set  $\lambda_c \in \{8, 10, 15\}$  to iteratively find the counterfactuals. We consider that our method failed if we do not find any explanation after exhausting the values of  $\lambda_c$ .

## 4 Experiments

**Experimental goals.** In this section, we evaluate DiME, our CE approach, using standard metrics. Also, we develop new tools to go beyond the current evaluation practices. Let us then recap the principles of current evaluation metrics, following previous works [48, 55]. The first goal of CEs is to create *realistic* explanations that *mislead* the classifier under observation. The capacity to change the classifier decision is typically exposed as a flip ratio (FR). Following the image synthesis research literature, the Frechet Inception Distance [18] (FID) measures the fidelity of the image distribution. The second goal of CE methods is to create proximal and sparse images. Among other tools, the XAI community adopted the Face Verification Accuracy [7] (FVA) and Mean Number of Attributes Changed (MNAC) [48]. On the one hand, the MNAC metric looks at the face attributes that changed between the input image and its counterfactual explanation, disregarding if the individual’s identity changed. On the other hand, the FVA looks at the individual’s identity without considering the difference of attributes.

Despite their usefulness, the previous metrics miss two important properties of CEs. Indeed, following [48], to give a sense of trust in a classifier, the CEs must also produce diverse explanations and ensure that the classifier is not subject to spurious correlations. On the one hand, generating diverse explanations is useful to discover the brittleness of CNNs. Mothilal *et al.* [39] propose a pair-wise distance to evaluate the diversity of counterfactual examples. Nevertheless, this

	Smile			Young		
Method	FID ( $\downarrow$ )	FVA ( $\uparrow$ )	MNAC ( $\downarrow$ )	FID ( $\downarrow$ )	FVA ( $\uparrow$ )	MNAC ( $\downarrow$ )
xGEM+ [28]	66.9	91.2	-	59.5	97.5	6.70
PE [55]	35.8	85.3	-	53.4	72.2	3.74
DiVE [48]	29.4	97.3	-	33.8	<b>98.2</b>	4.58
DiVE <sup>100</sup>	36.8	73.4	4.63	39.9	52.2	4.27
DiME	<b>3.17</b>	<b>98.3</b>	<b>3.72</b>	<b>4.15</b>	95.3	<b>3.13</b>

Table 1: **State-of-the-Art results.** We compare our model performance against the State-of-the-Art on the FID, FVA and MNAC metrics. The values in **bold** are the best results. All metrics were extracted from [48]. Our model has a 10 fold improvement on the FID metric. We extracted all results from Rodriguez *et al.*' work [48].

work is exclusively dedicated to tabular data. We propose a simple adaptation to images based on the LPIPS metric [70]. On the other hand, current assessments to detect spurious correlations (e.g., in [55]) are quite extreme. They rely on modified datasets by entangling two attributes artificially to a full extent, e.g., all males are smiling, and all women are not. They also assume that in standard benchmarks, attributes are not entangled at all. Under this assumption, a classifier trained in this setting can be safely used as an oracle for the attributes, as proposed for computing MNAC. Actually, we show that this assumption can be largely erroneous and, therefore, challenge the derived metrics' validity. Based on our analysis, we designed a metric called *Correlation Difference* to assess if a counterfactual approach adequately reveals subtle "spurious correlations" (see Section 4.3).

**Dataset.** In this paper, we study the CelebA dataset [35]. Following standard practices, we preprocess all images to a  $128 \times 128$  resolution. CelebA contains 200k images, labeled with 40 binary attributes. Previous works validate their methods on the *smile* and *young* binary attributes, ignoring all other features. Finally, the architecture to explain is a DenseNet121 [21]. Given the binary nature of the task, the target label is always the opposite of the prediction. If the model correctly estimates an instance's label, we flip the model's forecast. Else, we modify the input image to classify the image correctly.

#### 4.1 Realism, Proximity and Sparsity Evaluation

To compute the FID, the FVA, and the MNAC, we consider only those successful counterfactual examples, following previous studies [48, 55]. The FVA is the standard metric for face recognition. To measure this value, we used the cosine similarity between the input image and its produced counterfactual on the feature space of a ResNet50 [16] pretrained model on VGGFace2 [7]. The instance and the explanation share the same identity if the similarity is higher

than 0.5. So, the FVA is the mean number of faces sharing the same identity with their corresponding CE. To compute the MNAC, we fine-tuned the VGGFace2 model on the CelebA dataset. We refer to the fine-tuned model as the *oracle*. Thus, the MNAC is the mean number of attributes for which the oracle switch decision under the action of the CE. For a fair comparison with the State-of-the-Art, we trained all classifiers, including the fine-tuned ResNet50 for the MNAC assessment, using the DiVE’s [48] available code.

DiVE do not report their flip rate (FR). This raises a concern over the fairness of comparing our methods. Since some metrics depend highly on the number of samples, especially FID, we recomputed their CEs. To our surprise, their flip ratio was relatively low (44.6% for the smile category). In contrast, we achieved a success rate of 97.6 and 98.9 for the smile and young attributes, respectively. Therefore, we calculated the counterfactual explanations with 100 optimization steps and reported the results as DiVE<sup>100</sup>. DiVE<sup>100</sup>’s success rates are 92.0% for smile and 93.4% for young, which is comparable to ours.

We show DiME’s performance in Table 1. Our method beats the previous literature in five out of six metrics. For instance, we have a  $\sim 10$  fold improvement on the FID metric for the smile category, while the young attribute has an  $\sim 8$  fold improvement. We credit these gains to the generative capabilities of the diffusion model. Further, our generation process does not require entirely corrupting the input instance; hence, the coarse details of the image remain. The other methods rely on latent space-based architectures. Thus, they require to compact essential information removing outlier data. Consequently, the generated CEs cannot reconstruct the missing information, losing significant visual components of the image statistics.

Despite the previous advantages, we cannot fail to notice that DiME is less effective in targetting the young attribute than the smile. The smile and young attributes have distinct features. The former is delineated by localized regions, while the latter scatters throughout the entire face. Thus, the gradients produced by the classifier differ between the attributes of choice; for the smile attribute, the gradients are centralized while they are outspread for the young attribute. We believe that this subtle difference underpins the slight drop of performance (especially with respect to FVA) in the young attribute case. This hypothetical explanation should be confirmed by a more systematic study of various attributes, though this phenomenon is out of scope of the paper.

## 4.2 Diversity Assessment

One of the most crucial traits of counterfactual explanations methodologies is the ability to create multiple and diverse examples [39, 48]. As stated in the methodology section, DiME’s stochastic properties enable the sampling of diverse counterfactuals. To measure the capabilities of different algorithms to produce multiple explanations, we computed the mean pair-wise LPIPS [70] metric between five independent runs. A higher LPIPS means increased perceptual dissimilarities between the explanations, hence, more diversity. To compute the



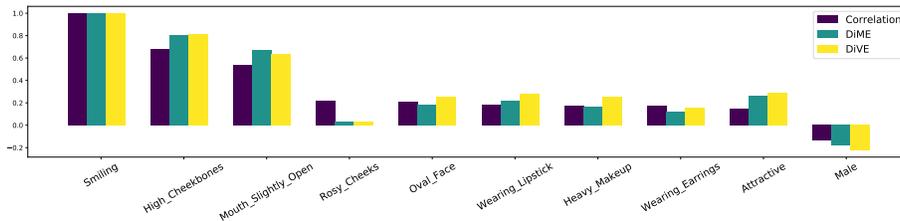
Fig. 2: **Diversity Counterfactual examples.** The classifier predicts first two input images as non-smiley and the last two as smiley. In this example all explanations fool the classifier. Our CE pipeline is capable of synthesising diverse counterfactuals without any additional mechanism.

evaluation metric, we use all counterfactual examples, even the unsuccessful instances, because we search the capacity of exploring different traits. Note that we exclude the input instance to compute the metric since we search for the dissimilarities between the counterfactuals. We compared DiME’s performance with DiVE<sup>100</sup> and its Fisher Spectral variant on a small partition of the validation subset.

We visualize some examples in Fig. 2 and show the performance of the five runs on the supplementary material. We obtained an LPIPS value of 0.213. In contrast, DiVE [48] and its Spectral Fisher variant obtained an LPIPS of 0.044 and 0.086, respectively. Recall that DiME does not have an explicit mechanism to create diverse counterfactuals. Its only mechanism is the stochasticity within the sampling process (Eqs. 3 and 4). In contrast, DiVE relies on a diversity loss when optimizing the eight explanations. Yet, our methodology achieves higher diversity with the LPIPS metric even without an explicit mechanism.

### 4.3 Discovering Spurious Correlations

The end goal of counterfactual examples is to uncover the modes of error of a target model. Current evaluation protocols [55] search to assess the spurious correlations by inducing artificial entanglements between two supposedly uncorrelated attributes. Conventionally, the experiment involves mixing the smile and gender attributes. The goal is then to evaluate whether or not the CE algorithm



**Fig. 3: Spurious Correlation Detection.** We show the top 9 most correlated attributes in the label space with “smile”. We obtained the Pearson Correlation Coefficient from the ground truth on the training set. Albeit the difference in the MNAC performance, DiME and DiVE achieve to detect the spurious correlations similarly. We show all the remaining attributes in the supplementary material.

is able to reveal the correlation. To assess this capability, it is common to verify if both the target and the entangled attributes change when producing the CEs. In our opinion, such an extreme experiment does not shed light on the ability to reveal spurious correlations in real situations. Indeed, the considered configuration assumes that only two labels are entangled and that this entanglement is complete.

In fact, as depicted in Fig. 3, in real datasets such as CelebA, many labels are correlated at multiple levels. As a result, this phenomenon calls the previously proposed correlation experiment into question. It also raises concerns about the value of the MNAC metric, or measurement tools such as the LVS [24]. As a matter of fact, while small MNAC values are often considered desirable (see [48, 55, 62]), the presence of spurious correlations challenges this interpretation. Indeed, consider the following illustrating scenario comparing two CE algorithms: the first one exposes all spurious correlations correctly; the second one can solely edit the main feature. Since the first method produces many changes, it will display a *high* MNAC while the alternate algorithm reaches a *low* MNAC value. This false sense of high performance does not reflect the true accomplishment of the first model: detecting spurious correlations. So, we propose to amend the MNAC measurement into a new metric called the Correlation Difference (CD), more adapted to assess the capacity of CEs to reveal spurious correlations.

The goal of CD is to measure the difference between the true correlations and the changes produced by the explanations. To achieve this, let  $c_{qa}$  be the true correlation computed on the label space between the attribute labels  $q$  and  $a$ . To measure the correlations in the prediction space, first we define

$$\delta_a(x) = O_a(M_q(x)) - O_a(x), \quad (9)$$

where  $O_a(\cdot)$  is the oracle’s binary prediction of its input for the attribute  $a$ ,  $M_q$  is the counterfactual method targeting the query attribute  $q$  on an image  $x$ . This measure looks at the signed changes implied by  $M_q$  on  $x$ . So, now we can measure

the relative changes on the attributes when computing a counterfactual example. Therefore, we can calculate the correlation coefficient  $c_{qa}^M$  between  $\{\delta_q(x)\}_x$  and  $\{\delta_a(x)\}_x$  to compare it with  $c_{qa}$ , the true correlation<sup>2</sup>. Accordingly, CD is:

$$CD = \sum_a |c_{qa} - c_{qa}^M|. \quad (10)$$

We apply our proposed metric on DiME and DiVE<sup>100</sup>'s counterfactual explanations. We got a CD of 2.30 while DiVE<sup>100</sup> 2.36 on CelebA's validation set, meaning that DiVE<sup>100</sup> lags behind DiME. However, the margin between the two approaches is only slender. This reveals our suspicions: the MNAC results presented in Table 1 give a misleading impression of a robust superiority of DiME over DiVE<sup>100</sup>.

#### 4.4 Impact of the noise-free input of the classifier

In this section, we assess the impact of our main adjustment over the original guided diffusion process. Recall that we argued that it is important to apply the classifier on noise-free images  $x_t$  and not on the current noisy version  $z_t$  in order to obtain a robust gradient direction. To validate this claim, we compare our approach to an alternative, dubbed *Direct*. It uses the gradient of the classifier applied directly to the noisy instance  $z_t$ . In this case, we removed the constant  $1/\sqrt{\alpha_t}$  since we compute directly the gradients with respect to  $z_t$ . To complete the picture, we also consider two additional variations of our approach. The first one, called *Naive*, uses the gradient of the input image at each time step to guide the optimization process. Therefore, it is not subject to noise issues, but it disregards the guidance that was already applied until time step  $t$ . The second variation is a near duplicate of DiME except for the fact that it ends the guided diffusion process as soon as  $x_t$  fools the classifier. We name this approach *Early Stopping*. Eventually, we will also evaluate the DDPM generation without any guiding and beginning from the corrupted image at time-step  $\tau$  to mark a reference of the performance of the DDPM model.

*Notes on considered metrics:* In addition to FID, FR and  $\ell_1$  metrics, we also evaluate the following metric:  $BKL(y||M_q(x)) = 1 - C(y|M_q(x))$ . It is the complement of the target label's probability, but whose origin is a bounded remapping of a KL divergence, hence the notation *BKL*. A low BKL means that the classifier under observation classifies the counterfactual example  $M_q(x)$  with high confidence and is effectively fooled by the CE.

Also, given that many variants are considered, we created a small and randomly selected mini-val to evaluate the various metrics. Besides, given that the different baselines can display varying levels of FR, we condition the FID computation on the successful CEs only. However, it is well known that FID is strongly biased, especially when using a low number of samples. To mitigate this bias,

<sup>2</sup> The series  $\{\delta_q(x)\}_x$  and  $\{\delta_a(x)\}_x$  uses solely the successful counterfactual explanations. Further, we used the Pearson Correlation Coefficient to compute  $c_{qa}$  and  $c_{qa}^M$ .

Method	FR ( $\uparrow$ )	FID <sup>+</sup> ( $\downarrow$ )	$\ell_1$ ( $\downarrow$ )	BKL ( $\downarrow$ )
Direct	19.7	<b>50.51</b>	0.0454	0.297
Naive	70.0	98.93 $\pm$ 2.36	0.0624	0.139
Early Stopping	97.3	51.97 $\pm$ 0.77	0.0467	0.350
Unconditional	8.6	53.22 <sup>§</sup> $\pm$ 0.98	0.0492	0.265
DiME	<b>97.9</b>	<b>50.20</b> $\pm$ 1.00	<b>0.0430</b>	<b>0.076</b>

Table 2: **DiME vs “Naive” variations.** This table shows the advantages of the proposed adjustment to incorporate the classifier under observation. We clearly see that including the clean gradients benefits DiME on all metrics, especially the FR.

we use the same number of CEs for each baseline (the least number of successful CEs) and repeat this computation 10 times to report a mean and standard deviation. We denote this fair FID as FID<sup>+</sup>. Similarly, we compute the  $\ell_1$  and BKL solely for successful counterfactuals.

We show the results of the different variations in Table 2. The most striking point is that when compared to the Naive and Direct approaches, the unimpaired version of DiME is the most effective in terms of FR by a large margin. This observation validates the need for our adjustment of the guided diffusion process. Further, our approach is also superior to all other variations in terms of the other metrics. At first glance, we expected the unconditional generation to have better FID than DiME and the ablated methods. However, we believe that the perceptual part of our loss is beneficial in terms of FID. Therefore, the unconditional FID is higher. Similarly, the early stopping variant is also impacted in terms of FID and BKL because the optimization is brought to an end prematurely.

We complement this ablation study in the supplementary material. In particular, we explore the role of the initial noise level  $\tau$  and the mixing factor  $\lambda_c$ . Other quantitative and qualitative results are presented therein.

#### 4.5 Qualitative Results

We visualize some inputs (left) and the counterfactual examples (right) produced by DiME in Fig. 4. We show visualizations for the attributes smile and young, yet we will include visualizations for other categories in the supplementary material. At first glance, the results reveal that the model performs semantical editings into the input image. In addition, uncorrelated features and coarse structure remain almost unaltered. We observe slight variations on some items, such as the pendants, or out-of-distribution shapes such as hands. DiME fails to reconstruct the exact shape of these objects, but the essential aspect remains the same.

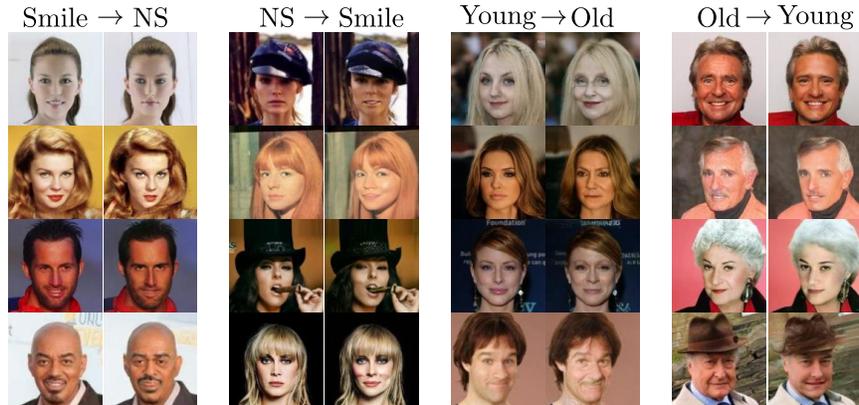


Fig. 4: **Qualitative Results.** We visualize some images and its corresponding counterfactual explanation produced by our proposed approach. Our methodology achieves to incorporate small but perceptually tangible changes in the image. NS stands for Non-Smiley.

#### 4.6 Limitations

Our pipeline for counterfactual explanation has several limitations. Although we show the benefits of using our model to generate CEs, we are far from accomplishing all aspects crucial for the XAI community. First of all, our method is slow and computationally demanding. Since we are using DDPMs, we adopt most of their limitations. For instance, we need to use the DDPM model  $\sim 1800$  times to generate a single explanation. This aspect is undesired whenever the user requires an explanation on the fly. Finally, we require access to the training data. This limitation is common in many previous studies. However, this aspect is vital in fields where data is sensitive. Although access to the training data is permitted in many cases, we restrict ourselves to using the data without any labels.

## 5 Conclusion

In this paper, we explore the novel diffusion models in the context of counterfactual explanations. By harnessing the conditional generation of the guided diffusion, we achieve successful counterfactual explanations through DiME. These explanations follow the requirements given by the XAI community: they produce a small but tangible change in the image while remaining realistic. The performance of DiME is confirmed based on a battery of standard metrics. DiME also exhibits strong diversity in the produced explanation. This is partly inherited from the intrinsic features of diffusion models, but it also results from a careful design of our approach. Further, we show that the current approach to validate the sparsity of CE has significant conflicts with the assessment of spurious

correlation detection. Finally, our proposed metric, Correlation Difference, correctly measures the impact of measuring the subtle correlation between labels. We hope that our work opens new ways to compute and evaluate counterfactual explanations.

## Acknowledgements

Research reported in this publication was supported by the Agence Nationale pour la Recherche (ANR) under award number ANR-19-CHIA-0017.

## References

1. Akula, A., Wang, S., Zhu, S.C.: Cocox: Generating conceptual and counterfactual explanations via fault-lines. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**(03), 2594–2601 (Apr 2020)
2. Alaniz, S., Marcos, D., Schiele, B., Akata, Z.: Learning decision trees recurrently through communication. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 13518–13527 (June 2021)
3. Alvarez-Melis, D., Jaakkola, T.S.: Towards robust interpretability with self-explaining neural networks. In: *Advances in neural information processing systems (NeurIPS)* (2018)
4. Arora, S., Risteski, A., Zhang, Y.: Do GANs learn the distribution? some theory and empirics. In: *International Conference on Learning Representations* (2018), <https://openreview.net/forum?id=BJehNfW0->
5. Baranchuk, D., Voynov, A., Rubachev, I., Khrukov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: *International Conference on Learning Representations* (2022)
6. Bohle, M., Fritz, M., Schiele, B.: Convolutional dynamic alignment networks for interpretable classifications. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 10029–10038 (June 2021)
7. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. pp. 67–74 (2018). <https://doi.org/10.1109/FG.2018.00020>
8. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. pp. 839–847 (2018). <https://doi.org/10.1109/WACV.2018.00097>
9. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: Deep learning for interpretable image recognition. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 32. Curran Associates, Inc. (2019)
10. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 14367–14376 (October 2021)
11. Dhariwal, P., Nichol, A.Q.: Diffusion models beat GANs on image synthesis. In: *Thirty-Fifth Conference on Neural Information Processing Systems* (2021)

12. Dhurandhar, A., Chen, P.Y., Luss, R., Tu, C.C., Ting, P., Shanmugam, K., Das, P.: Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 31. Curran Associates, Inc. (2018)
13. Ge, Y., Xiao, Y., Xu, Z., Zheng, M., Karanam, S., Chen, T., Itti, L., Wu, Z.: A peek into the reasoning of neural networks: Interpreting with structural visual concepts. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2195–2204 (June 2021)
14. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 32. Curran Associates, Inc. (2019)
15. Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S.: Counterfactual visual explanations. In: *ICML*. pp. 2376–2384 (2019)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 770–778 (2016)
17. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *Computer Vision – ECCV 2016*. pp. 3–19. Springer International Publishing (2016)
18. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 30. Curran Associates, Inc. (2017), <https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf>
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) *Advances in Neural Information Processing Systems*. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
20. Huang, C.W., Lim, J.H., Courville, A.: A variational perspective on diffusion-based generative models and score matching. In: *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models (2021)*
21. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017)
22. Huang, Z., Li, Y.: Interpretable and accurate fine-grained recognition via region grouping. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020)
23. Hvilshøj, F., Iosifidis, A., Assent, I.: Ecinn: efficient counterfactuals from invertible neural networks. *British Machine Vision Conference 2018, BMVC 2021* (2021)
24. Hvilshoj, F., Iosifidis, A., Assent, I.: On quantitative evaluations of counterfactuals. *ArXiv abs/2111.00177* (2021)
25. Ignatiev, A., Narodytska, N., Marques-Silva, J.: On relating explanations and adversarial examples. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 32. Curran Associates, Inc. (2019)
26. Jacob, P., Éloi Zablocki, Ben-Younes, H., Chen, M., Pérez, P., Cord, M.: Steex: Steering counterfactual explanations with semantics (2021)

27. Jalwana, M.A.A.K., Akhtar, N., Bennamoun, M., Mian, A.: Cameras: Enhanced resolution and sanity preserving class activation mapping for image saliency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16327–16336 (June 2021)
28. Joshi, S., Koyejo, O., Kim, B., Ghosh, J.: xgems: Generating exemplars to explain black-box models. ArXiv [abs/1806.08867](https://arxiv.org/abs/1806.08867) (2018)
29. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2668–2677. PMLR (10–15 Jul 2018)
30. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings (2014)
31. Kong, Z., Ping, W.: On fast sampling of diffusion probabilistic models. In: ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models (2021)
32. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. CoRR [abs/1904.06991](https://arxiv.org/abs/1904.06991) (2019)
33. Lee, J.R., Kim, S., Park, I., Eo, T., Hwang, D.: Relevance-cam: Your model already knows where to look. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14944–14953 (June 2021)
34. Liu, S., Kailkhura, B., Loveland, D., Han, Y.: Generative counterfactual introspection for explainable deep learning. In: 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). pp. 1–5 (2019)
35. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
36. Looveren, A.V., Klaise, J.: Interpretable counterfactual explanations guided by prototypes. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 650–665. Springer (2021)
37. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018)
38. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022)
39. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020)
40. Nauta, M., van Bree, R., Seifert, C.: Neural prototype trees for interpretable fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14933–14943 (June 2021)
41. Nemirovsky, D., Thiebaut, N., Xu, Y., Gupta, A.: CounterGAN: Generating realistic counterfactuals with residual generative adversarial nets. arXiv preprint [arXiv:2009.05199](https://arxiv.org/abs/2009.05199) (2020)
42. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models (2021)
43. Park, D.H., Hendricks, L.A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M.: Multimodal explanations: Justifying decisions and pointing to the

- evidence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
44. Pawelczyk, M., Agarwal, C., Joshi, S., Upadhyay, S., Lakkaraju, H.: Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis. arXiv:2106.09992 [cs] (Jun 2021)
  45. Petsiuk, V., Das, A., Saenko, K.: RISE: randomized input sampling for explanation of black-box models. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018. p. 151. BMVA Press (2018)
  46. Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V.I., Mehra, A., Ordonez, V., Saenko, K.: Black-box explanation of object detectors via saliency maps. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 11443–11452. Computer Vision Foundation / IEEE (2021)
  47. Poyiadzi, R., Sokol, K., Santos-Rodríguez, R., Bie, T.D., Flach, P.A.: Face: Feasible and actionable counterfactual explanations. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2020)
  48. Rodríguez, P., Caccia, M., Lacoste, A., Zamparo, L., Laradji, I., Charlin, L., Vazquez, D.: Beyond trivial counterfactual explanations with diverse valuable explanations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1056–1065 (October 2021)
  49. Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-image diffusion models. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
  50. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D., Norouzi, M.: Image super-resolution via iterative refinement. ArXiv **abs/2104.07636** (2021)
  51. Sauer, A., Geiger, A.: Counterfactual generative networks. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
  52. Schut, L., Key, O., Mc Grath, R., Costabello, L., Sacaleanu, B., Corcoran, M., Gal, Y.: Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In: Banerjee, A., Fukumizu, K. (eds.) Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 130, pp. 1756–1764. PMLR (13–15 Apr 2021)
  53. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
  54. Shih, S.M., Tien, P.J., Karnin, Z.: GANMEX: One-vs-one attributions using GAN-based model explainability. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 9592–9602. PMLR (2021)
  55. Singla, S., Pollack, B., Chen, J., Batmanghelich, K.: Explanation by progressive exaggeration. In: International Conference on Learning Representations (2020)
  56. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
  57. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021)
  58. Tan, S., Caruana, R., Hooker, G., Koch, P., Gordo, A.: Learning global additive explanations for neural nets using model distillation (2018)

59. Thiagarajan, J.J., Narayanaswamy, V., Rajan, D., Liang, J., Chaudhari, A., Spanias, A.: Designing counterfactual generators using deep model inversion. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems* (2021), <https://openreview.net/forum?id=iHisgL7PFj2>
60. Van Looveren, A., Klaise, J., Vacanti, G., Cobb, O.: Conditional generative models for counterfactual explanations. arXiv preprint arXiv:2101.10123 (2021)
61. Vasu, B., Long, C.: Iterative and adaptive sampling with spatial attention for black-box model explanations. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (March 2020)
62. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *arvard Journal of Law and Technology*, **31**(2), 841–887 (2018). <https://doi.org/10.2139/ssrn.3063289>
63. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (June 2020)
64. Wang, P., Li, Y., Singh, K.K., Lu, J., Vasconcelos, N.: Imagine: Image synthesis by image-guided model inversion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3681–3690 (June 2021)
65. Watson, D., Ho, J., Norouzi, M., Chan, W.: Learning to efficiently sample from diffusion probabilistic models. *CoRR* **abs/2106.03802** (2021)
66. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: F-vaegan-d2: A feature generating framework for any-shot learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019)
67. Yeh, C.K., Kim, B., Arik, S.Ö., Li, C.L., Pfister, T., Ravikumar, P.: On completeness-aware concept-based explanations in deep neural networks. *arXiv: Learning* (2020)
68. Yuhki Hatakeyama, Hiroki Sakuma, Y.K., Suenaga, K.: Visualizing color-wise saliency of black-box image classification models. In: *Asian Conference on Computer Vision (ACCV)* (2020)
69. Zhang, Q., Wu, Y.N., Zhu, S.C.: Interpretable convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018)
70. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018)
71. Zhao, Z., Dua, D., Singh, S.: Generating natural adversarial examples. In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net (2018)
72. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016)
73. Zhou, B., Sun, Y., Bau, D., Torralba, A.: Interpretable basis decomposition for visual explanation. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (September 2018)

## Supplementary material

### A Implementation Details

**DDPM Architectural and Training Details.** We trained the unconditional DDPM using the publicly available code of [11]. Our model has the same architecture as the ImageNet’s Unconditional DDPM of [11], except for two differences. (i) The number of diffusion steps for [11] is 1000 while we use 500 steps only. (ii) we reduced the number of inner channels from 256 to 128 given that CelebA’s complexity is far lower than ImageNet’s. We trained our model for 270.000 iterations with a batch size of 75 on 5 GPUs, *i.e.* a batch size of 15 per GPU. We set the learning rate to  $1 \times 10^{-4}$  with a weight decay of 0.05 and no dropout. Although we selected this configuration for the architecture and the training, we did not perform an exhaustive exploration since we are not searching to evaluate the diffusion model performance.

**Loss selection.** The selection of the losses influences the convergence of the stochastic optimization process for the CE. We chose the standard VGG19 perceptual loss as the  $L_{perc}$  loss. For the classification loss  $L_{class}$ , we opted to maximize directly logits of the target class instead of the log probability. More specifically, we minimize the negative logits.

### B Variability Evaluation

We report the performances of the five different runs in Table 3. Even when we set different initial conditions for each iteration, DiME is robust to many instantiations. We visualize more images for the variability in section E. Many results vary significantly, yet DiME solves the counterfactuals in most cases.

### C Correlation Discovery

In section 4.3 of the main manuscript we discussed the importance of our proposed metric CD. Nevertheless, we visualize only the top 9 attributes given that

Seed	FID(↓)	FR(↑)	$\ell_1$ (↓)	BKL(↓)
1	20.51	97.9	0.0430	0.076
2	20.60	97.6	0.0430	0.073
3	20.72	97.9	0.0431	0.067
4	20.67	97.7	0.0431	0.073
5	20.46	98.2	0.0430	0.076

Table 3: **Diversity experiments.** We ran our method five times, varying the initial seed. The results show that our method is robust to the initial conditions, although the visual elements vary significantly.

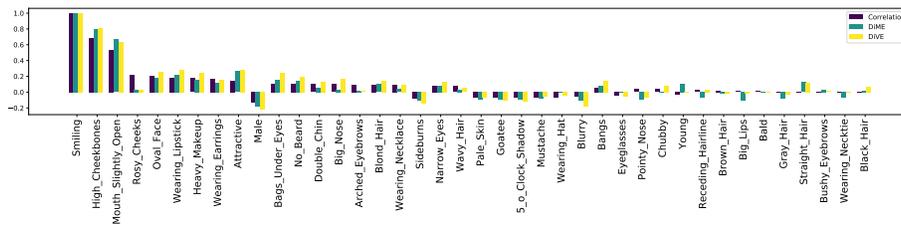


Fig. 5: **Correlation discovery.** We visualize all the correlation discovered by DiME and DiVE for all attributes on CelebA.

$\lambda_c$	FID <sup>+</sup> (↓)	FR(↑)	$\ell_1$ (↓)	BKL(↓)
8	22.93	80.1	<b>0.0427</b>	0.058
10	23.32	88.0	0.0432	0.041
15	25.87	97.7	0.0446	<b>0.019</b>
DiME	<b>22.48</b>	<b>97.9</b>	0.0430	0.076

Table 4: **Gradient Scales.** We show the impact of different scales choices. We computed the BKL and FID metrics solely from the successfully counterfactual explanations. Increasing the gradient scale  $\lambda_c$  decreases the FID and  $\ell_1$ . From the Flip Ratio results, we see that most explanations are produced with a low scale value, hence producing similar results in the pixel space with high fidelity. Harder instances require the use of an increased scale to successfully produce the counterfactual example. The FID<sup>+</sup> is computed taking the the same number of samples for the experiment  $\lambda_c = 8$ .

the other attributes are far less correlated. For the sake of completing the study, we added the rest of attributes on Fig. 5. Similarly, we observe that DiME and DiVE have similar capabilities finding correlations in the data.

## D Ablations Studies

To complement the ablation analysis on the components of DiME, we explore the variables that affect the generation of counterfactual explanations. On the one side, we analyze the impact of calibrating the gradients’ scale. On the other side, we study the effects of varying the initial noise level, considering that adding more noise implies removing more details. Finally, we visualize the evolution of the clean images produced at each time step of the guiding process.

### D.1 Gradients’ Scale Ablation

The work Dhariwal and Nichol [11] ablates the scale of the classifier. Their results indicate a positive correlation between the quality of the images, measured with

Steps	FID <sup>+</sup> (↓)	FR(↑)	$\ell_1$ (↓)	BKL(↓)
50	<b>20.19</b>	92.4	<b>0.0406</b>	0.100
60	20.94	97.9	0.0430	0.076
70	23.21	<b>99.7</b>	0.0479	<b>0.048</b>

Table 5: **Initialization Step.** We show the result of different  $\tau$  choices. The BKL and FID metrics are computed solely from the successfully counterfactual explanations. Using  $\tau = 60$  provides the best trade-off between image quality, Flip Ratio and similarity. We computed the FID<sup>+</sup> taking the the same number of samples for the experiment with less instances ( $\tau = 50$ ).

the FID, and the gradients’ scale. So, we perform a similar study; to find a counterfactual explanation, we optimize the image formation with three different scales  $\lambda_c \in \{8, 10, 15\}$  and choose the generated image with the smallest scale. Therefore, we analyze the individual contribution of each scale.

We report the results in Table 4. In opposition to [11], we observe that when increasing the scale, the quality as measured by FID drops; more precisely the FID value increases. This inconsistency with the observation of [11] remains to elucidate. Although it is out of the scope of our work, we can at least point out a few potential sources of discrepancy. First, the type of image edition that we perform is fundamentally different from the one considered in [11]. In their work, the task correspond to generate an image conditionally to a categorical label. This categorical, hence discrete, aspect of the condition may be at odd with soft constraints (i.e. small gradient scale). It is not present in our context. Among other differences, one can note that we start the denoising process from an intermediate step  $\tau \leq T$  while they start from the very last step  $T$ . Eventually, we have a specific way of computing the gradient.

In our particular context for CEs, we noticed a trade-off between the success rate and the quality of the images. Since we seek to produce sparse modifications, we created most explanations using the lowest scale. Thus, we enjoy the benefits of high-quality images. Further, we boost the FR by using higher scales at the cost of lowering the quality. Adding too much gradient produces out-of-distribution noise. The DDPM cannot recognize this noise, and therefore it produces artifacts on the image. However, these artifacts may coincide with patterns that impact the classifier response.

## D.2 Initial Step Ablation

Following the previous experiment, we seek to find the main variables to produce valid yet sparse CEs. The other variable of interest is the initialization step  $\tau$ . On the one hand, a higher  $\tau$  opens more opportunities to modify the image. But on the other hand, this increased generation power can be detrimental to the resulting image quality. We report the results in Table 5. The hyperparameter  $\tau$  has a similar effect to the gradient scale  $\lambda_c$ ; we observe a negative correlation

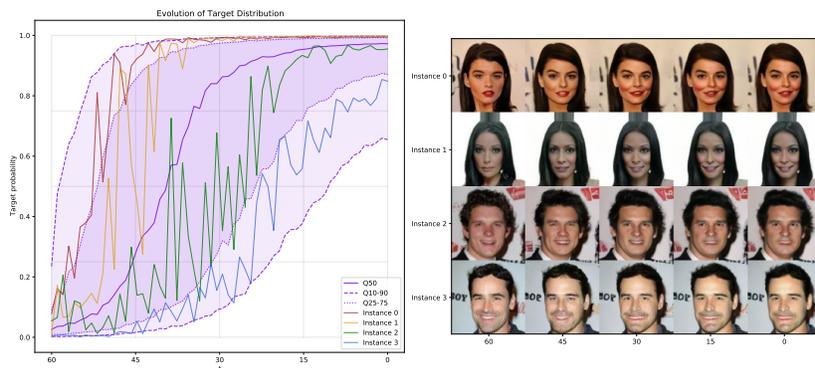


Fig. 6: **Target distribution at each time-step.** We visualize the evolution of the target labels’ probability. Each purple line represents a quantile of the probabilities. The color curves are cases shown on the right. In expectancy, the clean image at each time step increases. Nonetheless, typically the curves are sporadic. Yet, we observe an increase.

between  $\tau$  and the FID, and a positive one between  $\tau$  and the FR. The image generation has more optimization steps when increasing the initial noise level. Thus, it easily reaches a counterfactual that fools the classifier at the cost of decreasing the CE sparsity, an unwanted effect in the CE community. Similarly, a low  $\tau$  increases the sparsity, but the CEs are not as successful. Choosing  $\tau = 60$  finds an optimal equilibrium between both factors.

### D.3 Distribution Overtime

Our pipeline uses the unconditional DDPM to enable the use of the classifier under observation. At each step, the classifier uses the generated image to compute the gradient with respect to the target label. Therefore, this image gives information on the optimization process at each time step. Hence, in this experiment, we explore the behavior of these images at each stage of the denoising process. We plot the probability of the target class given by the inspected architecture at each step to accomplish this.

In Fig. 6 we visualize the evolution of the target labels’ probability over time, along with some examples. We see that the probability increases overtime on average. Nevertheless, the example instances show sporadic and non-steady development. Yet, we still observe an ascending behavior. Near the first steps, we see the most unstable conduct. Nonetheless, the optimization begins to settle when reaching a time steps near 0 (approx. at  $t = 20$ ). We attribute this behavior to an averaging effect along time; when the image generation reaches the final steps, the variance nearly vanishes. Hence, the unconditional generation does not vary much, reaching an equilibrium. This observation relates to the comments of Equation 5 in the main text, where we argue for using a single realization of  $x_t$  at each time step. As mentioned in the main text, the absence of averaging

at every step is partly mitigated in terms of the optimization objective by an averaging effect over time. But thanks to the randomness inherited from the early steps ( $t \approx \tau$ ), the overall CE creation process still displays some diversity.

## E Qualitative Results

In this section, we visualize some qualitative results from our proposed benchmark for counterfactual explanations. We include cases for the smile, young and other attributes from the CelebA dataset. Also, we compare our results with DiVE’s explanations in Figures 7 to 16. Further, we show some examples of the evolution at each time step of the noisy and clean instances in Figures 17 and 18. Finally, we visualize more examples on the variability of DiME in Figures 19 and 20.

In general, we see a clear pattern comparing DiME and DiVE. DiME’s generated instances are closer to the query image than DiVE’s. Further, DiVE uses a VAE as the generative model, so their CEs are blurrier than ours.

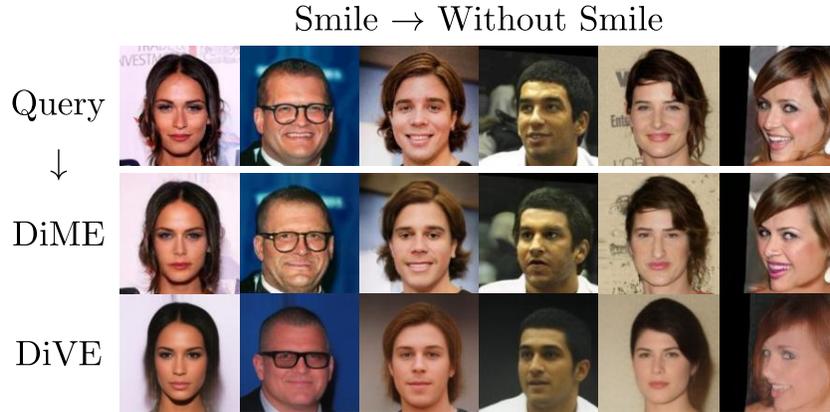


Fig. 7: Counterfactual explanations for the *Smile*. We visualize DiME and DiVE explanations targeting the label Smile.



Fig. 8: **Counterfactual explanations for the *Smile*.** We visualize DiME and DiVE explanations targeting the label No Bags under the eyes.



Fig. 9: **Counterfactual explanations for the *Age* attribute.** We visualize DiME and DiVE explanations targeting the label Young.



Fig. 10: **Counterfactual explanations for the *Age* attribute.** We visualize DiME and DiVE explanations targeting the label Old.

With respect to the gender attribute, we visualize two differences between each gender. For the *male to female* case, DiME exposes a clear correlation between the female label and the attributes *heavy makeup* and *lipstick*. We suspect that the classifier mainly relies on these attributes to classify an image as a woman. In contrast, DiVE adds “women-like” features to flip the prediction. For the *female to male* counterfactuals, major changes in the image are done to add female qualities for both models. The last two examples show that removing the makeup is enough to flip the classifier prediction.



Fig. 11: Counterfactual explanations from *male to female*. We visualize DiME and DiVE explanations.



Fig. 12: Counterfactual explanations from *female to male*. We visualize DiME and DiVE explanations.

Regarding the blurry attribute, at first glance, we see that DiVE’s VAE helps blur the input instance. Nevertheless, as we see in Fig. 13, CelebA’s inherited blurry attribute is different from the one produced by DiVE.

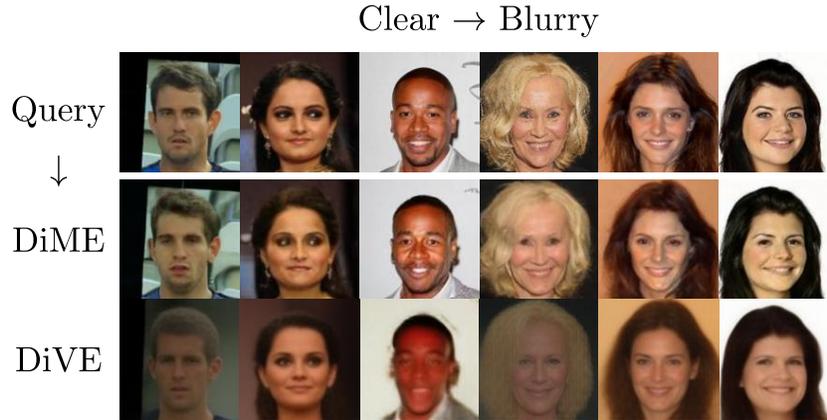


Fig. 13: Counterfactual explanations for the attribute *Blurry*. We visualize DiME and DiVE explanations using as target to deblur the input instance.



Fig. 14: Counterfactual explanations for the attribute *Blurry*. We visualize DiME and DiVE explanations using as target the blurry.

The attribute *Bags under the Eyes* has a clear and punctual location in the image: the region below the eyes. Both algorithms provide successful explanations when targeting this attribute. The main difference between DiME and DiVE performances is the capacity of DiME to retain as much fine-grained information as possible such as the hair, hands, and the background.



Fig. 15: Counterfactual explanations for the *Bag Under Eye*. We visualize DiME and DiVE explanations targeting the label Bags under the eyes.



Fig. 16: Counterfactual explanations for the *Bag Under Eye*. We visualize DiME and DiVE explanations targeting the label No Bags under the eyes.

Following the study on the evolution of the clean images  $x_t$  on time, we display more examples along with their noisy version. We see that, when  $t = 48$  and  $t = 36$ , the clean images present the most changes, while the last images do not vary much.

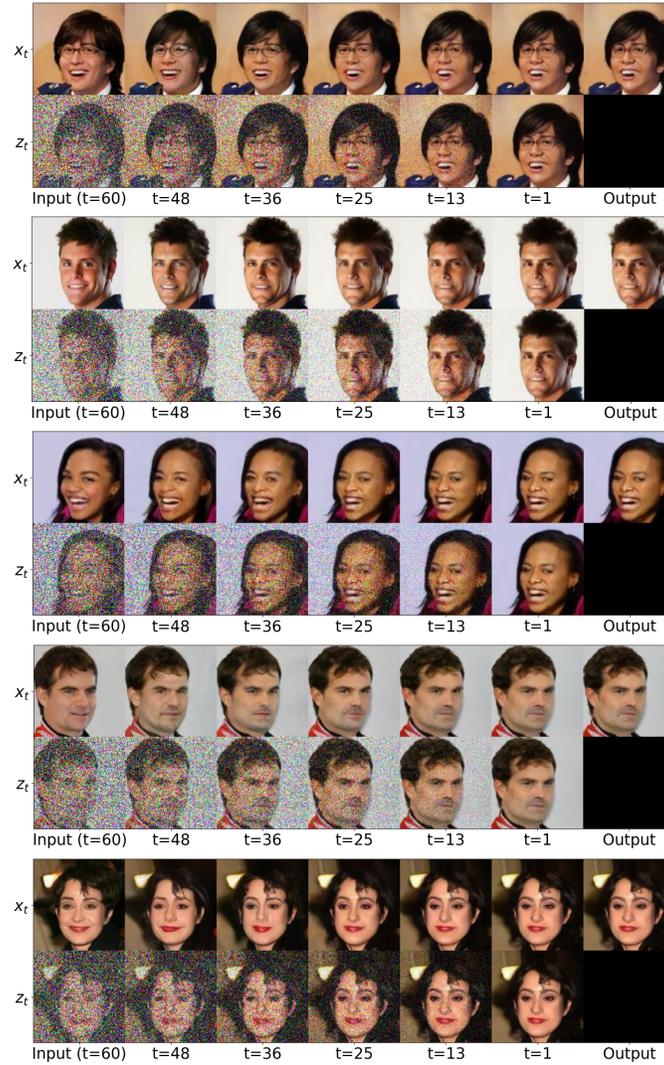


Fig. 17: **Visual inspection over  $t$ .** We visualize the evolution of the noisy  $z_t$  and clean instances  $x_t$  over time.

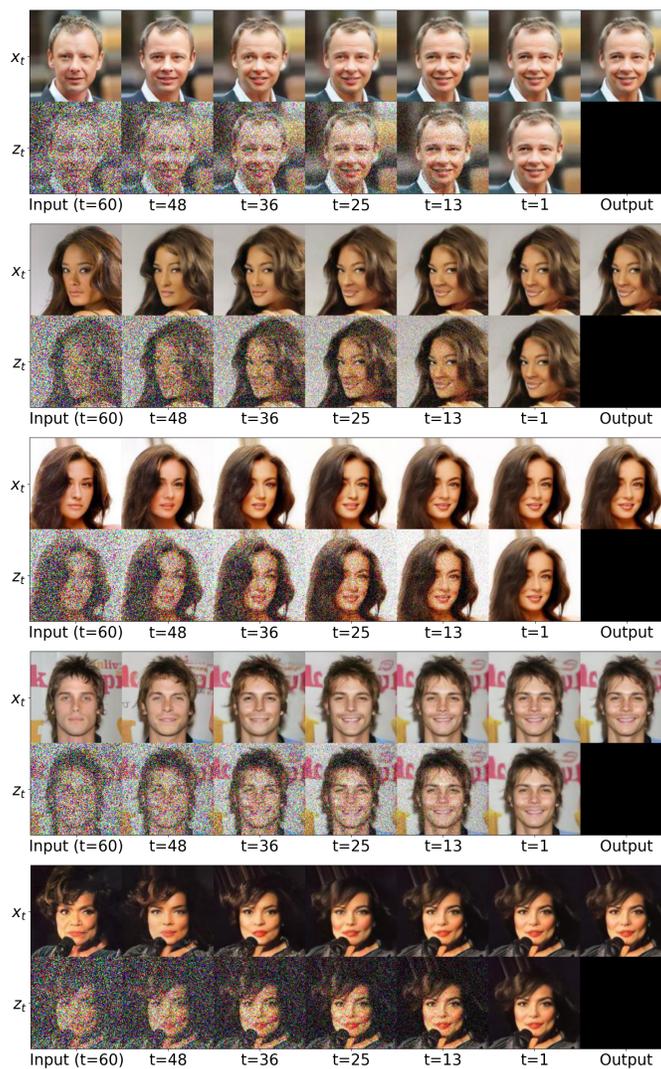


Fig. 18: **Visual inspection over  $t$ .** We visualize the evolution of the noisy  $z_t$  and clean instances  $x_t$  over time.

The following Figures visualize more examples of DiME’s capacity to create diverse counterfactual explanations. The visualizations show that DiME retains most details when generating counterfactuals.

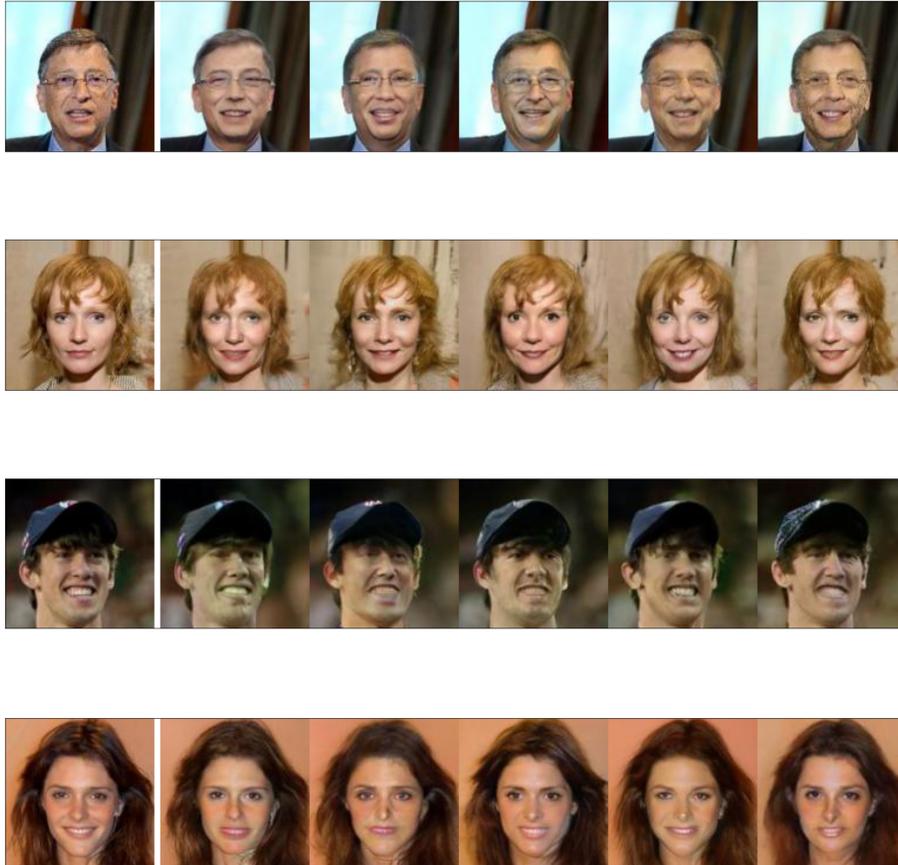


Fig. 19: **Variability examples.** We visualize the effects of the stochasticity of DiME to produce multiple explanations.

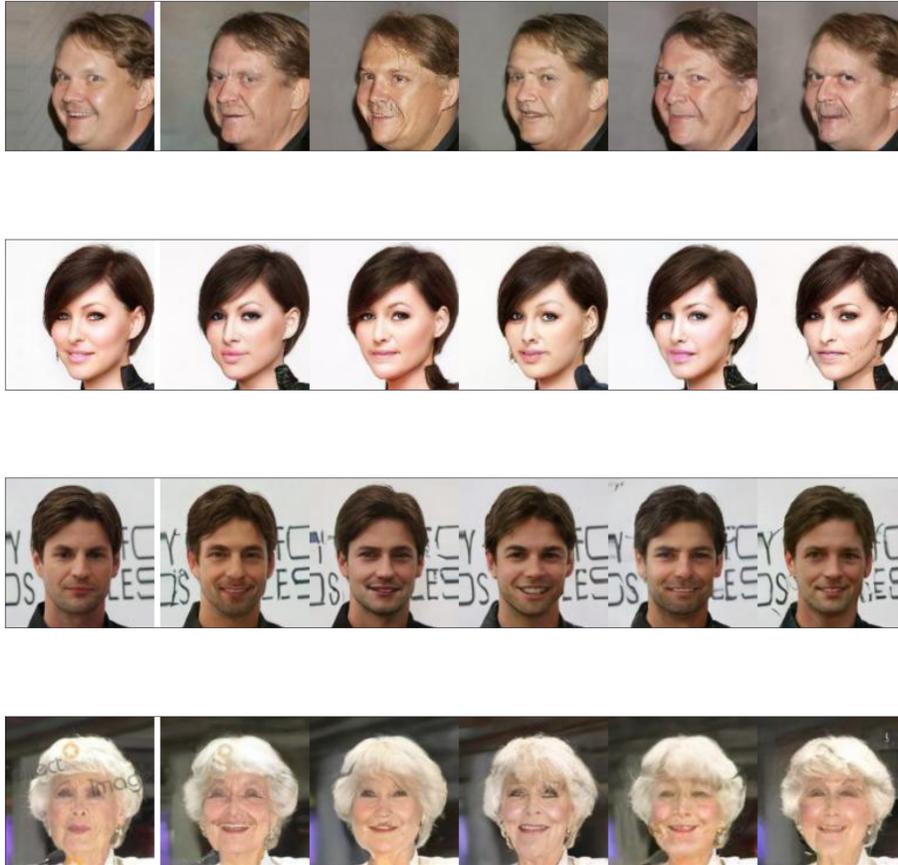


Fig. 20: **Variability examples.** We visualize the effects of the stochasticity of DiME to produce multiple explanations.