DISENTANGLED COUNTERFACTUAL DIFFUSION FOR IDENTITY-PRESERVING EXPLANATIONS OF MEDICAL IMAGE CLASSIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical image classifiers can be accurate while relying on clinically irrelevant evidence. Local saliency maps expose association or sensitivity, but do not test whether changing the highlighted evidence changes the decision; unconstrained counterfactual generators can cross a decision boundary by altering patient-specific anatomy. We introduce Disentangled Counterfactual Diffusion (DCD), an identity-preserving explanation framework that separates an image into a compact class-associated code and a multiscale individual representation. A conditional diffusion editor retains the source identity features while moving the class code toward an opposite-class donor. Training combines diffusion reconstruction, class supervision, prototype geometry, adversarial removal of class information from the identity branch, cross-covariance decorrelation, exchange consistency, structural preservation, cycle reconstruction, and realism constraints. At inference, DCD samples six points on a linear class-code path and selects the nearest generated image that changes the prediction; its difference from the source defines the explanation. On retinal OCT, chest X-ray, and two brain-tumor benchmarks, DCD achieves the best AOPC on all four datasets and the best peak degradation on three of four datasets among eleven baselines, while preserving source identity in the generated counterfactuals. These results show that explicitly separating decision-related semantics from individual structure can produce more faithful, controlled explanations of medical image classifiers.

1 INTRODUCTION

Deep neural networks increasingly support decisions in radiology and ophthalmology, yet predictive performance alone does not show that a model uses medically appropriate evidence. Classifiers can exploit acquisition artifacts, demographic signals, support devices, or site-specific correlations and still perform well in-distribution (DeGrave et al., 2021; Gichoya et al., 2022; Vaidya et al., 2024). This gap is consequential in medicine: before a prediction can inform care, clinicians need evidence that the model responded to pathology rather than to a shortcut.

Most post-hoc methods return a saliency map. Gradient approaches propagate sensitivity through the classifier (Selvaraju et al., 2017; Srinivas & Fleuret, 2019), while perturbation approaches infer importance from output changes after masking image regions (Ribeiro et al., 2016). Such maps are useful summaries, but neither sensitivity nor association constitutes an intervention. A highlighted region may correlate with the output without being sufficient to change it, and saliency rankings can depend strongly on the architecture, metric, and perturbation design (Adebayo et al., 2018; Hesse et al., 2024; Klein et al., 2024).

Counterfactual explanations ask a sharper question: how must this image change for the classifier to return another decision? Early methods exchange regions or move through an autoencoder latent space (Goyal et al., 2019; Cohen et al., 2021). Medical counterfactuals remain difficult because a class-changing image is not automatically an explanation. If generation alters anatomy, acquisition style, or unrelated background, the resulting difference map mixes the intended decision factor with identity changes. The central requirement is therefore dual: produce a valid decision change while preserving source-specific structure.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We propose **Disentangled Counterfactual Diffusion (DCD)**. DCD factorizes an image into a low-dimensional class-associated code and a multiscale identity pyramid, then conditions a diffusion U-Net on both. The class branch is supervised and prototype-structured; gradient reversal and cross-covariance regularization suppress class leakage from the identity branch. During cross-class exchange, a frozen target classifier constrains the generated decision, while re-encoding, anatomy, edge, cycle, and realism losses constrain what may change. At inference, a linear path between source and donor class codes yields a sequence of controlled counterfactuals. The first point that changes the classifier decision defines a minimal sampled intervention and an instance-level difference map.

Our contributions are threefold. First, we introduce a source-anchored diffusion editor that combines a compact class code with multiscale identity features at every denoising resolution. Second, we couple representation separation with exchange consistency and explicit identity-preservation objectives so that a class intervention need not rewrite patient anatomy. Third, we evaluate the method under the matched AOPC and peak-degradation protocol of Xie et al. (2026) on four medical imaging datasets and eleven competing explanation methods. DCD ranks first in seven of eight dataset–metric comparisons and attains the best AOPC on every dataset.

2 Related Work

Local attribution and faithfulness. Grad-CAM combines gradients with convolutional activations to produce class-discriminative maps (Selvaraju et al., 2017); FullGrad aggregates input and bias gradients across a network (Srinivas & Fleuret, 2019); and LIME fits a sparse local surrogate to perturbed samples (Ribeiro et al., 2016). Their visual plausibility is insufficient evidence of fidelity: some maps remain similar after model or label randomization (Adebayo et al., 2018). Deletion metrics test explanations more directly by perturbing highly ranked locations and measuring confidence loss (Samek et al., 2017; Hooker et al., 2019). However, rankings can change with the perturbation process and evaluation metric (Hesse et al., 2024; Klein et al., 2024). We therefore use exactly the AOPC and peak-degradation definitions adopted by the matched medical benchmark (Xie et al., 2024; 2026).

Generative counterfactual explanations. Counterfactual visual explanations replace or synthesize evidence rather than differentiating the classifier (Goyal et al., 2019; Cohen et al., 2021). ICAM-Reg disentangles attribute and content codes in neurological scans (Bass et al., 2022); DiffExplainer generates medical images that alter prognosis (Fang et al., 2024); and DeCoDEx adds confounder-detector guidance to diffusion counterfactuals (Fathi et al., 2024). Diffusion autoencoders enable continuous transitions across classification boundaries (Preechakul et al., 2022; Atad et al., 2024), while semantic abduction formalizes increasingly structured diffusion interventions (Rasal et al., 2025). These methods motivate diffusion as a high-fidelity editor, but do not jointly impose DCD's explicit two-branch factorization, multiscale source anchoring, and exchange consistency.

Class-associated representations. CAE separates class-associated information from individual information and recombines them to generate class-changing samples (Xie et al., 2024). CAML adds topological analysis and geodesic paths, linking global class-transition structure to instance explanations (Xie et al., 2026). DCD retains the useful principle that decision-related semantics should be isolated, but replaces adversarial decoding with conditional diffusion and adds explicit constraints against class leakage and identity drift. This is an operational, classifier-centered factorization rather than an identifiable causal model: without stronger assumptions, counterfactual image editing cannot establish biological causation (Pan & Bareinboim, 2024).

3 Disentangled Counterfactual Diffusion

3.1 Problem formulation and architecture

Let a fixed classifier f map an image $x \in \mathbb{R}^{C\times H\times W}$ to class probabilities. We seek a spatially coherent image change that alters f while retaining source anatomy and acquisition characteristics. DCD decomposes x into

$$z_x^c = E_c(x), \qquad s^x = E_i(x) = \{s_x^{(\ell)}\}_{\ell=0}^{L-1}, \tag{1}$$

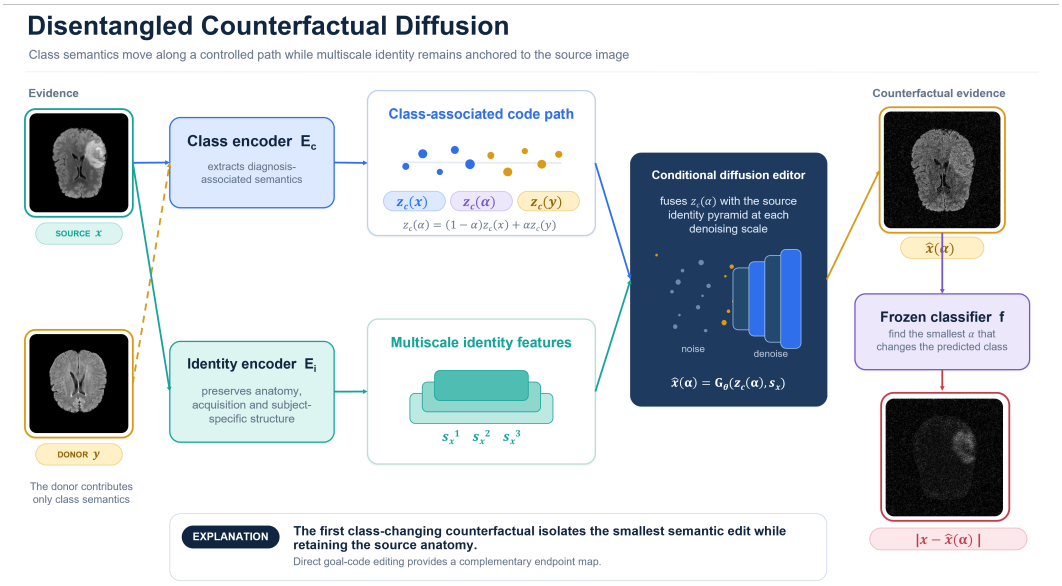

Figure 1: Overview of DCD. The source image supplies the multiscale individual representation, while the class code moves toward an opposite-class donor. The conditional diffusion editor synthesizes counterfactuals; the nearest sampled point that changes the frozen classifier defines the explanation.

where $z_x^c \in \mathbb{R}^8$ is a compact class-associated vector and s^x is a four-level spatial feature pyramid. The class encoder uses residual downsampling, global average pooling, and a normalized linear projection. The identity encoder preserves feature maps at four resolutions. A conditional U-Net G_θ receives projected identity features at the matching resolutions; timestep and class embeddings modulate residual blocks by feature-wise scale and shift, and self-attention is applied at the bottleneck.

We use a cosine schedule with $T = 1000$ diffusion steps and velocity prediction. With noised image x_t and velocity target v_t, the diffusion objective is

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x,t,\epsilon}\left[\|v_t - v_\theta(x_t, t, s^x, z_x^c)\|_2^2\right]. \tag{2}$$

Classifier-free guidance is trained by occasionally replacing z_x^c with a learned null code. At inference, deterministic DDIM (Song et al., 2021) starts from a noised version of the source rather than an unconditional sample, anchoring the reverse process to source structure.

3.2 Disentanglement and Exchange Training

The class code predicts the label and is attracted to a learnable class prototype; prototype separation penalizes positive cosine similarity between different classes. Conversely, the coarsest identity feature passes through a class head after a gradient-reversal layer, encouraging E_i to remove label information. A cross-covariance penalty further discourages redundant information across branches:

$$\mathcal{L}_{\text{dis}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{proto}}\mathcal{L}_{\text{proto}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{cov}}\mathcal{L}_{\text{cov}}. \tag{3}$$

After warm-up, source x is paired with an image y from another class. Keeping s^x while supplying the detached donor code z_y^c produces

$$\hat{x}^{\cdot, y\leftarrow x} = G_\theta(x_t, t, s^x, z_y^c), \qquad y_y \neq y_x. \tag{4}$$

The frozen classifier requires $\hat{x}^{\cdot, y\leftarrow x}$ to predict the donor class. Re-encoding matches its class code to z_y^c and the donor prototype. A normalized multiscale identity loss matches the source and generated identity pyramids; multiscale ℓ_1 and gradient losses preserve anatomy and edges. Same-class exchange is reconstructed, and a cycle edit restores the original class code. A spectrally normalized patch discriminator adds a low-weight realism objective. Figure 2 summarizes the coupled signals. The target classifier is never updated, although the present training implementation propagates gradients through it; DCD is therefore output-only at inference but requires differentiable classifier access during training.

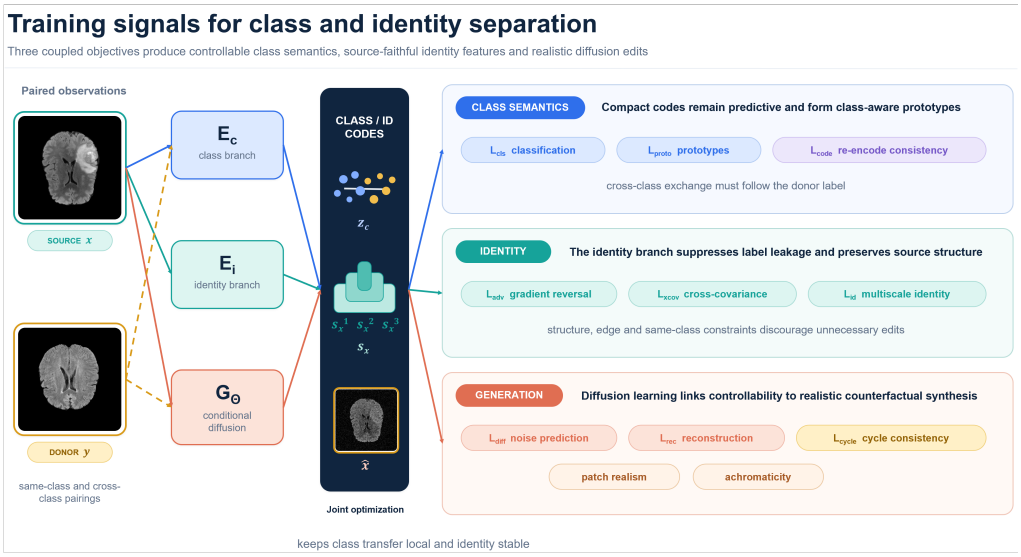

Figure 2: Training signals for representation separation and controlled generation. Class semantics are supervised and prototype-structured; label leakage is suppressed in the identity branch; and exchanged samples are constrained by diffusion, reconstruction, cycle, structure, and realism objectives.

3.3 Class-code paths and saliency

For test image x, we sample a donor y from a different dataset class, encode z_x^c, z_y^c, and s^x, and construct

$$z(\alpha) = (1-\alpha)z_x^c + \alpha z_y^c, \quad \alpha \in \{0, .2, .4, .6, .8, 1\}. \tag{5}$$

All six edits share the source, identity pyramid, and noise realization. Let $c_x = \arg\max f(x)$ and $\widehat{x}_\alpha = G_\theta(x_t, t, s^x, z(\alpha))$. We select

$$\alpha^* = \min\{\alpha : \arg\max f(\widehat{x}_\alpha) \neq c_x\}, \qquad S(x) = \frac{1}{C}\sum_{c=1}^{C} |x^{(c)} - \widehat{x}_{\alpha^*}^{(c)}|. \tag{6}$$

This direct-success map (D_1) is the nearest sampled boundary-crossing edit. The goal map $D_2 = |x - \widehat{x}_1|$ is always defined and represents the full donor-code displacement. For OCT, the external classifier remains four-way: raw argmax class 0 is normal and raw argmax classes 1–3 form the diseased superclass; disease probabilities are not summed. Figure 4 illustrates the procedure.

4 Experimental Setup

4.1 Datasets and target classifiers

We follow the four datasets and train/test partitions used by the matched CAE/CAML benchmark (Xie et al., 2024; 2026). Images are center-cropped or resized to 256×256, and training applies random horizontal flipping with probability 0.5. OCT contains normal scans and three retinal disease categories (CNV, DME, and drusen) (Kermany et al., 2018). Chest X-rays contain normal and pneumonia images. Brain Tumor1 uses Br35H (Hamada, 2020); Brain Tumor2 is derived from BraTS 2021 (Baid et al., 2021). Table 1 reports the benchmark splits.

Each dataset has a separately trained ResNet-50 target classifier whose parameters remain fixed. Binary counterfactual success is an argmax change. For OCT, the normal/diseased decision uses the raw four-class argmax as described above, while probability-based evaluation tracks the original raw classifier class. DCD uses an eight-dimensional class code, base width 64, four identity levels, AdamW, cosine learning-rate decay, mixed precision, gradient clipping, and an exponential moving average. The reference configuration uses learning rate 8×10^{-5}, weight decay 10^{-4}, effective batch size 8, 3,000 warm-up updates, and 100,000 optimizer updates. Evaluation uses deterministic DDIM ($\eta = 0$).

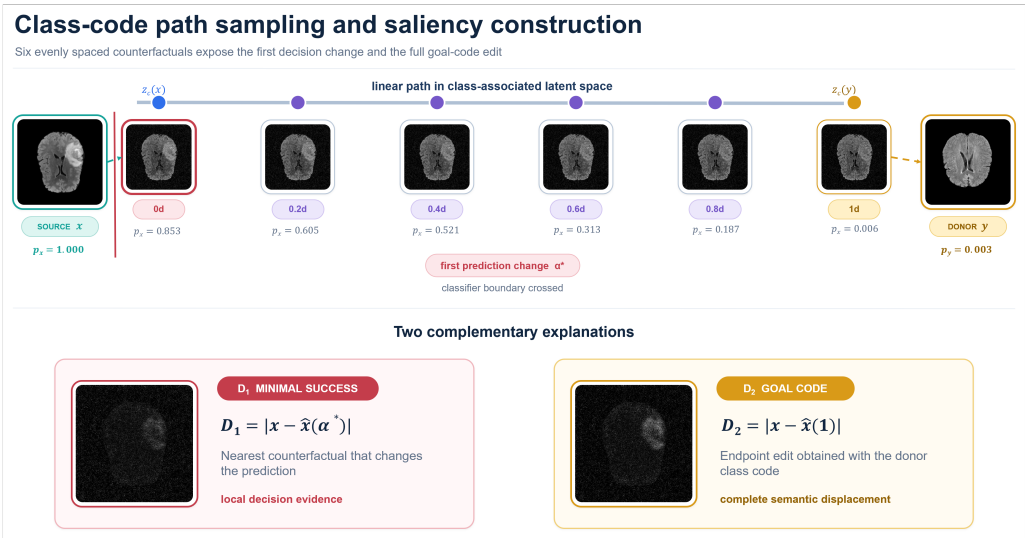

Figure 3: Class-code path sampling. Six evenly spaced counterfactuals expose the first decision change. D_1 uses the nearest successful point; D_2 uses the donor-code endpoint.

Table 1: Dataset tasks and benchmark split sizes.

Dataset	Task	Train		Test	
		Normal	Abnormal	Normal	Abnormal
OCT	normal vs. disease	8,000	24,000	250	750
Chest X-ray	normal vs. pneumonia	1,349	3,883	234	390
Brain Tumor1	normal vs. tumor	1,200	1,200	300	300
Brain Tumor2	normal vs. tumor	710	4,398	302	1,623

4.2 Baselines and Metrics

We compare with all methods in Figure 5b of Xie et al. (2026): LIME, FullGrad, Simple FullGrad, Smooth FullGrad, Grad-CAM, LAGAN, ICAM, DiffExplainer, CBED, CAE, and CAML. This set spans local perturbation, gradient attribution, adversarial counterfactual generation, case-based explanation, and class-associated representation learning.

We evaluate saliency faithfulness using area over the perturbation curve (AOPC) and peak degradation (PD). Ranked pixels are progressively replaced by random values; $x_{i,p}$ is image i after step p, and c_i is its original predicted class:

$$\mathrm{AOPC} = \frac{1}{KN} \sum_{i=1}^{K} \sum_{p=1}^{N} \left[f_{c_i}(x_{i,0}) - f_{c_i}(x_{i,p}) \right], \tag{7}$$

$$\mathrm{PD} = \max_{p} \frac{1}{K} \sum_{i=1}^{K} \left[f_{c_i}(x_{i,0}) - f_{c_i}(x_{i,p}) \right]. \tag{8}$$

Higher values mean that deleting highly ranked locations causes a larger confidence decrease. The terminology is *AOPC* (not "APOC"), following the cited benchmark.

5 Results

5.1 Faithfulness Across Datasets

Tables 2 and 3 report the matched comparison. DCD obtains the highest AOPC on all four datasets. Relative to the strongest prior class-associated method, CAML, the absolute gains are 0.100 on OCT, 0.019 on Chest, 0.026 on Brain Tumor1, and 0.082 on Brain Tumor2. These gains span retinal OCT,

Table 2: AOPC comparison (higher is better). Best values are bold.

Method	OCT	Chest	Brain1	Brain2
LIME	0.513	0.695	0.046	0.387
FullGrad	0.307	0.767	0.092	0.446
Simple FullGrad	0.302	0.782	0.090	0.466
Smooth FullGrad	0.314	0.669	0.079	0.302
Grad-CAM	0.365	0.485	0.014	0.267
LAGAN	0.500	0.912	0.105	0.578
ICAM	0.490	0.883	0.098	0.510
DiffExplainer	0.557	0.898	0.079	0.571
CBED	0.408	0.814	0.086	0.562
CAE	0.571	0.917	0.116	0.605
CAML	0.582	0.926	0.123	0.609
DCD (ours)	**0.682**	**0.945**	**0.149**	**0.691**

Table 3: Peak degradation (PD; higher is better). Best values are bold and the runner-up is underlined.

Method	OCT	Chest	Brain1	Brain2
LIME	0.682	0.931	0.067	0.681
FullGrad	0.454	0.954	0.138	0.582
Simple FullGrad	0.439	0.961	0.136	0.607
Smooth FullGrad	0.461	0.891	0.111	0.448
Grad-CAM	0.498	0.769	0.032	0.520
LAGAN	0.669	0.963	0.151	0.663
ICAM	0.670	0.961	0.153	0.595
DiffExplainer	0.705	0.965	0.132	0.621
CBED	0.658	0.952	0.135	0.615
CAE	0.708	0.968	0.182	0.689
CAML	0.717	**0.972**	0.210	0.709
DCD (ours)	**0.820**	<u>0.963</u>	**0.233**	**0.768**

radiography, and brain MRI, suggesting that the benefit is not confined to one modality or lesion morphology.

DCD also achieves the best PD on OCT and both brain datasets, improving over CAML by 0.103, 0.023, and 0.059, respectively. On Chest, DCD reaches 0.963: competitive with the strongest generative baselines but below CAML's 0.972. This exception is informative. Chest abnormalities can be diffuse, so concentrating the ranking on a smaller set of counterfactual changes can improve average degradation (AOPC) without maximizing the single largest degradation point. We therefore avoid claiming uniform dominance and summarize DCD as best in seven of eight dataset–metric comparisons.

The pattern is consistent with the method design. Local gradients may emphasize sensitive but non-decisive pixels, whereas unconstrained generators can cross a boundary through broad appearance changes. DCD restricts both the intervention direction and the information retained as identity. Consequently, deleting its highest-ranked locations produces a large confidence loss early in the perturbation sequence, as reflected especially by the AOPC gains.

5.2 Qualitative Explanations

We conduct experiments on four medical imaging datasets covering retinal OCT, chest X-ray, and brain MRI. Figure 4 presents representative saliency maps generated by DCD and the competing explanation methods. Across these examples, DCD produces more spatially concentrated and anatomically coherent explanations, with high-saliency regions more closely aligned with the pathological patterns that influence the classifier's predictions. In contrast, several baseline methods yield diffuse responses, emphasize irrelevant background structures, or highlight regions beyond the apparent lesion boundaries. By preserving patient-specific anatomy while selectively modifying class-associated information, DCD suppresses incidental changes and isolates finer-grained diagnostic evidence, including subtle lesion boundaries and low-contrast pathological regions. The resulting saliency

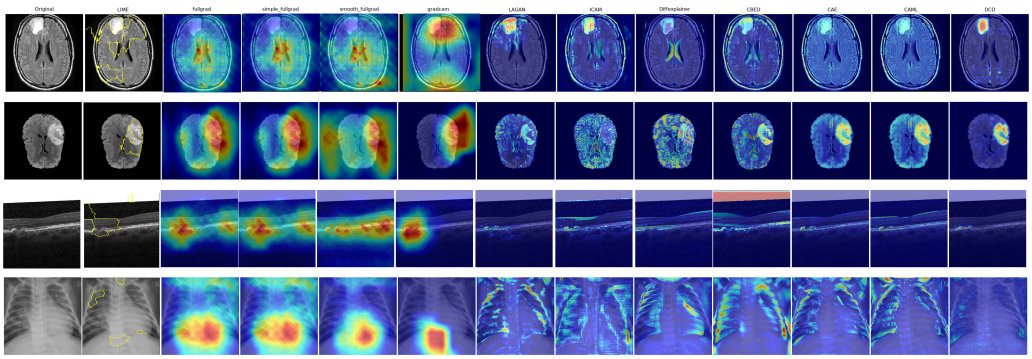

Figure 4: Qualitative comparison for OCT, chest X-rays, Brain Tumor1, and Brain Tumor2.

Table 4: Source preservation for evaluated DCD counterfactuals (lower is better).

Dataset	LPIPS ↓	Identity distance ↓
OCT	0.0269	0.0003
Chest X-ray	0.0843	0.0006
Brain Tumor1	0.0808	0.0114
Brain Tumor2	0.0544	0.0033

maps therefore provide clearer visual evidence of why the classifier changes its decision and make the generated counterfactual explanations easier to interpret. Although these qualitative observations do not by themselves establish clinical correctness, together with the improvements in AOPC and PD, they provide complementary evidence that DCD captures more decision-relevant pathological features than the compared methods.

5.3 Identity preservation of counterfactuals

Faithfulness scores alone cannot establish that a counterfactual remains source-consistent. We therefore measure LPIPS between the source and evaluated counterfactual and the distance between their identity-encoder features. Table 4 summarizes these preservation diagnostics. LPIPS is low on every benchmark, while identity distance ranges from 0.0003 to 0.0114. The modest increase from OCT to Brain Tumor1 indicates that preservation difficulty varies by dataset, but the distances remain close to the source in the learned identity space. These results complement AOPC/PD: the class-code intervention changes classifier evidence while the multiscale identity representation remains comparatively stable.

6 Discussion and Limitations

DCD treats explanation as a constrained intervention: the class-code path proposes a semantic direction, the frozen classifier verifies whether the decision changes, and the identity branch constrains what remains stable. The comparison with CAE and CAML is particularly informative because all three methods share the class-association premise. DCD's gains therefore support the value of its source-anchored diffusion editor and stronger separation constraints rather than a different evaluation target.

Several limitations qualify the results. First, the learned factorization is not causally identifiable; DCD explains the target classifier and may expose dataset correlations rather than biological mechanisms. Second, the current implementation requires gradients through the frozen classifier during exchange training, so it is not a strict query-only black-box method, although inference and path selection use classifier outputs. Third, linear interpolation is pragmatic rather than guaranteed to follow the data manifold, and six samples provide only a coarse boundary estimate. Fourth, AOPC and PD inherit the distribution-shift sensitivity of deletion metrics. Finally, the four benchmarks are two-dimensional curated datasets with fixed ResNet-50 classifiers; generalization to volumetric scans, multi-label tasks, external hospitals, and other classifier families remains untested.

These limitations motivate donor- and noise-sensitivity studies, adaptive path search, class-leakage probes on the identity representation, and clinician-blinded assessment. When dense annotations are available, overlap with lesion masks can complement faithfulness metrics, but localization agreement should not replace the central question of whether the explanation is faithful to the classifier.

7 CONCLUSION

We introduced DCD, an identity-preserving framework that explains medical image classifiers through controlled class-code interventions and source-anchored diffusion. Explicit class/identity separation, exchange consistency, and multiscale preservation produce counterfactual difference maps that achieve the best AOPC on all four benchmarks and the best PD on three. The results support diffusion-based, representation-controlled counterfactuals as a promising route to faithful classifier auditing, while highlighting the need for careful causal interpretation and broader clinical validation.

AI USE STATEMENT

Generative AI tools were used for language editing, translation, LaTeX preparation, and assisting with figure generation.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Matan Atad, David Schinz, Hendrik Moeller, Robert Graf, Benedikt Wiestler, Daniel Rueckert, Nassir Navab, Jan S. Kirschke, and Matthias Keicher. Counterfactual explanations for medical image classification and regression using diffusion autoencoder. *arXiv preprint arXiv:2408.01571*, 2024.

Ujjwal Baid et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Cher Bass et al. ICAM-Reg: Interpretable classification and regression with feature attribution for mapping neurological phenotypes in individual scans. *IEEE Transactions on Medical Imaging*, 42: 959–970, 2022.

Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P. Lungren, and Akshay Chaudhari. Gifsplanation via latent shift: A simple autoencoder approach to counterfactual generation for chest X-rays. In *Proceedings of MIDL*, volume 143 of *PMLR*, pp. 74–104, 2021.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3:610–619, 2021.

Yinan Fang, Shuang Wu, Zhe Jin, Chen Xu, Shuo Wang, Sean Walsh, and Guang Yang. DiffExplainer: Unveiling black box models via counterfactual generation. In *Medical Image Computing and Computer Assisted Intervention*, pp. 208–218, 2024.

Navid Fathi, Abhishek Kumar, Brennan Nichyporuk, Mohammad Havaei, and Talia Arbel. DeCoDEx: Confounder detector guidance for improved diffusion-based counterfactual explanations. In *Proceedings of MIDL*, volume 250 of *PMLR*, pp. 439–451, 2024.

Judy Wawira Gichoya et al. AI recognition of patient race in medical imaging: A modelling study. *The Lancet Digital Health*, 4:e406–e414, 2022.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of ICML*, volume 97 of *PMLR*, pp. 2376–2384, 2019.

Ahmed Hamada. Br35H brain tumor detection 2020. Kaggle dataset, 2020.

Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Benchmarking the attribution quality of vision models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Daniel S. Kermany et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172:1122–1131, 2018.

Lukas Klein, Christian Lueth, Udo Schlegel, Till Bungert, Mennatallah El-Assady, and Paul Jaeger. Navigating the maze of explainable AI: A systematic approach to evaluating methods and metrics. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yikun Pan and Elias Bareinboim. Counterfactual image editing. In *Proceedings of ICML*, volume 235 of *PMLR*, pp. 39087–39101, 2024.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of CVPR*, pp. 10619–10629, 2022.

Rahul Ramesh Rasal, Avinash Kori, Fabio De Sousa Ribeiro, Tian Xia, and Ben Glocker. Diffusion counterfactual generation with semantic abduction. In *Proceedings of ICML*, volume 267 of *PMLR*, pp. 51201–51228, 2025.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you? explaining the predictions of any classifier. In *Proceedings of KDD*, pp. 1135–1144, 2016.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28:2660–2673, 2017.

Ramprasaath R. Selvaraju et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of ICCV*, pp. 618–626, 2017.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of ICLR*, 2021.

Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Anchal Vaidya et al. Demographic bias in misdiagnosis by computational pathology models. *Nature Medicine*, 30:1174–1190, 2024.

Rui Xie, Jun Chen, Lin Jiang, Rui Xiao, Yikun Pan, and Yizhou Cai. Accurate explanation model for image classifiers using class association embedding. In *Proceedings of the IEEE International Conference on Data Engineering*, pp. 2271–2284, 2024.

Rui Xie et al. Bridging the interpretability gap for medical artificial intelligence models using class-association manifold learning. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-026-01676-w.