RETHINKING VISUAL COUNTERFACTUAL EXPLANATIONS THROUGH REGION CONSTRAINT

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual counterfactual explanations (VCEs) have recently gained immense popularity as a tool for clarifying the decision-making process of image classifiers. This trend is largely motivated by what these explanations promise to deliver – indicate semantically meaningful factors that change the classifier's decision. However, we argue that current state-of-the-art approaches lack a crucial component – the *region constraint* – whose absence prevents from drawing explicit conclusions, and may even lead to faulty reasoning due to phenomenons like confirmation bias. To address the issue of previous methods, which modify images in a very entangled and widely dispersed manner, we propose *region-constrained* VCEs (RVCEs), which assume that only a predefined image region can be modified to influence the model's prediction. To effectively sample from this subclass of VCEs, we propose *Region-Constrained Counterfactual Schrödinger Bridges* (RCSB), an adaptation of a tractable subclass of Schrödinger Bridges to the problem of conditional inpainting, where the conditioning signal originates from the classifier of interest. In addition to setting a new state-of-the-art by a large margin, we extend RCSB to allow for *exact* counterfactual reasoning, where the predefined region contains *only* the factor of interest, and incorporating the user to actively interact with the RVCE by predefining the regions manually.

1 INTRODUCTION

Visual counterfactual explanations (VCEs) aim at explaining the decision-making process of an image classifier by modifying the input image in a semantically meaningful and minimal way so that its decision changes. Over time, they have become an independent research direction with the latest methods presenting impressive and visually appealing results. Nevertheless, in this work we show that they possess a fundamental flaw at a conceptual level – the lack of *region constraint* and its proper utilization.

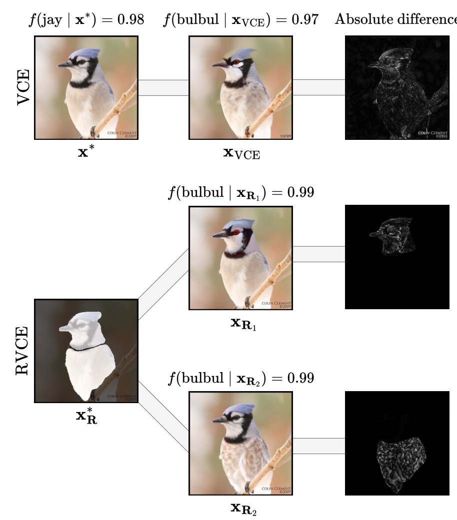

Figure 1: Previous methods create VCEs with unconstrained changes, making it virtually impossible to understand the decision-making process of a model. We propose *region-constrained* VCEs, establishing a new paradigm for comprehensible and actionable explanatory process.

Consider the image $\mathbf{x}^*$ in Fig. 1, which the classifier f correctly predicts to be a jay. In essence, VCEs focus on semantically editing $\mathbf{x}^*$ so that the prediction of f changes to some target class – bulbul in this case – hence providing an answer to a specific *what-if* question, through which the model's reasoning is explained. Consider now an example VCE for $\mathbf{x}^*$, denoted as $\mathbf{x}_{\text{VCE}}$, obtained with a recent state-of-the-art (SOTA) method. While $\mathbf{x}_{\text{VCE}}$ is successful at changing the prediction of f and can be considered both realistic and semantically close to $\mathbf{x}^*$, answering *why* f now predicts it as a bulbul is close to impossible. The algorithm simultaneously modifies the bird's head and feathers, changes the texture of the branch and even modifies the copyright caption. The entanglement and dispersion of introduced changes hence leaves the question unanswered. We argue that to circumvent these fundamental difficulties, VCEs should be synthesized with a hard constraint on the *region*, where the changes are allowed to appear, while leaving the rest of the image unchanged. For example, consider the image $\mathbf{x}_{\mathbf{R}}^*$ with regions of the bird's head ($\mathbf{R}_1$) and body ($\mathbf{R}_2$) overlayed. Constraining the VCEs to introduce changes *only* to predetermined regions leads to two distinct explanations, $\mathbf{x}_{\mathbf{R}_1}$ and $\mathbf{x}_{\mathbf{R}_2}$, of why the decision changes to bulbul. By isolating the modified factors, the explanatory process greatly simplifies – one can now state with certainty that f's new prediction is based either on the modified feathers ($\mathbf{x}_{\mathbf{R}_2}$) or the changed characteristics of its head ($\mathbf{x}_{\mathbf{R}_1}$). Region-constrained VCEs (RVCEs) allow, therefore, to reason about the model's thought process in a *causal* and principled manner, mitigating the potential *confirmation bias* and clarifying the explanatory process.

By putting RVCEs in the spotlight, our work establishes new frontiers in the field of VCE generation. First, we define the objective of finding RVCEs as solving a conditional inpainting task. By building on top of the Image-to-Image Schrödinger Bridge (I^2SB, Liu et al. (2023a)) approach and adapting it to the classifier guidance scheme, we develop an efficient algorithm which synthesizes RVCEs with extreme realism, sparsity and closeness to the original image. Specifically, we set a new quantitative state-of-the-art (SOTA) on ImageNet (Deng et al., 2009) with up to 4 times better scores in FID and 3 times better sFID (realism), up to 2 times higher COUT (sparsity), and match or exceed S^3 (similarity) and Flip Rate (efficiency) achieved by previous methods. Through large-scale experiments, we demonstrate that, besides a fully automated way of synthesizing meaningful and highly interpretable RVCEs, our approach, *Region-constrained Counterfactual Schrödinger Bridge* (RCSB), allows to infer causally about the model's change in prediction and enables the user to actively interact with the explanatory process by manually defining the region of interest. Moreover, our results highlight the importance of RVCEs in future research, indicating potential pitfalls of unconstrained methods that could lead to drawing misleading conclusions.

2 Background & Related Work

In this section, we introduce the necessary background knowledge connected with score-based generative models (SGMs) and I^2SB, which forms the foundation of our method. We then present an overview of recent methods for VCE generation based on SGMs. For an extended literature review and detailed description of the theoretical basis, please refer to the Appendix.

SGM. Following the work of Song et al. (2021), SGMs can be constructed through the framework of stochastic differential equations (SDEs), where samples from a complex distribution p_0 (*e.g.*, natural images) are mapped to a Gaussian distribution p_1, while the model is trained to reverse this mapping. Formally, converting data to noise is performed by following the *forward* SDE (Eq. (1a)), while denoising happens through the *reverse* SDE (Eq. (1b), Anderson (1982)):

$$\mathrm{d}\mathbf{x}_t = \mathbf{F}_t(\mathbf{x_t})\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\mathbf{w}, \tag{1a}$$

$$\mathrm{d}\mathbf{x}_t = (\mathbf{F}_t(\mathbf{x_t}) - \beta_t \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t))\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\bar{\mathbf{w}}, \tag{1b}$$

where $\mathbf{x}_t$ is the noisy version of a clean image $\mathbf{x} \in \mathbb{R}^n$ for some $n \in \mathbb{N}$ at timestep $t \in [0, 1]$, $\mathbf{w}$ and $\bar{\mathbf{w}}$ denote the Wiener process and its reversed (in time) counterpart, $\mathbf{F}_t(\mathbf{x}_t) : \mathbb{R}^n \to \mathbb{R}^n$ is the *drift* coefficient, $\beta_t \in \mathbb{R}$ is the *diffusion* coefficient and $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t)$ is the *score function*. An SGM $\mathbf{s}_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ denotes the model's parameters, is trained to approximate the score, *i.e.*, $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) \approx \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t)$. During sampling, denoising begins from pure noise $\mathbf{x}_1 \sim p_1$ and follows some discretized version of Eq. (1b) with the approximate score $\mathbf{s}_{\boldsymbol{\theta}}$.

SGMs can also be adapted to *conditional* generation, where $\mathbf{y}$ represents the conditioning variable. In this case, the score $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t)$ is replaced by $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t \mid \mathbf{y})$, which can be decomposed with Bayes' Theorem into $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t \mid \mathbf{y}) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \mathbf{x}_t, t)$. While $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t)$ can be approximated with an already trained $\mathbf{s}_{\boldsymbol{\theta}}$, $\nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \mathbf{x}_t, t)$ must be modeled additionally. For $\mathbf{y}$ representing class labels, $p(\mathbf{y} \mid \mathbf{x}_t, t)$ can be approximated with an auxiliary time-dependent classifier $p_{\boldsymbol{\phi}}(\mathbf{y} \mid \mathbf{x}_t, t)$ trained on noisy images $\{\mathbf{x}_t\}_{t \in [0,1]}$. Incorporating $p_{\boldsymbol{\phi}}$ into the sampling process is termed as *classifier guidance* (CG), and can be strengthened (or weakened) with *guidance scale* s through $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t) + s \cdot \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t \mid \mathbf{y})$. Therefore, class-conditional sampling in SGMs amounts to additionally maximizing the likelihood $p_{\boldsymbol{\phi}}(y \mid \mathbf{x}_t, t)$ of the classifier throughout the generative process to arrive at images from the data manifold, which resemble (according to $p_{\boldsymbol{\phi}}$) instances of a specific class. We emphasize this fact here for further reference.

I^{2}SB. The framework of I^{2}SB extends SGMs to p_1 representing an *arbitrary* data distribution. For training, I^{2}SB requires paired data, *e.g.*, in the form of clean and partially masked samples for inpainting, where it learns to infill the missing parts. While SGMs can also be adapted to solve inverse problems like inpainting, I^{2}SB maps these samples *directly* (see Fig. 2 for a comparison of their generative trajectories). Therefore, I^{2}SB follows the same theoretical paradigm, where sampling is achieved by discretizing Eq. (1b) and using a score approximator $\mathbf{s}_\psi$, but the generative process begins from a corrupted (*e.g.*, masked) image instead of pure noise. Hence, I^{2}SB can also be adapted to conditional generation in the same manner as SGMs, especially for class-conditioning with an auxiliary classifier. Importantly, a special case of I^{2}SB follows an optimal transport ordinary differential equation (OT-ODE) when $\beta_t \to 0$, eliminating stochasticity beyond the initial sampling step (see Appendix). We utilize the OT-ODE version of I^{2}SB in our implementation.

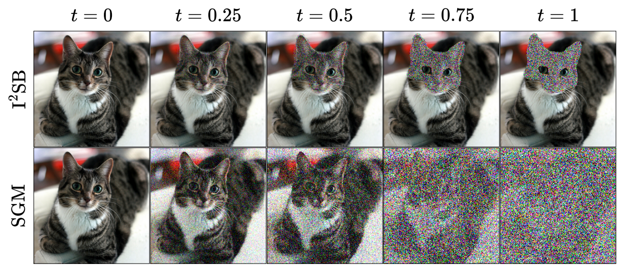

Figure 2: Generative trajectories of I^{2}SB and SGM. Intermediate images of I^{2}SB are much closer to the data manifold.

SGM-based VCEs. The initial approach of adapting SGMs to VCE generation, DiME (Jeanneret et al., 2022), obtains the classifier's gradient by mapping the noised image to its clean version at each step through the reverse process. Augustin et al. (2022) incorporate the gradient of a robust classifier and a cone projection scheme. Jeanneret et al. (2023) decompose the VCE generation into pre-explanation construction and refinement using RePaint (Lugmayr et al., 2022). Jeanneret et al. (2024) utilize a foundation model, Stable Diffusion (SD, Rombach et al. (2022)), to generate VCEs in a black-box scenario. Farid et al. (2023) and Motzkus et al. (2024) utilize Latent Diffusion Models (LDMs), including SD, in a white-box context. Weng et al. (2024) propose FastDiME to accelerate the generation process in a shortcut learning scenario. Also in black-box context, Sobieski & Biecek (2024) utilize a Diffusion Autoencoder (Preechakul et al., 2022) to find semantic latent directions that globally flip the classifier's decision. Finally, Augustin et al. (2024) also make use of SD in various contexts, including classifier disagreement and neuron activation besides VCEs.

3 Method

In this section, we describe the details of our approach, beginning with the formulation of RVCEs as solutions to conditional inpainting task. Next, we motivate the use of I^{2}SB as an effective prior for synthesizing meaningful RVCEs and follow with a series of steps that better align the gradients of a standard classifier w.r.t corrupted images from its generative trajectory. We conclude with a description of the automated region extraction method, forming the basis of our algorithm.

RVCEs through conditional inpainting. We define the problem of finding RVCEs for the classifier f from a given image $\mathbf{x}^*$, a region $\mathbf{R}$ and *target class* label y, where $\arg\max_{y'} f(y' \mid \mathbf{x}^*) \neq y$, as the task of sampling from

$$p(\mathbf{x} \mid \arg\max_{y'} f(y' \mid \mathbf{x}) = y, (\mathbf{1} - \mathbf{R}) \odot \mathbf{x} = (\mathbf{1} - \mathbf{R}) \odot \mathbf{x}^*), \tag{2}$$

where $\mathbf{R}$ is a binary mask with 1 indicating the region. Intuitively, sampling from Eq. (2) means obtaining $\mathbf{x}$ with the complement of $\mathbf{R}$ unchanged and the content of $\mathbf{R}$ modified in a way that changes the decision of f to y, *i.e.*, performing inpainting with additional condition coming from the classifier f.

Synthesizing meaningful RVCEs. Looking at Eq. (2), one quickly realizes that obtaining semantically meaningful RVCEs requires maximizing the likelihood $f(y \mid \mathbf{x})$ of the classifier while inpainting $\mathbf{R}$ with content that keeps $\mathbf{x}$ in the data manifold. These conditions greatly resemble the CG scheme in the context of I^{2}SB, since the score estimate $\boldsymbol{s}_\psi$ serves as an effective prior for generating in-manifold infills, while the likelihood $p_\phi(y \mid \mathbf{x})$ of an auxiliary classifier is maximized to ensure that p_ϕ predicts them as instances of y. Moreover, I^{2}SB maps masked images *directly* to clean samples, leaving the content outside $\mathbf{R}$ unchanged in the final image.

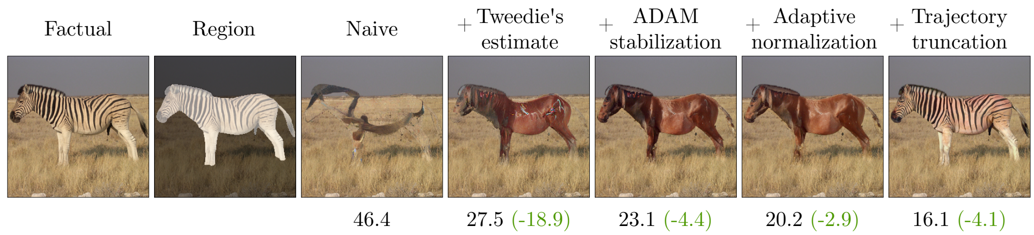

Figure 3: Series of proposed improvements to better align the gradient's of the classifier of interest with the generative trajectory. Changes to the *factual* image are constrained to the indicated *region*. Subsequent images illustrate the influence of each new adaptation. Numbers below images correspond to FID (↓) values obtained in a larger-scale experiment (for details, see Appendix).

The above arguments suggest that inserting f in place of p_{ϕ} should function as an effective mechanism for sampling meaningful RVCEs. However, a fundamental drawback of this *naive* approach is that, throughout the generative process, f's gradients originate from evaluating it on images with highly noised infills inside $\mathbf{R}$ (see Fig. 2). Such corrupted images are far from what f observed during training, hence leading to a *misalignment* of its gradients with the correct trajectory and generation of out-of-manifold samples. Similar issue has been identified by previously mentioned SGM-based methods for VCEs, which can be generally unified as attempts to replace the auxiliary classifier p_{ϕ} with f in the CG scheme in SGMs and *correct* f's gradients. Following Fig. 2, one should expect the misalignment in these methods to be of great extent, as the generative trajectory consists of highly noised images, leaving no meaningful content for f to provide accurate gradients. There, as shown in Fig. 2, I^{2}SB provides a crucial advantage, which stems from its generative trajectory being *much closer* to the data manifold. Moreover, by using I^{2}SB, f is able to effectively utilize the readily available context outside $\mathbf{R}$. Hence, in the following, we focus on reducing the misalignment problem caused by the noised content inside $\mathbf{R}$, in the end arriving at a highly effective algorithm for meaningful RVCEs.

Aligning the gradients. We propose to adapt the gradients of f to properly align with the generative trajectory of I^{2}SB through a series of incremental steps. To provide the intuition standing behind the introduction of each consecutive improvement, Fig. 3 provides an example RVCE task, where the factual image depicts a zebra correctly predicted by the model (ResNet50 (He et al., 2016)), and the goal is to change the decision to 'sorrel'. We set the region constraint to include the entire animal to make the task challenging enough and verify the improvements quantitatively through a large-scale experiment with around 2000 images. For each step, we compute FID between the RVCEs and original images to assess their realism. For details on the experimental setup, see Appendix.

Naive. We first verify that naively plugging f in place of p_{ϕ} does not provide meaningful results. Indeed, as shown in Fig. 3, the method struggles to include the information from f. The unrealistic infill also suggests that the classifier's signal negatively influences the score from I^{2}SB.

Tweedie's formula. To begin with closing the gap between the data manifold and the generative trajectory, we refer to a classic result of Tweedie's formula (Robbins, 1992; Chung et al., 2022), which states that a *denoised estimate* of the final image at step t can be achieved by computing the posterior expectation

$$\hat{\mathbf{x}}_0(\mathbf{x}_t) := \mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t] = \mathbf{x}_t + \sigma_t^2 \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t), \tag{3}$$

where $\sigma_t^2 = \int_0^t \beta_\tau \mathrm{d}\tau$. For visual differences between $\mathbf{x}_t$ and $\hat{\mathbf{x}}_0(\mathbf{x}_t)$, see Appendix. Crucially, one has access to approximate $\hat{\mathbf{x}}_0(\mathbf{x}_t)$ at every step t by utilizing I^{2}SB as the approximate score. Replacing $\nabla_{\mathbf{x}_t} \log f(y \mid \mathbf{x}_t)$ with $\nabla_{\mathbf{x}_t} \log f(y \mid \hat{\mathbf{x}}_0(\mathbf{x}_t))$ brings the inputs of f much closer to what it expects, improving the conditional inpainting process as indicated by Fig. 3, which now shows a structure resembling a sorrel and a much smaller FID.

ADAM stabilization. Despite utilizing the Tweedie's estimate, we observed the norms of f's gradients to have a very noisy tendency throughout the generation process, pointing out a possible cause for visible artifacts and the missing parts of the animal. Hence, we propose to smooth out the gradients by applying the ADAM update rule at each step (Vaeth et al., 2024; Kingma, 2019), to which we simply refer as ADAM stabilization. Figure 3 indicates that this modification allows for filling in the missing parts of the sorrel and further lowering FID.

Adaptive normalization. Incorporating ADAM stabilization required greatly lowering the guidance scale to values on the order of $1\mathrm{e}{-2}$, as using standard $s = 1$ led to extreme artifacts. This phenomenon suggested that the step size could also be adjusted throughout the generation process. While we initially experimented with various types of schedulers (see Appendix), using *adaptive normalization* has empirically proven to be the most effective approach. Specifically, at the beginning of the conditional inpainting process, we register the norm of the first encountered gradient of the log-likelihood of f. We then use it as a normalizing constant for each subsequent gradient, meaning that the generation begins with gradient of unit norm. This simple modification not only further lowered FID, but also reduced the final visible artifacts and improved color balance (Fig. 3).

Trajectory truncation. Up until this point, we relied solely on the ability of I^{2}SB and the classifier's signal to correctly infill the missing regions with semantically meaningful content, with no knowledge of the structure of the missing objects. Since a possible infill of the region is always available from the original image, one can begin the inpainting process from some *intermediate* step instead of the final one. This intervention allows for mixing the available information with the one coming from the classifier, and gives direct control over the preservation of the original content. As our approach does not bias the conditional score with signal from any additional losses (like Learned Perceptual Image Patch Similarity (LPIPS Zhang et al. (2018)) or l_2 in other works), we can fully rely on the *conceptual* compression of I^{2}SB, similarly to SGMs (Ho et al., 2020), which decomposes the generation process into initial phases responsible for the overall structure of objects and later ones responsible for small details. Figure 3 showcases the effect of using this *trajectory truncation* (τ) at the 0.4 level, meaning that the infilling process starts from $t = \tau \cdot T$, where T denotes the final timestep. Understandably, trajectory truncation greatly lowers the FID score, as much more information is available from the very beginning of the process, and introduces much more subtle changes to the image. We explore the effect of manipulating τ further in the Appendix, showing that it functions as a very interpretable mechanism for controlling the content preservation.

Automated region extraction. While the introduced algorithmical improvements effectively incorporate the classifier's signal into the inpainting process, they do not address the issue of predetermining the region for the resulting explanation. To this end, the optimal strategy would be fully automated and focus on regions that are both important to the classifier's prediction and point to semantically meaningful concepts. This description closely resembles the role of visual attribution methods, which assign importance values to pixels based on their relevance to the model's output (Holzinger et al., 2022). Figure 4 shows an example attribution map obtained with Integrated Gradients (IG, Sundararajan et al. (2017)) method for the squirrel prediction of a ResNet50 model. Perceptually, highest attributions are focused around the squirrel's head. To extract a region from such attributions, one can threshold them to cover a specific fraction a of the total image area. However, after binarizing the attributions with $a = 0.05$, we observe that the resulting region is highly scattered, losing focus from semantic concepts. To address this issue, we divide the image into a grid of square cells of size $c \times c$, where each cell receives the value equal to the sum of the absolute pixel attributions inside it. Figure 4 shows that this postprocessing mechanism (here with $c = 16$) greatly amplifies the focus of the resulting map. By thresholding it with $a = 0.05$, we observe the extracted region to focus solely on the squirrel's head. This leads to a fully automated strategy for obtaining regions that are both aligned with semantically meaningful concepts and based on pixels that are important for the classifier.

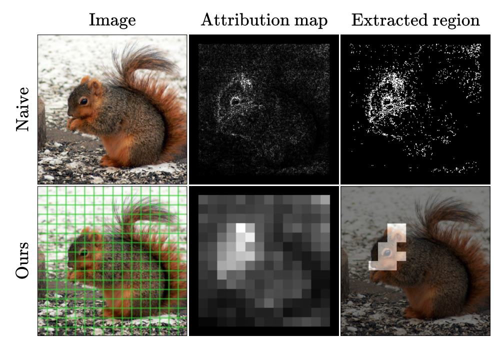

Figure 4: Example region obtained with our automated region extraction. Instead of directly binarizing an attribution map (upper row), we amplify the focus on semantic concepts (bottom row) with a simple approach based on grid cells.

We term the final version of the algorithm which combines all of the aforementioned improvements with the automated region extraction as RCSB. For the pseudocode of the entire procedure, see Appendix. We include our implementation in the Supplementary Material.

4 Experiments

Following previous works for VCEs on ImageNet, we base the quantitative evaluation on 3 challenging **main** VCE generation tasks: **Zebra – Sorrel**, **Cheetah – Cougar**, **Egyptian Cat – Persian Cat**, where each task requires creating VCEs for images from both classes and flipping the decision to their counterparts. We treat it as a general benchmark for evaluating the effectiveness of RCSB in various scenarios. We use FID (↓) and sFID (↓) to assess realism (Heusel et al., 2017), S^3 (↑) for representation similarity (Chen & He, 2021), COUT $\in [-1, 1]$(↑) (Khorram & Fuxin, 2022) for sparsity and Flip Rate (FR) (↑) for efficiency. For qualitative examples, we extend the main tasks with a large array of **other** tasks, which we show throughout the paper and the Appendix, where more details regarding the experimental setup and the metrics description can be found.

RCSB sets new SOTA for VCEs. We first verify that synthesizing RVCEs with RCSB leads to new SOTA in VCE generation. Table 1 quantitatively compares RCSB with recent SOTA approaches to VCEs on ImageNet. Our RVCEs are much more realistic (at least $2-4\times$ decrease in FID and sFID), stay close to original images (match or exceed best values of S^3) and almost always flip the model's decision (FR $\approx$ 1.0). RCSB also solves a long-standing challenge of achieving extremely *sparse* explanations on ImageNet, especially on **Zebra – Sorrel** task. While all other methods fail to achieve nonnegative values, RCSB approaches the upper bound of COUT. Our method is clearly the most balanced, as it does not struggle on any specific metric like, *e.g.*, DVCE on S^3. In the Appendix, we show that it is also the most computationally efficient.

Method	FID	sFID	S^3	COUT	FR
			Zebra – Sorrel		
ACE l_1	84.5	122.7	**0.92**	−0.45	47.0
ACE l_2	67.7	98.4	$\underline{0.90}$	−0.25	81.0
LDCE-cls	84.2	107.2	0.78	−0.06	88.0
LDCE-txt	82.4	107.2	0.71	−0.21	81.0
DVCE	33.1	43.9	0.62	−0.21	57.8
RCSB^C	13.0	20.4	0.82	0.70	**99.7**
RCSB^B	$\underline{9.51}$	$\underline{17.4}$	0.86	$\underline{0.72}$	97.4
RCSB^A	**8.0**	**16.2**	0.88	**0.74**	$\underline{94.7}$
			Cheetah – Cougar		
ACE l_1	70.2	100.5	$\underline{0.91}$	0.02	77.0
ACE l_2	74.1	102.5	0.88	0.12	95.0
LDCE-cls	71.0	91.8	0.62	0.51	$\underline{100.0}$
LDCE-txt	91.2	117.0	0.59	0.34	98.0
DVCE	46.9	54.1	0.70	0.49	99.0
RCSB^C	30.2	39.2	0.87	0.79	$\underline{100.0}$
RCSB^B	$\underline{23.4}$	$\underline{32.4}$	0.90	$\underline{0.85}$	99.9
RCSB^A	**17.2**	**26.6**	**0.92**	**0.92**	**100.0**
			Egyptian Cat – Persian Cat		
ACE l_1	93.6	156.7	$\underline{0.85}$	0.25	85.0
ACE l_2	107.3	160.4	0.78	0.34	97.0
LDCE-cls	102.7	140.7	0.63	0.52	99.0
LDCE-txt	121.7	162.4	0.61	0.56	99.0
DVCE	46.6	59.2	0.59	0.60	98.5
RCSB^C	41.1	56.3	0.79	0.82	$\underline{100.0}$
RCSB^B	$\underline{31.3}$	$\underline{48.1}$	0.84	$\underline{0.87}$	$\underline{100.0}$
RCSB^A	**23.0**	**40.0**	**0.87**	**0.92**	**100.0**

Table 1: Quantitative comparison with SOTA. RCSB outperforms previous methods by a large margin across all metrics. The best results are obtained with $A(a = 0.1, c = 4, s = 3, \tau = 0.6)$, but the superiority is clear for various configurations, including $B(a = 0.2, c = 4, s = 1.5, \tau = 0.6)$, $C(a = 0.3, c = 4, s = 1.5, \tau = 0.6)$.

Figure 5 shows example explanations obtained with RCSB, greatly highlighting the importance of synthesizing RVCEs instead of standard VCEs. Our region extraction approach is able to precisely localize semantic concepts responsible for the model's decision. For example, in the Guacamole → Cabbage task, RCSB detects the guacamole bowl in the background and, guided by the classifier, infills it with cabbage while leaving the rest of the image unchanged. RCSB is capable of performing a wide range of editing tasks with various levels of difficulty, beginning with textural and color-based edits (*e.g.*, Tench → Goldfish, Mashed Potato → Cauliflower) to partially changing the object's structure (*e.g.*, Limpkin → Flamingo) to infilling the region with new, realistically looking concepts (*e.g.*, Cougar → Lynx, Green Mamba → Indian Cobra, Cougar → Lynx). Most importantly, thanks to the region constraint, our RVCEs allow for greatly limiting the potential factors that influenced the model's decision, making the explanations much more interpretable.

RCSB allows for causal inference about the model's reasoning. Drawing definite conclusions about the model's reasoning from an unconstrained VCE is not possible, as one cannot be certain that modifying potentially irrelevant factors did not in fact influence the prediction. RVCEs overcome this limitation when constrained on the region connected with the sole factor of interest, *e.g.*, the body of an animal in a species prediction task. To adapt RCSB to such scenario, we replace the automated region extraction method with a foundation text-to-object-segmentation model [1]. Using

[1]Language Segment Anything (LangSAM) combines Segment Anything Model (Kirillov et al., 2023) with GroundingDINO (Liu et al., 2023b) to allow object segmentation from text prompts.

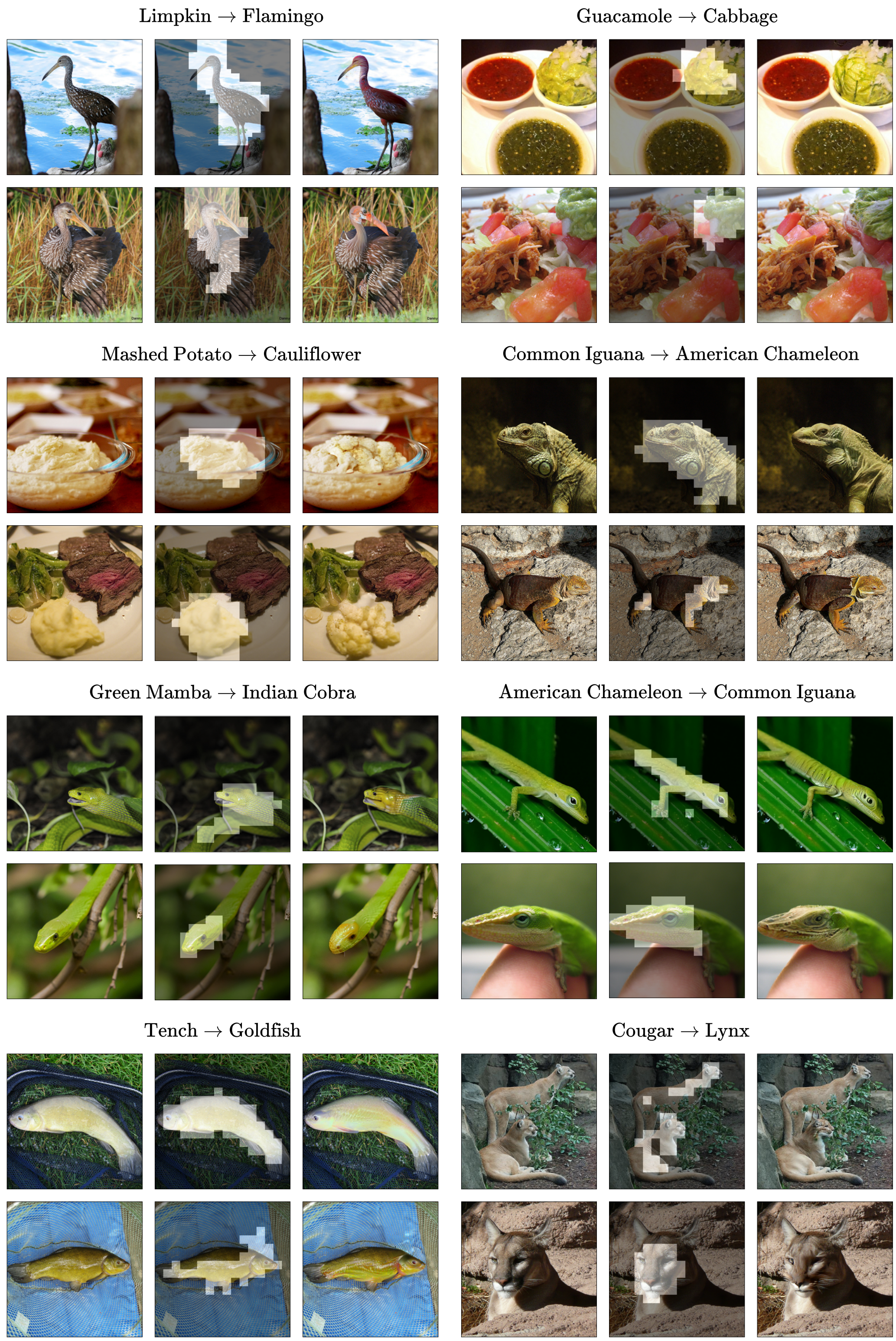

Figure 5: Qualitative examples obtained with RCSB using automated region extraction. Each task of the form *predicted class* → *target class* shows the factual image, the extracted region and the RVCE obtained with RCSB.

Metric	FID	sFID	S^3	COUT	FR	FID	sFID	S^3	COUT	FR	FID	sFID	S^3	COUT	FR
Task	**Zebra – Sorrel**					**Cheetah – Cougar**					**Egyptian Cat – Persian Cat**				
A	Exact regions obtained with LangSAM and prompts: zebra / horse, cheetah / cougar, cat respectively														
Values	32.8	41.5	0.87	0.74	98.9	37.2	50.6	0.91	0.84	99.4	52.0	82.8	0.81	0.84	99.2
B	Regions based on freeform masks with the area in the indicated range														
$10-20\%$	6.7	15.0	0.85	0.85	87.6	9.0	19.1	0.89	0.72	96.6	12.4	29.6	0.80	0.73	96.9
$20-30\%$	7.8	15.8	0.84	0.53	92.2	11.6	21.3	0.88	0.71	99.6	17.7	34.0	0.78	0.74	99.3
C	Ablation study with adaptations of other inpainting algorithms														
RePaint	63.8	76.0	0.55	0.77	99.3	129.3	144.2	0.50	0.77	99.0	148.7	175.2	0.38	0.76	99.5
MCG	43.2	55.6	0.73	0.45	96.0	76.6	91.4	0.74	0.64	100.0	93.7	117.5	0.62	0.65	99.9
DDRM	42.5	49.4	0.69	0.72	99.6	60.5	68.4	0.72	0.76	100.0	59.2	73.0	0.63	0.76	100.0

Table 2: Quantitative results from various experiments. **A**: regions extracted from LangSAM with text prompt connected to the initial class name. **B**: regions based on freeform masks that cover the fraction of the total area from the indicated range. **C**: automatically extracted regions used with adaptations of other inpainting algorithms.

the class name from a given task as the text prompt allows us to obtain highly precise segmentation masks of the relevant objects, enabling the identification of the cause behind the model's prediction change based solely on factors related to the object of interest.

We first quantitatively assess that RCSB is capable of utilizing regions provided by a generic object detector at scale. Table 2(A) shows the results of this evaluation together with the used text prompt. Here, the metrics are computed by first discarding images with a mask that covers area larger than 40%. Despite I^2SB being trained on masks covering at most 30% of the image area, we observed that it generalizes well beyond this threshold with 40% starting to pose a challenge. Crucially, despite the regions being classifier-agnostic and hence not necessarily focused on the most influential pixels, Table 2(A) indicates that RCSB is versatile enough to maintain most of the performance from the automated approach. The efficiency, sparsity and representation similarity of the obtained RVCEs remain very close to the values achieved by the closest configuration (in terms of hyperparameters) from Table 1, as the region area is often close to or exceeds 30%. The slight increase in FID and sFID stems mainly from the regions covering complex objects, whose modification may naturally move RVCEs further from original data at a distribution level, and a lower number of images used for these metrics' computation (as both are sensitive to sample size) due to the rejection of samples from the area constraint.

Regions that contain *exactly* the objects of interest provide novel insights about the model's reasoning. For example, consider the Lemon $\rightarrow$ Orange task from Fig. 6, where the lemons were correctly identified by the ResNet50 model. One would require the VCE for this task to indicate the sole determining factor of 'why lemons and not oranges'. However, with unconstrained VCEs, this identification process quickly becomes incomprehensible due to small changes added to each object in the image, such as other fruits. By constraining VCEs to the region occupied by the lemons, the reasoning process can be disentangled and simplified, as one can now look for this factor in the modifications of the lemons only. In this case, RCSB allows for increasing trust in the model, as making the lemons more orange correctly modifies its decision.

RVCEs also allow for clarifying the model's decision-making when its reasoning is not initially understandable. In the Volcano $\rightarrow$ Seashore task, the image shows both objects, while the model predicts it as the former. Applying RCSB to the exact region of the seashore results in a RVCE that changes the model's decision when the water's color becomes more light blue and structures like stones start to appear. Hence, one is able to better understand what the model *actually* identifies as a seashore. In other examples, the method introduces class-specific characteristics when the changes are constrained *precisely* and *exclusively* to the object of interest, ensuring the receiver about the general cause of the model's decision change. Such cases are also especially relevant when the generative model used to synthesize explanations is prone to systematic errors like, *e.g.*, SGMs struggling with correctly generating hands. In the Night Snake $\rightarrow$ Kingsnake task, this error can be bypassed with the region constraint by not allowing the generative model to affect anything other than the animal, hence alleviating the evaluation of the classifier on out-of-manifold samples.

Discovering complex patterns with interactive RVCEs. Despite the impressive capabilities of deep models in object localization, the receiver of the explanation may be interested in testing the

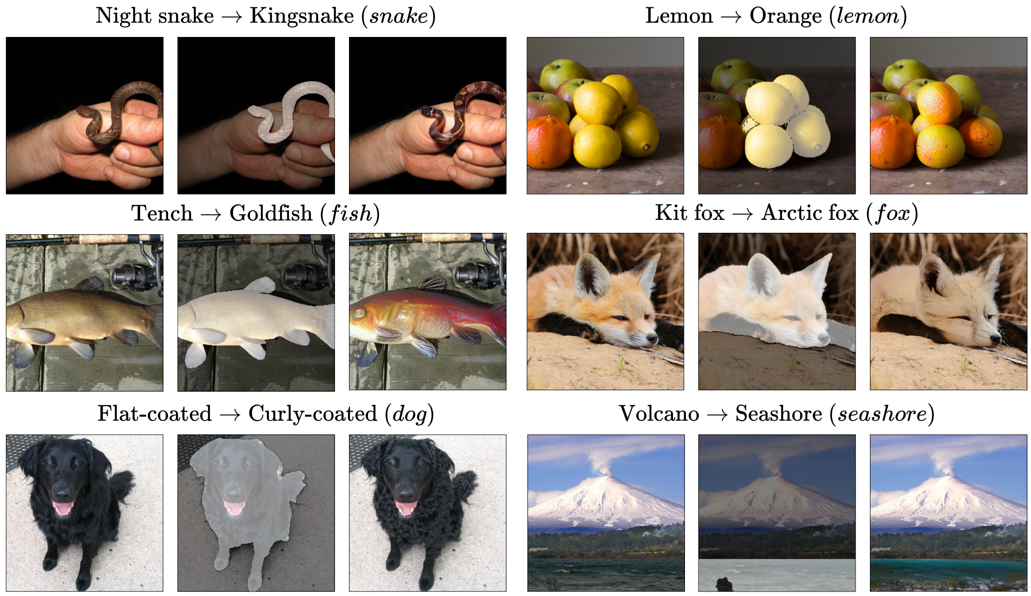

Figure 6: Qualitative examples obtained with RCSB using *exact* regions extracted from LangSAM using text prompt of the predicted class. For each task of the form *predicted class* → *target class*, a factual image together with the used region and the resulting RVCE are shown. The used text prompts are *emphasized*.

model for highly abstract and complex concepts that cannot be localized automatically and must be provided manually by the user. We begin with verifying the capability of RCSB in generating RVCEs based on user-defined regions by simulating such scenario at scale. Specifically, we randomly match images from the main tasks with regions given by the $10\% - 20\%$ and $20\% - 30\%$ freeform masks from the I^{2}SB training data (Saharia et al., 2022). We argue that this serves as a very challenging benchmark, since the algorithm's access to the most influential pixels (for the classifier) might often be very restricted.

Despite the task's difficulty, quantitative results from Table 2(B) highlight the versatility of RCSB, which is able to effectively utilize the restricted resources to influence the classifier's prediction. While S^3, COUT and FR are not significantly different from previous results, we observe a decrease in FID and sFID, indicating higher realism and closeness to the data distribution. This is largely due to the fact that freeform masks are often not connected to entire complex objects and do not contain the pixels most important to the classifier. Hence, RCSB may often leave large portions of the regions unchanged, which boosts the realism evaluation.

To allow for true interaction of the user with the explanatory process, we implement a simple interface that allows for manual image segmentation using a brush-like cursor. Figure 7 shows example results, where we manually predefine the regions on different images. This exploration gives important insights about the added value provided by RVCEs. In the Cat → Tiger task, we discover that the classifier's decision can be flipped by independently modifying either the cat's paws or snout, in both cases introducing a tiger's coloration. Similarly in the Arctic Fox → Red Fox task, choosing either the ears and muzzle or paws and stomach area allows for changing the model's decision with the features of a red fox. User-defined regions also allow to discover unusual reasoning patterns of the model. In the Cucumber → Zucchini task, the model's decision can be influenced by modifying only one of the cucumbers to zucchini, leaving the other unchanged. This observation connects with recent positions on the topic of contextual and spatial understanding of predictive models (Tomaszewska & Biecek, 2024), providing new rationale in further exploring how image classifiers *actually* reason.

Ablating RCSB's components. We empirically verified that combining our novel guidance mechanism with the I^{2}SB prior leads to highly effective RVCEs. To better understand the benefits provided by each component of our framework, we perform an ablation study, where we adapt the proposed improvements to SGM-based inpainters, aiming to assess the influence of the guidance scheme and I^{2}SB in isolation. Specifically, we pick RePaint (Lugmayr et al., 2022), one of the first adaptations of SGMs to inpainting, MCG (Chung et al., 2022) and DDRM (Kawar et al., 2022), two different adaptations of SGMs to linear inverse problems, which also include inpainting. We manually tune

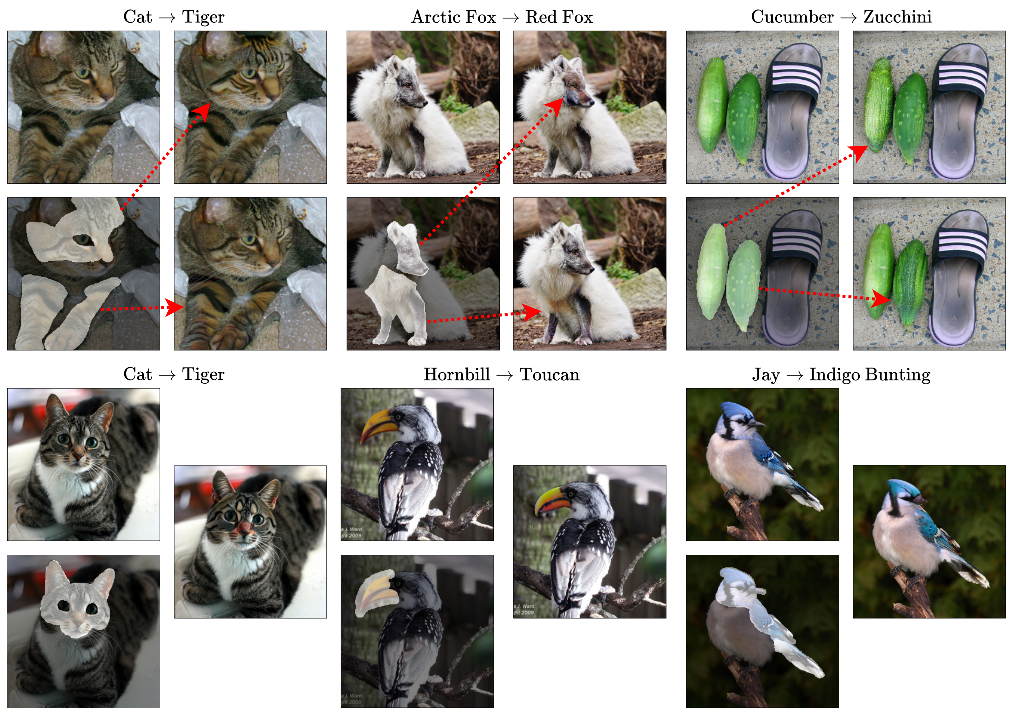

Figure 7: Qualitative examples obtained with RCSB from *user-defined* regions. For each task of the form *predicted class* → *target class*, a factual image together with the provided regions are shown. Arrows point to RVCEs obtained by modifying only the indicated region.

our guidance scheme to each method on a small subset of images and repeat the same evaluation protocol with the automated region extraction method (see Appendix for details of each adaptation). As these methods are much less compute-efficient, we cap their computational budget on each task to 24 A100 GPU hours.

Table 2(C) shows the results of the ablation study. Despite the fact that the used methods were never explicitly trained for inpainting, combining them with our guidance mechanism and region extraction allows for matching or even exceeding previous SOTA. For example, all adaptations achieve very high sparsity, almost always flip the classifier's decision and keep the explanation close to the original. This indicates the benefits of utilizing only the pixels from the extracted region and a proper utilization of the classifier's gradients without biasing them with additional components like LPIPS or l_2 loss. RCSB differentiates itself from the adaptations with a much higher realism of the obtained RVCEs (significantly lower FID and sFID), more balanced results and much smaller computational burden, *e.g.* $24\times$ less NFEs than RePaint. These benefits stem from the I^2SB prior, which is trained to map corrupted images directly to clean samples and the resulting trajectory being much closer to the data manifold, allowing the classifier to more effectively influence the inpainting process.

For extended quantitative and qualitative results, including RVCEs obtained for another 5 non-robust, 2 robust and a zero-shot CLIP classifier (Radford et al., 2021), and evaluation of 10 other attribution methods, we refer to the Appendix.

5 Conclusions

Our work advances the SOTA in VCE generation by constraining the explanations to differ from the factual image exclusively within a predetermined region. RCSB is not only very effective in sampling such explanations, proven by new quantitative records, but also showcases the novel capabilities for explaining image classifiers enabled by RVCEs. Specifically, to properly reason about the model's decision-making process, one must ensure that the potential confounding factors are limited to the greatest possible extent. RVCEs obtained with RCSB allow to do that in a wide range of scenarios, ranging from a fully automated approach to incorporating the user directly into the interactive explanation creation process. Our work establishes a new paradigm for explaining image classifiers in a much more principled manner, allowing the receiver to infer causally about the model's reasoning.

REFERENCES

Brian DO Anderson. *Reverse-time diffusion equation models*, volume 12. Stochastic Processes and their Applications, Elsevier, 1982.

Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. In *Advances in Neural Information Processing Systems*, 2022.

Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. Analyzing and explaining image classifiers via diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. In *PloS one*, 2015.

Przemyslaw Biecek and Wojciech Samek. Position: Explain to question not to justify. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 3996–4006, 21–27 Jul 2024.

Valentyn Boreiko, Maximilian Augustin, Francesco Croce, Philipp Berens, and Matthias Hein. Sparse visual counterfactual explanations in image space. In *DAGM German Conference on Pattern Recognition*, pp. 133–148. Springer, 2022.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2019.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems*, 2022.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023a.

Hyungjin Chung, Jeongsol Kim, and Jong Chul Ye. Direct diffusion bridge using data consistency for inverse problems. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library). In *https://github.com/MadryLab/robustness*, 2019.

Karim Farid, Simon Schrodi, Max Argus, and Thomas Brox. Latent diffusion counterfactual explanations. In *arXiv preprint arXiv:2310.06668*, 2023.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pp. 2376–2384. PMLR, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, 2020.

Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. Explainable AI methods-a brief overview. In *International workshop on extending explainable AI beyond deep models and classifiers*, pp. 13–38. Springer, 2022.

Paul Jacob, Éloi Zablocki, Hedi Ben-Younes, Mickaël Chen, Patrick Pérez, and Matthieu Cord. Steex: steering counterfactual explanations with semantics. In *European Conference on Computer Vision*, pp. 387–403. Springer, 2022.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision*, pp. 858–876, 2022.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16425–16435, 2023.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Text-to-image models for counterfactual explanations: a black-box approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4757–4767, 2024.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022.

Saeed Khorram and Li Fuxin. Cycle-consistent counterfactuals by latent transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10203–10212, 2022.

Ba Jimmy Kingma, Diederik P. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Andreas Kirsch. *An introduction to the mathematical theory of inverse problems*. Applied mathematical sciences. Springer, New York, 2nd ed edition, 2011.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. In *arXiv preprint arXiv:2009.07896*, 2020.

Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 693–702, 2021.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos Theodorou, Weili Nie, and Anima Anandkumar. I2SB: Image-to-Image Schrödinger Bridge. In *International Conference on Machine Learning*, pp. 22042–22062. PMLR, 2023a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *arXiv preprint arXiv:2303.05499*, 2023b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11461–11471, 2022.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Robert J. McCann. A convexity principle for interacting gases. In *Advances in Mathematics*, 1997.

Franz Motzkus, Christian Hellert, and Ute Schmid. Cola-dce–concept-guided latent diffusion counterfactual explanations. In *arXiv preprint arXiv:2406.01649*, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, 2019.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10619–10629, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Herbert E Robbins. An empirical Bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, 1992.

Pau Rodríguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam Laradji, Laurent Charlin, and David Vazquez. Beyond trivial counterfactual explanations with diverse valuable explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1056–1065, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Jonathan Scarlett, Reinhard Heckel, Miguel RD Rodrigues, Paul Hand, and Yonina C Eldar. Theoretical perspectives on deep learning methods in inverse problems. In *IEEE journal on selected areas in information theory*, 2022.

Erwin Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, pp. 269–310, 1932.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, 2022.

Lisa Schut, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In *International Conference on Artificial Intelligence and Statistics*, pp. 1756–1764. PMLR, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. In *International journal of computer vision*, 2020.

Sheng-Min Shih, Pin-Ju Tien, and Zohar Karnin. Ganmex: One-vs-one attributions using gan-based model explainability. In *International Conference on Machine Learning*, pp. 9592–9602. PMLR, 2021.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, Workshop Track Proceedings*, 2014.

Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. In *International Conference on Learning Representations*, 2020.

Bartlomiej Sobieski and Przemysław Biecek. Global counterfactual directions. In *European Conference on Computer Vision*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations, Workshop Track Proceedings*, 2015.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Jayaraman Thiagarajan, Vivek Sivaraman Narayanaswamy, Deepta Rajan, Jia Liang, Akshay Chaudhari, and Andreas Spanias. Designing counterfactual generators using deep model inversion. In *Advances in Neural Information Processing Systems*, 2021.

Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, and Yu-Gang Jiang. Deeper insights into the robustness of vits towards common corruptions. In *arXiv preprint arXiv:2204.12143*, 2022.

Paulina Tomaszewska and Przemysław Biecek. Position paper: Do not explain (vision models) without context. In *International Conference on Machine Learning*, 2024.

Philipp Vaeth, Alexander M Fruehwald, Benjamin Paassen, and Magda Gregorova. Gradcheck: Analyzing classifier guidance gradients for conditional diffusion sampling. In *arXiv preprint arXiv:2406.17399*, 2024.

Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 650–665. Springer, 2021.

Nina Weng, Paraskevas Pegios, Aasa Feragen, Eike Petersen, and Siavash Bigdeli. Fast diffusion-based counterfactuals for shortcut removal and generation. In *European Conference on Computer Vision*, 2024.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer, 2014.

Mehdi Zemni, Mickaël Chen, Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. Octet: Object-aware counterfactual explanations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15062–15071, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *International Conference on Learning Representations*, 2018.