Counterfactual Contrastive Analysis

Yunlong He and Pietro Gori*

LTCI, Télécom Paris, Institut Polytechnique de Paris, France
pietro.gori@telecom-paris.fr

Abstract. Visual Counterfactual Explanations (VCEs) aim to explain image classifiers by generating minimally edited and realistic versions of an input image that change the classifier's prediction. Existing VCE methods are inherently classifier-dependent and therefore susceptible to classifier biases and failure modes, such as sensitivity to shortcut features and calibration errors. In this paper, we propose a classifier-free approach for visual counterfactual generation based on Contrastive Analysis (CA). Given two datasets corresponding to different classes (e.g., healthy and patients), we disentangle the generative factors that are common across the two datasets from those that are salient to each dataset, and generate counterfactual images by swapping only the salient factors. By operating directly on data distributions rather than decision boundaries, our method provides model-agnostic VCEs that are less sensitive to classifier biases. Our approach leverages the high-quality synthesis and well-structured latent space of StyleGAN2. We use the feature space F, instead than the usual W-space, to improve detail preservation. Unlike conventional CA approaches, which typically assume salient factors in only one dataset, we introduce an adapted framework and loss functions for VCE that allow multiple salient factors in each dataset. We evaluate our method on three medical imaging datasets and demonstrate superior counterfactual generation quality compared to existing approaches.

Keywords: Visual Counterfactual Explanations · Contrastive Analysis

1 Introduction

Lately, the ability to explain decisions of black-box AI models has become critical, particularly in medical imaging. Visual counterfactual explanations (VCEs) [2, 9, 14, 15] are widely studied as a way to probe a classifier's behavior, seeking the smallest realistic changes to an input image that would alter its prediction.

VCEs for image classifiers are typically realized by generating a counterfactual (CF) image as an edited version of the original input. This task is often posed as a constrained generation problem with three objectives. First, the CF must be **valid**, meaning that it changes the classifier's prediction to the desired outcome. Second, the edited image should remain **realistic**, staying on the image manifold and avoiding unnecessary changes to irrelevant content. Third, the

* Corresponding author.

induced changes should be **interpretable** and **minimal** in a *semantic* sense, so that the explanation clearly indicates the key evidence responsible for the classifier's decision change. Most VCE methods instantiate these objectives within a generative framework, leveraging VAEs [27], GANs [20, 32], or diffusion models (DMs) [14–16, 35]. Among them, DMs currently achieve state-of-the-art performance for realistic CF image generation, typically guiding the generation process with classifier-driven losses to reach a label-flipping outcome, while penalizing deviations to preserve plausibility and similarity to the input. However, classifier guidance can induce changes aligned with the classifier biases and failure modes, such as sensitivity to shortcut features [35] and calibration errors, rather than realistic changes in the underlying data distribution, potentially reducing edit fidelity. In addition, these methods offer limited insight into which semantic factors drive the prediction flip, thereby limiting their interpretability.

Here, we introduce a **classifier-free** approach for visual counterfactual generation grounded in the *Contrastive Analysis* (CA) framework [1,7,10,22,23,34]. Instead of generating CFs based on a classifier's decision boundary, we operate at the data-distribution level, as in CA. Given two class-specific datasets, the proposed method aims to uncover the common and salient (i.e., class-specific) generative factors. CFs can then be generated by swapping only the salient factors, while preserving the common ones. Unlike classical VCE approaches, our method is a **generative editing framework** for producing CF examples. Most CA methods are based on deep generative models, such as VAEs [3,6,19,30,34,38] and GAN [7, 8]. While reconstruction can regularize the salient space by suppressing common information, it is insufficient on its own and typically requires additional constraints for clean latent separation [22,23,28,34]. A recent work [10] extended CA to modern GANs and diffusion models and reported that StyleGAN2 [17] achieves diffusion-level quality while enabling faster inference. In parallel, advances in StyleGAN editing have shown that translating edits from the W-space to the generator's intermediate feature space (F-space) improves fine-grained details [5]. Inspired by these findings, we propose a StyleGAN-based CA framework for VCEs on medical imaging datasets. Our contributions are:

1. A classifier-free generative framework grounded in CA that enables controllable CF generation via disentangled latent representations, yielding high-fidelity and semantically interpretable CF edits.
2. A CA-based framework, supporting both background-target and multi-salient assumptions, based on StyleGAN2's high-quality synthesis and well-structured F-space, instead of standard W-space, for better detail preservation.
3. Experiments on three medical imaging datasets demonstrate improved disentanglement performance over CA baselines and superior generation quality to prior state-of-the-art VCE approaches with faster image editing (0.25s/image).

2 Method

Problem Statement. In the CA setting, we consider two datasets $X = \{x_i\}_{i=1}^{N}$ and $Y = \{y_j\}_{j=1}^{M}$. Most prior work adopts a *background-target (BT)* assumption,

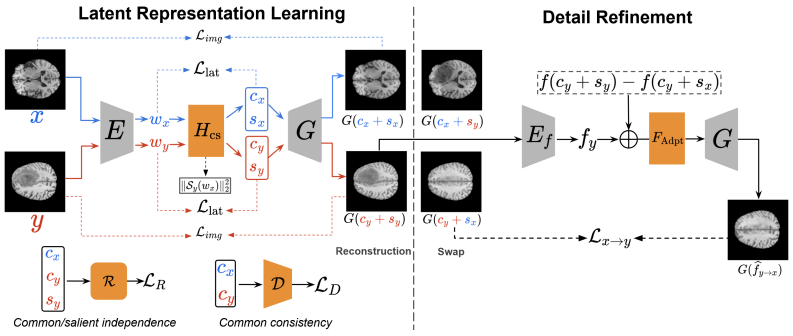

Fig. 1. Overview of the proposed CA-based framework for CF generations. Left: we first disentangle each input into a common c and a salient factor s, enabling reconstruction and CF editing by swapping the salient factors between images from X and Y. Right: we then refine the synthesized image by adjusting F-space features according to the feature shift $\Delta f_{y\to x} = f(c_y + s_y) - f(c_y + s_x)$.

where only one dataset (typically Y) contains a dataset-specific (or *salient*) pattern (e.g., a pathology) absent from the other, while both share common content (e.g., healthy anatomy). We generalize CA to a *multiple-salient (MS)* setting, where both X and Y contain their own *salient* patterns that are absent in the other. Our goal is to learn a *common* generative factor $c \in \mathbb{R}^{d_c}$ capturing information shared across X and Y, and *salient* generative factors $s_x \in \mathbb{R}^{d_{sx}}$, $s_y \in \mathbb{R}^{d_{sy}}$ that capture variations unique to X and Y, respectively.

Common and Salient (CS)-StyleGAN. In the context of StyleGAN, given samples $x \in X$ and $y \in Y$, an encoder E (i.e., pSp [26]) maps them to latent codes $w_x = E(x)$ and $w_y = E(y)$, where $w_x, w_y \in \mathbb{R}^{d_w}$. These codes lie in the extended StyleGAN latent space W^+, which is more regularized and editable than the generator feature space [5,26]. We introduce a CS separator $H_{\text{cs},\phi}$, which consists of one common branch $\mathcal{C}_{\phi_c}$ and two salient branches $\mathcal{S}_{x,\phi_{sx}}$ and $\mathcal{S}_{y,\phi_{sy}}$. For brevity, we omit the parameters ϕ in the following. We define the latent factors as: $c_x = \mathcal{C}(w_x)$ and $c_y = \mathcal{C}(w_y)$. In the MS setting, we set $s_x = \mathcal{S}_x(w_x)$ and $s_y = \mathcal{S}_y(w_y)$, while in the BT setting, we remove $\mathcal{S}_x$ and use $\mathcal{S}_y$ only (all $\mathcal{S}_x$-related terms are omitted).

Latent Regularization. We first train the H_{cs} separator with the following reconstruction loss in the StyleGAN2 latent space:

$$\mathcal{L}_{\text{lat}} = \|\mathcal{C}(w_x) + \mathcal{S}_x(w_x) - w_x\|_2^2 + \|\mathcal{C}(w_y) + \mathcal{S}_y(w_y) - w_y\|_2^2, \tag{1}$$

where each term encourages the sum of the common output and the corresponding salient output to reconstruct the StyleGAN2 latent (w_x or w_y). Following previous CA works [1, 7, 10, 22, 34], we combine common and salient factors by summation ($c + s$) rather than concatenation to keep the reconstructed code in the same W^+ space and dimensionality, avoiding extra mapping parameters. To enforce that the salient factors capture only dataset-specific information, we additionally penalize cross-dataset activations, by minimizing $\|\mathcal{S}_y(w_x)\|_2^2$ and

$\|\mathcal{S}_x(w_y)\|_2^2$ (we use only $\|\mathcal{S}_y(w_x)\|_2^2$ under the BT assumption, as explained in Fig. 1).

The loss in Eq. (1) is necessary but insufficient. In particular, it does not enforce (i) *common consistency*, i.e., that the common factors capture the same information across X and Y, nor (ii) *common-salient independence*, i.e., that common and salient factors do not share information. To promote *common consistency*, we impose distributional alignment of the common factors by introducing a discriminator $\mathcal{D}$ that predicts whether $\mathcal{C}(w)$ originates from X or Y, while the separator H_{cs} is trained to make this prediction impossible:

$$\mathcal{L}_D = \min_{H_{\text{cs}}} \max_{\mathcal{D}} \Big(- \mathbb{E}_{w_y}[\log \mathcal{D}(\mathcal{C}(w_y))] - \mathbb{E}_{w_x}[\log(1 - \mathcal{D}(\mathcal{C}(w_x)))]\Big). \tag{2}$$

Here, $\mathcal{D}$ is optimized to discriminate X from Y using the common codes, whereas H_{cs} is optimized adversarially so that $\mathcal{C}(w_x)$ and $\mathcal{C}(w_y)$ become indistinguishable.

To encourage ***common–salient statistical independence***, we introduce two regressors, $\mathcal{R}_x$ and $\mathcal{R}_y$, that attempt to predict the salient outputs $\mathcal{S}_x(w_x)$ and $\mathcal{S}_y(w_y)$ from the corresponding common outputs $\mathcal{C}(w_x)$ and $\mathcal{C}(w_y)$, respectively. If the regressors can accurately predict salient output from the common one, this indicates that $\mathcal{C}(w)$ still contains salient information and the separation is incomplete. We therefore train the separator adversarially against the regressors by minimizing the following objective:

$$\begin{aligned}\mathcal{L}_R = \min_{H_{\text{cs}}} \max_{\mathcal{R}} \Big(&- \mathbb{E}_{w_x}\big[\|\mathcal{R}_x(\mathcal{C}(w_x)) - \mathcal{S}_x(w_x)\|_2^2\big] - \mathbb{E}_{w_y}\big[\|\mathcal{R}_x(\mathcal{C}(w_y))\big] \\ &- \mathbb{E}_{w_y}\big[\|\mathcal{R}_y(\mathcal{C}(w_y)) - \mathcal{S}_y(w_y)\|_2^2\big] - \mathbb{E}_{w_x}\big[\|\mathcal{R}_y(\mathcal{C}(w_x))\big]\Big).\end{aligned} \tag{3}$$

This discourages $\mathcal{C}(w)$ from encoding salient patterns. We regress s from c because c is typically bigger (more information) and thus tends to absorb s.

Image-Space Losses. We adopt an image-space reconstruction loss combining pixel and perceptual terms. Given a real image i and its reconstruction $\widehat{i}$, we use $\mathcal{L}_{\text{rec}}(i, \hat{i}) = \|i - \hat{i}\|_2^2 + \|V(i) - V(\hat{i})\|_2^2$, where the first term is a pixel-wise ℓ_2 loss, and the second term is the LPIPS perceptual loss [37]. The function $V(\cdot)$ extracts perceptual features using a pretrained VGG16 network [31]. We use this Eq. to reconstruct both x and y, minimizing the image-space loss:

$$\mathcal{L}_{\text{img}} = \mathcal{L}_{\text{rec}}(x, G(c_x + s_x)) + \mathcal{L}_{\text{rec}}(y, G(c_y + s_y)), \tag{4}$$

where G is the pretrained StyleGAN2 generator. The final loss to train H_{cs} is:

$$\mathcal{L}_{H_{cs}} = \lambda_{lat}\mathcal{L}_{\text{lat}} + \lambda_D\mathcal{L}_D + \lambda_R\mathcal{L}_R + \lambda_{img}\mathcal{L}_{\text{img}} \tag{5}$$

F-space Refinement. After learning common and salient factors, we refine the generated images in the higher-dimensional generator feature space (F space), following [5]. This improves fine-grained detail preservation without changing the semantic edit direction learned in the W^+ space. Let $f(\cdot)$ denote the intermediate feature of the generator G at layer l=8, as in [5]. Given the learned latent factors, we define F-space feature shifts for the two swap directions as:

$$\Delta f_{x\to y} = f(c_x + s_x) - f(c_x + s_y); \;\; \Delta f_{y\to x} = f(c_y + s_y) - f(c_y + s_x). \tag{6}$$

Intuitively, $\Delta f_{x\to y}$ captures the change induced by replacing the salient factor of x with the one of y while keeping c_x fixed, and $\Delta f_{y\to x}$ is defined analogously. To condition the shifts on the source image, we extract features from reconstructions using a pretrained F-space encoder E_f: $f_x = E_f(G(c_x + s_x))$ and $f_y = E_f(G(c_y + s_y))$. As in [5], we concatenate f and Δf, and pass them to an adapter F_{Adpt}, yielding: $\widehat{f}_{x\to y} = F_{\text{Adpt}}(f_x \oplus \Delta f_{x\to y})$ and $\widehat{f}_{y\to x} = F_{\text{Adpt}}(f_y \oplus \Delta f_{y\to x})$, where $\oplus$ denotes concatenation. Decoding the adapted features with G yields the edited images. F_{Adpt} is trained with: $\mathcal{L}_{x\to y} = \mathcal{L}_{\text{rec}}(G(c_x + s_y),\ G(\widehat{f}_{x\to y}))$ and $\mathcal{L}_{y\to x} = \mathcal{L}_{\text{rec}}(G(c_y + s_x),\ G(\widehat{f}_{y\to x}))$. We also include a real-image reconstruction term to preserve performance on real inputs: $\mathcal{L}_{\text{real}} = \mathcal{L}_{\text{rec}}(x,\ \widehat{x}) + \mathcal{L}_{\text{rec}}(y,\ \widehat{y})$, where $\widehat{x} = G(F_{\text{Adpt}}(E_f(x)))$ and $\widehat{y} = G(F_{\text{Adpt}}(E_f(y)))$. Finally, we encourage realism with an adversarial loss $\mathcal{L}_{\text{adv}} = \mathcal{L}_{\text{adv-}x} + \mathcal{L}_{\text{adv-}y}$, where $\mathcal{L}_{\text{adv-}x}$ is applied to $G(\widehat{f}_{y\to x})$ against x, and $\mathcal{L}_{\text{adv-}y}$ to $G(\widehat{f}_{x\to y})$ against y. The overall refinement objective is: $\mathcal{L}_{\text{refine}} = \mathcal{L}_{x\to y} + \mathcal{L}_{y\to x} + \mathcal{L}_{\text{real}} + \lambda_{adv}\mathcal{L}_{adv}$.

3 Experiments

Datasets. We evaluate our method on three 2D medical imaging datasets: **BloodMNIST** [36] (Eosinophil Vs Neutrophil; Train=2500/Test=500), **OCTMNIST** [36] (Normal Vs Choroidal neovascularization; Train=17000/Test=3000 + Diabetic macular edema Vs Drusen; Train=7000/Test=2000), and **BraTS2023** [24] (Healthy Vs Tumor; Train=8000/Test=2000). For each dataset, we define two groups, denoted X and Y, based on the provided labels. We construct balanced splits such that X and Y contain the same number of images in both training and test sets. For BloodMNIST and OCTMNIST, we use the 224×224 images from the MedMNIST+ release and resize them to 256×256 to match the input resolution of StyleGAN2. For BraTS MRI scans, we zero-pad each slice to 256×256 to preserve the original aspect ratio.

Implementation Details. We implement our method using a two-stage training strategy similar to [5]: (1) learning common and salient factors in the W^+ space, followed by (2) refining image details in F space. In Stage 1, we optimize the separator H_{cs} to disentangle common and salient factors. Training alternates between updating $\mathcal{D}$ and $\mathcal{R}$ by maximizing Eqs. (2)–(3) and updating H_{cs} by minimizing Eq. (5) with $\mathcal{D}$ and $\mathcal{R}$ frozen. In Stage 2, we refine the results in the F-space (right panel of Fig. 1) to enhance image quality. The learned latent factors and their shifts (Δf in Eq. (6)) are mapped by E_f and F_{Adpt} and injected into G to minimize $\mathcal{L}_{\text{rec}}$. As architectures, we use the pSp E from [26], the separator H_{cs} from [10], and employ StyleGAN2 G [17] together with the F-space module (E_f, F_{Adpt}) from [5], all pretrained on the corresponding training dataset. The StyleGAN2 generator is pre-trained separately for each dataset and then kept frozen when training the separator and the F-space refinement module. Hyperparameters are selected on validation sets, not tuned on test sets, and kept the same across datasets. More details are provided in the code repository: `https://github.com/BioMedTP/CF_Contrastive_Analysis`.

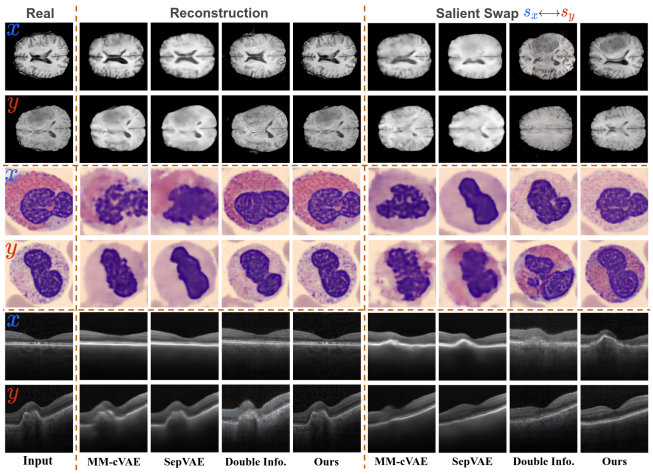

Fig. 2. Comparisons with CA baselines on BraTS, BloodMNIST, and OCTMNIST (Normal vs. CNV) (top to bottom). *Reconstruction*: use the common and salient factors from original images. *Salient Swap*: swap salient factors between x and y while keeping common factors fixed during generation.

Baselines and Evaluation Metrics. We compare our method with CA baselines MM-cVAE [34], SepVAE [22], and Double InfoGAN [7], and with diffusion-based VCE methods: ACE [15], DiME [14], FastDiME [35], and TIME [16]. We also consider 2 off-the-shelf T2I DMs (not CF-specific methods): FLUX [4] and SDXL [25] adapted with LoRA [12]. In Table 1, we evaluate the proposed method on CA and CF generation using widely adopted metrics to evaluate synthesis quality: L2, LPIPS [37], MS-SSIM [33], and FID [11]. We also assess whether swapped outputs match the post-edit domain distribution reporting $\text{FID}_{X\to Y}$ (FID between X images after swapping to Y and real Y) and $\text{FID}_{Y\to X}$ (swapped Y vs. real X). We also use a U-Net-based classifier pre-trained on real training images, using the same architecture as in [35], and report accuracy w.r.t. the desired post-edit label. About latent separation quality, we train a classifier on the learned latent factors to predict the class (X or Y). In Table 2, higher accuracy indicates the class is encoded in the corresponding (common or salient) space. Please note that in the BT assumption, the information about X is entirely encoded in c. Eventually, following VCE prior works [15, 35], we report in Table 4 FID, sFID, Flip Ratio (FR), Mean Absolute Difference (MAD) and Bidirectional KL Divergence (BKL) for measuring counterfactual quality, using the previously described U-Net-based classifier.

Comparison with CA Baselines. Fig. 2 qualitatively compares our method with CA baselines for reconstruction and salient swapping. Our reconstructions are sharper and better preserve fine anatomical structures. For salient swapping, our approach transfers the salient pattern (tumors in BraTS, eosinophil–neutrophil staining differences in BloodMNIST, and CNV-related abnormalities in OCTM-

Table 1. Quantitative comparison results with CA baselines on multiple datasets. Best results in bold, second-best underlined (within each dataset block). Ref: F-space refinement; BT/MS Ass.: background-target/multi-salient assumptions.

Dataset (X vs. Y)	Method	Reconstruction (X)					Reconstruction (Y)					Swap		
		PSNR↑	MSE↓	MS-SSIM↑	LPIPS↓	FID↓	PSNR↑	MSE↓	MS-SSIM↑	LPIPS↓	FID↓	$\text{FID}_{X\to Y}$ ↓	$\text{FID}_{Y\to X}$ ↓	Acc (X+Y)↑
BraTS (Healthy vs. Tumored)	**MMc-VAE**	20.32	0.008	0.852	0.163	89.86	21.06	0.007	0.951	0.157	86.36	98.63	92.77	0.753
	SepVAE	23.24	0.006	0.933	0.156	115.2	23.75	0.006	0.930	0.152	109.4	113.7	111.9	0.821
	Double InfoGAN	23.96	0.005	0.925	0.151	125.1	25.44	0.004	0.958	0.145	128.3	137.2	135.4	0.901
	Ours (pSp-cs)	24.03	0.004	0.91	0.054	44.46	24.75	0.003	0.90	0.055	30.10	37.10	45.50	**0.939**
	Ours (pSp-cs-Ref)	**38.19**	**0.0002**	**0.99**	**0.004**	**2.64**	**38.20**	**0.0002**	**0.99**	**0.005**	**2.57**	**25.17**	**32.56**	0.914
BloodMNIST (Eos. vs. Neutrophil)	**MMc-VAE**	18.27	0.015	0.556	0.378	175.1	19.42	0.011	0.643	0.357	179.6	181.7	183.9	0.819
	SepVAE	18.04	0.016	0.551	0.421	182.8	18.01	0.016	0.579	0.419	219.5	217.5	221.9	0.691
	Double InfoGAN	19.45	0.015	0.567	0.472	95.97	19.70	0.014	0.606	0.491	104.7	149.50	135.7	0.646
	Ours (pSp-cs)	22.35	0.006	0.725	0.139	73.46	22.43	0.006	0.760	0.139	70.42	84.05	88.61	0.979
	Ours (pSp-cs-Ref)	**41.99**	**0.0001**	**0.998**	**0.004**	**2.887**	**42.54**	**0.0001**	**0.998**	**0.004**	**2.590**	**43.37**	**45.90**	**0.981**
OCTMNIST (DME vs. Drusen)	**MMc-VAE**	23.44	0.005	0.805	0.404	128.05	26.19	0.003	0.867	0.390	117.71	126.49	128.17	0.665
	SepVAE	23.89	0.005	0.822	0.400	107.84	26.26	0.003	0.872	0.395	105.92	131.22	143.67	0.597
	Double InfoGAN	16.84	0.026	0.592	0.369	181.38	18.89	0.017	0.698	0.300	179.87	189.06	199.42	0.832
	Ours (BT Ass.)	23.79	0.005	0.806	0.111	51.10	24.99	0.004	0.830	0.084	42.62	65.66	79.13	0.627
	Ours (BT Ass. Ref)	**38.24**	0.0002	0.992	0.0143	5.832	39.02	0.0001	0.994	0.012	**3.294**	44.67	63.62	0.689
	Ours (MS Ass.)	26.50	0.005	0.814	0.108	48.64	24.70	0.004	0.837	0.083	39.43	59.21	73.79	**0.876**
	Ours (MS Ass. Ref)	38.09	**0.0001**	**0.994**	**0.014**	**5.356**	**39.00**	**0.0001**	**0.995**	**0.012**	3.911	**39.24**	**55.92**	0.863

Table 2. Latent separation results. Mean ± std of 5-fold cross-validated classification accuracy to distinguish between the class of X or Y using logistic regression and using as features either C, S_x or S_y. $\Delta = |0.5 - C| + |1.0 - S_1|(+|1.0 - S_2|)$.

Model	BraTS (Healthy vs. Tumored)			OCTMNIST (DME vs. Drusen)			
	C	S_y	Δ	C	S_x	S_y	Δ
MM-cVAE	0.68 ± 0.02	0.73 ± 0.02	0.45	0.61 ± 0.01	0.69± 0.03	–	0.42
SepVAE	0.67 ± 0.02	0.92 ± 0.01	0.25	0.64 ± 0.02	0.94 ± 0.01	–	0.20
Double Info.	0.65 ± 0.01	0.86 ± 0.01	0.29	0.76 ± 0.02	0.61 ± 0.02	–	0.65
Ours (BT Ass.)	**0.58 ± 0.08**	**0.95 ± 0.01**	**0.13**	0.57 ± 0.08	0.96 ± 0.006	–	0.11
Ours (MS Ass.)	–	–	–	**0.54 ± 0.05**	**0.97 ± 0.008**	**0.98 ± 0.003**	**0.09**
Expected	0.5	1.0	0	0.5	1.0	1.0	0

Table 3. Ablation study on the BraTS dataset: Healthy (X) vs. Tumored (Y).

Model	Latent Separation			Image Edit Quality		
	C	S	Δ	$\text{FID}_{X\to Y}$	$\text{FID}_{Y\to X}$	Acc.
Base ($\mathcal{L}_{\text{lat}}+\mathcal{L}_{\text{img}}$)	0.74 ± 0.010	0.73 ± 0.02	0.51	39.78	55.75	0.535
Base + $\mathcal{L}_D$	0.64 ± 0.02	0.92 ± 0.01	0.22	37.03	53.21	0.638
Base + $\mathcal{L}_D+\mathcal{L}_{\text{DiscMI}}$	0.68 ± 0.01	0.86 ± 0.01	0.32	38.78	55.04	0.593
Base + $\mathcal{L}_D+\mathcal{L}_R$	**0.58 ± 0.08**	**0.95 ± 0.01**	**0.13**	37.10	45.50	**0.939**
Base + $\mathcal{L}_D+\mathcal{L}_R+\mathcal{L}_{\text{refine}}$	–	–	–	**25.17**	**35.56**	0.913

NIST) while preserving shared anatomy (healthy brain tissues, global cell layout/nuclear morphology, and normal retinal layer structure, respectively). Table 1 reports quantitative results, where our pSp-based method, with or without F-space refinement (**-Ref**), consistently outperforms all CA baselines on both reconstruction and swapping. Notably, with the MS setting, our method better handles the challenging DME vs. Drusen case, achieving the best reconstruction metrics and the highest swapping quality. Finally, Table 2 evaluates latent separation between the learned common (C) and salient (S) factors. Salient patterns are mainly captured in S rather than in C, and our method attains the lowest separation gap Δ across all methods, indicating improved semantic disentanglement over prior approaches. In Table 3, we also report an ablation study to assess the contribution of each training component.

Counterfactual Generation Based on CA. Table 4 reports quantitative comparisons on BraTS Healthy (X) and Tumored (Y) for two counterfactual directions: adding tumors ($X \to Y$) and removing tumors ($Y \to X$). Here, the *counterfactual generation* is equivalent to the salient-factor swap operation in Table 1. FID differences arise because Table 1 uses random X, Y pairs, whereas Table 4 fixes either X or Y using the source images shown in Fig. 3. Overall, our CA-based swapping achieves the best image fidelity in both directions, as indicated by the lowest FID/sFID scores, and is better or on par with diffusion-based CF baselines (ACE, DiME, and FastDiME) in terms of counterfactual

Table 4. CF generation comparison with VCE-specific methods on BraTS. Best in bold, second-best underlined. Metrics for CF edits (X→Y: add tumors; Y→X: remove tumors) and runtime.

	CF: X→Y (add tumors)						CF: Y→X (remove tumors)						
Method	FID	sFID	L1	FR	MAD	BKL	FID	sFID	L1	FR	MAD	BKL	Time (s/img)
ACE	38.553	51.428	**0.003**	**0.977**	0.942	**0.043**	39.436	52.841	**0.005**	0.895	0.855	0.120	21.37 ± 0.34
DiME	38.378	50.421	0.026	0.947	0.934	0.054	48.752	60.722	0.025	0.741	0.723	0.228	294.12 ± 47.44
FastDiME	32.431	44.735	0.017	0.905	0.873	0.108	51.391	64.007	0.016	0.696	0.658	0.285	16.07 ± 0.64
FastDiME-2+	32.865	45.283	0.017	0.912	0.876	0.106	52.191	64.753	0.016	0.707	0.668	0.277	30.69 ± 0.91
TIME	76.704	91.064	0.012	0.389	0.293	0.475	88.706	100.074	0.012	0.341	0.247	0.413	17.59 ± 0.40
SDXL-LoRA	51.258	60.179	0.027	0.731	0.727	0.230	59.006	69.185	0.028	0.824	0.817	0.164	4.95 ± 0.67
Ours	**27.533**	**38.441**	0.021	0.953	**0.946**	0.047	**37.492**	**47.527**	0.022	**0.914**	**0.906**	**0.074**	**0.25** ± **0.005**

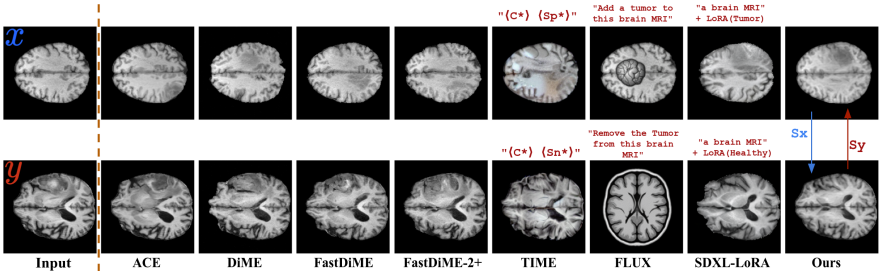

Fig. 3. Comparisons of CF generation. Cols. 2–5 show CF outputs from SOTA diffusion-based methods. Cols. 6–8 show CF outputs from T2I diffusion models.

effectiveness (i.e., FR/MAD/BKL), while achieving the shortest inference time (0.25 s/img). In Fig. 3, we observe that diffusion-based methods struggle to completely remove large tumors from Y images, and the added tumors in $X \to Y$ have often limited size and/or are misplaced. In addition, text-conditioned editing (e.g., FLUX) or learning per-image embeddings (e.g., TIME) can drift away from the real image manifold, while fine-tuning large diffusion models with LoRA (e.g., SDXL-LoRA) may alter common content beyond the desired tumor-related changes. By contrast, our approach explicitly swaps salient generative factors between X and Y, improving image fidelity and offering better controllability for tumor-specific counterfactual manipulation.

Thanks to the well structured latent space of StyleGAN2, we can also interpolate between salient factors. Given a pair of samples (x, y) from two domains/classes, we can generate a continuous counterfactual trajectory by scaling the salient change with a scalar $\alpha \in [0, 1]$, e.g., $G(\alpha \widehat{f}_{x \to y})$. This is shown in Fig. 4, where the predicted class probability changes smoothly along the trajectory, suggesting that the generated counterfactuals form a coherent explanation path.

4 Conclusion

We propose a classifier-free, CA-based framework for VCEs that yields strong common–salient disentanglement and high-quality CF generations. Future work will include a formal multi-expert reader study to assess clinical relevance, as well

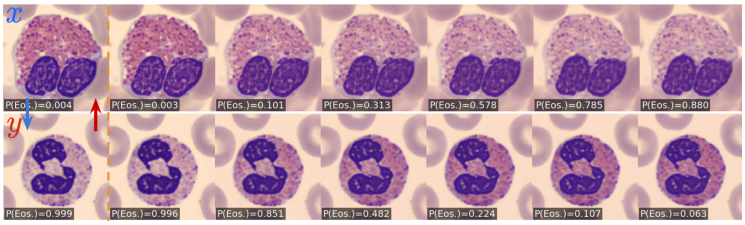

Fig. 4. Interpolation along salient factors. Gradual interpolation from s_x to s_y (and viceversa) using an interpolation weight α $(0 \to 1)$ during generation. The classifier prediction changes consistently along the trajectory, illustrating gradual insertion/removal of class-specific evidence while preserving common content.

as experiments on bias reproduction to evaluate whether classifier-free CF generation is less susceptible than classifier-guided VCEs [35]. Interesting perspectives also include adapting the proposed method to recent VLM, as in [18], and studying its identifiability. Similarly to all recent VCE [14, 15, 35] and CA [7, 22] works, we note that our method also, without strong assumptions, is not identifiable [13, 21] and does not enable causal discovery. This is generally impossible from purely observational single-modal data. An interesting direction, in line with causal representation learning theory [29], would be using additional prior or longitudinal data, to further constrain the learning process.

Acknowledgments. This work was performed using HPC resources from GENCI–IDRIS (A0160615058) and it was supported by the EUR BERTIP (ANR-18-EURE-0002).

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

References

1. Abid, A., Zhang, M.J., Bagaria, V.K., Zou, J.: Exploring patterns enriched in a dataset with contrastive principal component analysis. Nature Communications (2018)
2. Augustin, M., Boreiko, V., Croce, F., Hein, M.: Diffusion visual counterfactual explanations. In: NeurIPS (2022)
3. Benaim, S., Khaitov, M., Galanti, T., Wolf, L.: Domain intersection and domain difference. In: ICCV (2019)
4. Black Forest Labs, et al.: FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space (2025)
5. Bobkov, D., Titov, V., Alanov, A., Vetrov, D.: The devil is in the details: StyleFeatureEditor for detail-rich StyleGAN inversion and high quality image editing. In: CVPR (2024)
6. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain separation networks. In: NeurIPS (2016)
7. Carton, F., Louiset, R., Gori, P.: Double InfoGAN for contrastive analysis. In: AISTATS (2024)

8. Gonzalez-Garcia, A., van de Weijer, J., Bengio, Y.: Image-to-image translation for cross-domain disentanglement. In: NeurIPS (2018)
9. Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S.: Counterfactual Visual Explanations. In: ICML (2019)
10. He, Y., Lesné, G., Liu, Z., Soumm, M., Gori, P.: Learning common and salient generative factors between two image datasets (2025)
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
13. Hyvärinen, A., Sasaki, H., Turner, R.E.: Nonlinear ICA using auxiliary variables and generalized contrastive learning. In: AISTATS (2019)
14. Jeanneret, G., Simon, L., Jurie, F.: Diffusion models for counterfactual explanations. In: ACCV (2022)
15. Jeanneret, G., Simon, L., Jurie, F.: Adversarial counterfactual visual explanations. In: CVPR (2023)
16. Jeanneret, G., Simon, L., Jurie, F.: Text-to-image models for counterfactual explanations: a black-box approach. In: WACV. pp. 4757–4767 (2024)
17. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: CVPR (2020)
18. Kazimi, T., Allada, R., Yanardag, P.: Explaining in Diffusion: Explaining a Classifier with Diffusion Semantics. In: CVPR (2025)
19. Kleinman, M., Achille, A., Soatto, S., Kao, J.C.: Gács–körner common information variational autoencoder. In: NeurIPS (2023)
20. Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., Freeman, W.T., Isola, P., Globerson, A., Irani, M., Mosseri, I.: Explaining in style: Training a GAN to explain a classifier in stylespace. In: ICCV (2021)
21. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: ICML (2019)
22. Louiset, R., Duchesnay, E., Grigis, A., Dufumier, B., Gori, P.: SepVAE: A contrastive VAE to separate pathological patterns from healthy ones. In: MIDL (2024)
23. Louiset, R., Duchesnay, E., Grigis, A., Gori, P.: Separating common from salient patterns with contrastive representation learning. In: ICLR (2024)
24. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE TMI (2015)
25. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: ICLR (2024)
26. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a StyleGAN encoder for image-to-image translation. In: CVPR (2021)
27. Rodríguez, P., Caccia, M., Lacoste, A., Zamparo, L., Laradji, I., Charlin, L., Vazquez, D.: Beyond trivial counterfactual explanations with diverse valuable explanations. In: ICCV (2021)
28. Sánchez, E.H., Serrurier, M., Ortner, M.: Learning disentangled representations via mutual information estimation. In: ECCV (2020)
29. Schölkopf, B., Kügelgen, J.v.: From Statistical to Causal Learning. In: ICM (2022)

30. Severson, K.A., Ghosh, S., Ng, K.: Unsupervised learning with contrastive latent variable models. In: AAAI (2019)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
32. Singla, S., Pollack, B., Chen, J., Batmanghelich, K.: Explanation by progressive exaggeration. In: ICLR (2020)
33. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Asilomar Conference on Signals, Systems and Computers (2003)
34. Weinberger, E., Beebe-Wang, N., Lee, S.I.: Moment matching deep contrastive latent variable models. In: AISTATS (2022)
35. Weng, N., Pegios, P., Petersen, E., Feragen, A., Bigdeli, S.: Fast diffusion-based counterfactuals for shortcut removal and generation. In: ECCV (2024)
36. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data (2023)
37. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
38. Zou, K., Faisan, S., Heitz, F., Valette, S.: Joint disentanglement of labels and their features with VAE. In: ICIP (2022)