DARC: Dual Adjustment Reasoning with Counterfactuals for Trustworthy Chest X-ray Classification

Supplementary Material

A. Theoretical Derivations

We provide a complete theoretical derivation for the transformation of the Total Direct Effect (TDE) into the robust, additive prediction score presented in Equation (5) of the main text. Our derivation relies on a mild linear response assumption within the feature space, which we will formally define and justify.

A.1. Preliminaries: TDE and Counterfactuals

We begin by re-defining the objective. Our goal is to isolate the causal effect of a target pathology X on the model's prediction Y, disentangled from the confounding effect of a co-occurring pathology Z. As established in our Structural Causal Model (SCM), this co-occurrence is generated by an unobservable common cause U via the back-door path $X \leftarrow U \rightarrow Z \rightarrow F \rightarrow Y$. Due to the unobservability of U, we turn to counterfactual reasoning to estimate the Total Direct Effect (TDE) [1, 2].

The TDE is formally defined as the difference between two potential outcomes under a counterfactual scenario:

$$\text{TDE} \triangleq \mathbb{E}[Y_{x,Z=z}] - \mathbb{E}[Y_{x_0,Z=z}], \tag{S1}$$

where $Y_{x,Z=z}$ denotes the potential outcome of Y had X been set to x while Z was observed to be z. The first term, $\mathbb{E}[Y_{x,Z=z}]$, corresponds to the factual (observational) prediction, which can be estimated from data using $P(Y = 1|X = x, Z = z)$. The second term, $\mathbb{E}[Y_{x_0,Z=z}]$, is the counterfactual quantity of primary interest. It quantifies the prediction the model would make if, hypothetically, the target pathology X were absent (intervened to x_0), while the co-occurring pathology Z remained present. This term represents the bias introduced by the co-occurrence, as it measures the model's tendency to predict Y based solely on the presence of Z.

Directly computing this counterfactual term is intractable in deep learning models. Our goal is therefore to find a computable surrogate for the TDE that is monotonically related to it, thereby making it suitable for optimization. This requires decomposing the counterfactual bias term $\mathbb{E}[Y_{x_0,Z=z}]$ into observable or approximable quantities.

A.2. Linear Response Assumption and a Key Lemma

To proceed, we introduce a mild and plausible assumption regarding the model's behavior in the high-level feature space. Let $\phi(x, z)$ be the feature representation of the image right before the final classifier, where x and z denote the visual features corresponding to pathologies X and Z, respectively. We assume the final classifier is a linear layer (or its behavior can be locally approximated by one), such that the logit output is approximately $\text{logit}(Y) \approx \mathbf{w}^T\phi(x, z)+b$.

Assumption A.1 (Linear Response in Feature Space)**.** *Let $\phi(x_0, z)$ be the feature representation when pathology X is absent, and $\phi(x, z)$ be the representation when both X and Z are present. We assume that the feature representation of a pure pathology X in the absence of Z's visual influence, denoted as $\phi(x, z_0)$, can be related to the others through a linear combination. Specifically, the feature perturbation caused by removing X is approximately proportional to the pure feature signal of X itself. Formally, for some coefficients α, β that depend on the co-occurrence statistics but are constant for a given pair (X, Z):*

$$\phi(x_0, z) \approx \alpha \cdot \phi(x, z) - \beta \cdot \phi(x, z_0). \tag{S2}$$

Justification. This assumption posits that the feature vector for a co-occurring scene $\phi(x, z)$ can be thought of as a superposition of the pure signals $\phi(x, z_0)$ (X alone) and $\phi(x_0, z)$ (Z alone), though not necessarily a simple sum. The assumption states that the features of "Z alone" ($\phi(x_0, z)$) can be approximated by taking the features of "X and Z together" ($\phi(x, z)$) and subtracting a scaled version of the features of "X alone" ($\phi(x, z_0)$). This is a reasonable first-order approximation in high-dimensional vector spaces, where feature directions corresponding to different semantic concepts (such as different pathologies) tend to be separable.

Based on this assumption, we can now state a key lemma that connects the counterfactual probability to observable and intervenable probabilities.

Lemma A.1. *Under the Linear Response Assumption (Assumption A.1) and the assumption that the prediction probability $P(Y = 1|\cdot)$ is a smooth function of the feature representation ϕ (such as one given by a sigmoid function on a linear mapping of ϕ), the counterfactual bias term can be approximated as:*

$$P(Y = 1|X = x_0, Z = z) \approx \alpha' \cdot P(Y = 1|X = x, Z = z) - \beta' \cdot P(Y = 1|X = x, do(Z = z_0)), \tag{S3}$$

where α' and β' are effective coefficients that absorb the parameters of the final classifier and are treated as constants for a given co-occurrence relationship.

Proof. Let $P(Y = 1|\phi) = \sigma(\mathbf{w}^T\phi + b)$, where σ is the sigmoid function. We are interested in $P(Y = 1|X = x_0, Z = z)$, which is equivalent to $\sigma(\mathbf{w}^T\phi(x_0, z) + b)$.

Using Assumption A.1, we substitute the expression for $\phi(x_0, z)$:

$$P(Y = 1|X = x_0, Z = z) \approx \sigma\Big(\mathbf{w}^T(\alpha \cdot \phi(x, z) - \beta \cdot \phi(x, z_0)) + b\Big). \tag{S4}$$

The non-linearity of the sigmoid function complicates a direct decomposition. However, for optimization purposes, a monotonic relationship between quantities is often sufficient. We consider the logits directly, as the probability is a monotonic function of the logit. Let $L(\phi) = \mathbf{w}^T\phi + b$ be the logit function. Applying this linear function to Assumption A.1, we obtain:

$$\begin{aligned} L(\phi(x_0, z)) &\approx \mathbf{w}^T(\alpha \cdot \phi(x, z) - \beta \cdot \phi(x, z_0)) + b \\ &= \alpha(\mathbf{w}^T\phi(x, z) + b) - \beta(\mathbf{w}^T\phi(x, z_0) + b) \\ &\quad + (1 - \alpha + \beta)b \\ &\approx \alpha \cdot L(\phi(x, z)) - \beta \cdot L(\phi(x, z_0)) + c, \end{aligned} \tag{S5}$$

where c is a new constant bias term. Since probability is a monotonic transformation of the logit, a linear relationship between logits suggests an approximately linear relationship between the probabilities themselves within a typical operating range, or at least a relationship that preserves the ordering for optimization. By absorbing the scaling and shifting effects into new effective coefficients α' and β', we arrive at the form stated in Lemma A.1 (Eq. S3). The term $P(Y = 1|X = x, \text{do}(Z = z_0))$ is the neural network's estimation of the probability given the feature representation $\phi(x, z_0)$, which corresponds to an intervention on the visual features of Z. This completes the proof. □

A.3. Derivation of the Additive Debiased Prediction Score

With Lemma A.1 established, we can now substitute the decomposition of the counterfactual bias term back into the original TDE definition to derive our final, optimizable score.

Theorem A.1. *Under the conditions of Lemma A.1, the Total Direct Effect (TDE) is monotonically related to an additive prediction score, $\text{Score}_D(Y)$, given by:*

$$\begin{aligned} \text{Score}_D(Y) \propto\ & P(Y = 1|X = x, Z = z) \\ & + \lambda \cdot P(Y = 1|X = x, do(Z = z_0)), \end{aligned} \tag{S6}$$

where $\lambda = \beta'/(1 - \alpha')$ is a non-negative, data-dependent constant that balances the observational and counterfactual terms.

Proof. We start with the definition of TDE from Eq. (S1):

$$\begin{aligned} \text{TDE} = & P(Y = 1|X = x, Z = z) \\ & - P(Y = 1|X = x_0, Z = z). \end{aligned} \tag{S7}$$

Now, we apply Lemma A.1 to replace the second term, which represents the co-occurrence bias. For notational simplicity, we let $P(Y|A)$ denote $P(Y = 1|A)$ in the following steps:

$$\begin{aligned} \text{TDE} &\approx P(Y|x, z) - [\alpha' \cdot P(Y|x, z) - \beta' \cdot P(Y|x, \text{do}(z_0))] \\ &= P(Y|x, z) - \alpha' \cdot P(Y|x, z) + \beta' \cdot P(Y|x, \text{do}(z_0)). \end{aligned} \tag{S8}$$

Combining the terms related to the observational probability $P(Y|x, z)$:

$$\text{TDE} \approx (1 - \alpha') \cdot P(Y|x, z) + \beta' \cdot P(Y|x, \text{do}(z_0)). \tag{S9}$$

Eq. (S9) provides a computable expression for TDE based on two quantities our neural network can estimate: the standard observational prediction $P(Y|x, z)$ and the counterfactual prediction $P(Y|x, \text{do}(z_0))$.

For the purpose of optimization via gradient-based methods, we are interested in a function that is monotonically related to the TDE. The coefficients $(1 - \alpha')$ and β' are treated as constants for a given disease pair. Assuming $(1 - \alpha') > 0$, which is plausible as α' represents a proportional contribution and should thus be less than 1, we can divide the entire expression by this constant without changing the optimal solution for the model's parameters:

$$\frac{\text{TDE}}{1 - \alpha'} \approx P(Y|x, z) + \frac{\beta'}{1 - \alpha'} \cdot P(Y|x, \text{do}(z_0)). \tag{S10}$$

We now define the hyperparameter $\lambda \triangleq \frac{\beta'}{1-\alpha'}$. Since α' and β' are derived from the feature statistics and are expected to be non-negative, λ is a non-negative constant. This leads us to the final debiased prediction score, $\text{Score}_D(Y)$, which is proportional to the TDE:

$$\text{Score}_D(Y) \propto P(Y|x, z) + \lambda \cdot P(Y|x, \text{do}(z_0)). \tag{S11}$$

This completes the proof. □

B. More Details on Algorithms and Architecture

B.1. Detailed Algorithmic Pseudocode

While Algorithms 1 and 2 in the main text offer a high-level overview, Algorithm B.1 below presents more detailed,

PyTorch-style pseudocode for the Confounder-Aware Feature Extraction and Conditioned Prediction steps within the Global Stream. This detailed view clarifies the specific tensor manipulations and dimensional changes involved in the back-door adjustment process.

Algorithm B.1: Detailed Implementation of Steps 3 & 4 in Global Stream

Input: Global feature map $\mathbf{F}_{\text{map}} \in \mathbb{R}^{B\times D_g\times H'\times W'}$;
Confounder mask $\mathbf{M}_{\text{conf}} \in \mathbb{R}^{B\times N_c\times H\times W}$;
Global feature vector $\mathbf{f}_{\text{vec}} \in \mathbb{R}^{B\times D_g}$.
Output: Adjusted global feature $\mathbf{f}'_{\text{global}} \in \mathbb{R}^{B\times D_g}$.

▷ Step 3: Confounder-Aware Feature Extraction

$\mathbf{M}'_{\text{conf}} \leftarrow \text{Interpolate}(\mathbf{M}_{\text{conf}}, \text{size} = (H', W'))$
▷ Shape: $B \times N_c \times H' \times W'$

$\text{mask\_sum} \leftarrow \mathbf{M}'_{\text{conf}}.\text{sum}([-1, -2])$ ▷ Shape: $B \times N_c$

$\tilde{\mathbf{M}}' \leftarrow \mathbf{M}'_{\text{conf}}/(\text{mask\_sum.unsqueeze}(-1).\text{unsqueeze}(-1)+\epsilon)$
▷ Spatially normalize the mask

$\mathbf{h}_{\text{conf}} \leftarrow \text{einsum}('bchw, bkhw \rightarrow bkc', \mathbf{F}_{\text{map}}, \tilde{\mathbf{M}}')$
▷ Weighted pooling, Shape: $B \times N_c \times D_g$

$\mathbf{E}_{\text{conf}} \leftarrow W_{\text{proj\_conf}}(\mathbf{h}_{\text{conf}})$
▷ Project embeddings, Shape: $B \times N_c \times D_{conf}$

▷ Step 4: Conditioned Prediction via FiLM

$\mathbf{p}_{\text{conf}} \leftarrow \text{Softmax}(\text{mask\_sum})$
▷ Normalized confounder weights, Shape: $B \times N_c$

$\mathbf{e}_{\text{agg}} \leftarrow (\mathbf{E}_{\text{conf}} \odot \mathbf{p}_{\text{conf}}.\text{unsqueeze}(-1)).\text{sum}(1)$
▷ Softmax-weighted aggregation of per-class confounder embeddings

$[\gamma, \beta] \leftarrow W_{\text{film}}(\mathbf{e}_{\text{agg}})$
▷ Map aggregated confounder state to FiLM parameters

$\mathbf{f}'_{\text{global}} \leftarrow (1 + \gamma) \odot \mathbf{f}_{\text{vec}} + \beta$

return $\mathbf{f}'_{global}$

B.2. Network Architecture Parameters

The Global Stream's Confounder Handler module processes the $D_g = 1536$ dimensional global feature map. Its FiLM Layer consists of a single 'nn.Linear' layer that maps the aggregated confounder embedding $\mathbf{e}_{\text{agg}} \in \mathbb{R}^{B\times 256}$ to a $2 \times D_g = 3072$-dimensional vector, which is then reshaped into the modulation parameters $\gamma, \beta \in \mathbb{R}^{B\times 1536}$ without any intermediate non-linear activation.

The Local Stream employs a pre-trained Anatomical Landmark Detector based on the YOLOv8 architecture to identify six key anatomical points. The subsequent Anatomic-Causal Attention module is implemented as a single-head attention mechanism. The embeddings of the learnable disease queries, $\mathbf{Q}_{\text{disease}} \in \mathbb{R}^{14\times 256}$, are projected by a linear layer ($W_q : 256 \rightarrow 384$) to match the dimension of the keys. The keys and values are derived from the patch features, $\mathbf{M}_{\text{local}} \in \mathbb{R}^{6\times 384}$ (from ConvNeXt stage 1), augmented with a learnable positional embedding, $\mathbf{P}_{\text{pos}} \in \mathbb{R}^{6\times 384}$, before use, with no further projection.

The Causal Fusion Module receives the concatenated features of size $D_g + D_p = 1536 + 384 = 1920$. The MLP fusion, f_{fuse}, is a two-layer perceptron. The first 'nn.Linear' layer maps the 1920-dimensional input to a 512-dimensional hidden representation, followed by a GELU activation and a Dropout layer (p=0.1). A residual connection with a linear projection is used to handle the dimension mismatch between the input and output. The Classification Head is a final 'nn.Linear' layer that maps the 512-dimensional fused features to the 14-class output logits.

B.3. Upstream Modules: Training Sources, Performance, and Robustness

To prevent information leakage, the Anatomical Landmark Detector (A.L.D.) is trained only on an independent public dataset and is never fine-tuned on ChestX-ray14, CheXpert, or CheXconf. On its held-out evaluation set, A.L.D. achieves a mean radial error (MRE) of 9.22 pixels. The Confounder Segmentation module (C.S.) is implemented by fine-tuning SAM on CheXconf and achieves 87.3% mIoU on a held-out evaluation split.

In the Local Stream, we crop $s\times s$ patches centered at the detected anatomical landmarks and encode them with a shared ConvNeXt stage-1 feature extractor. In the Global Stream, the class-wise confounder masks predicted by C.S. are used for confounder-aware conditioning. These two upstream modules therefore provide the spatial priors required by the dual-stream debiasing process, but are not themselves optimized on the downstream test benchmarks.

To assess whether DARC is overly dependent on perfect upstream priors, we further perturb the detected landmarks with Gaussian jitter before local patch extraction, and corrupt the predicted confounder masks with random morphological operations before Global Stream conditioning. As shown in Figure S1, DARC remains stable under mild-to-moderate perturbations and degrades noticeably only under unrealistically large upstream errors. This indicates that the downstream gains of DARC do not rely on perfect upstream predictions, but arise from the debiasing mechanism itself.

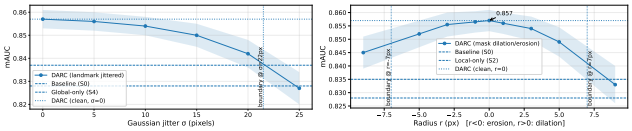


Figure S1. Upstream components sensitivity analysis.

C. More Details on the CheXconf Dataset

C.1. Confounder Class Definitions and Visual Examples

The CheXconf dataset comprises 11 distinct classes of non-pathological visual confounders commonly encountered in clinical practice. The selection of these classes was finalized in collaboration with two senior radiologists to ensure clinical relevance and identifiability. Each class was defined to be visually distinct to minimize ambiguity during annotation. Table S1 provides the definition for each class, and Figure S2 presents representative visual examples.

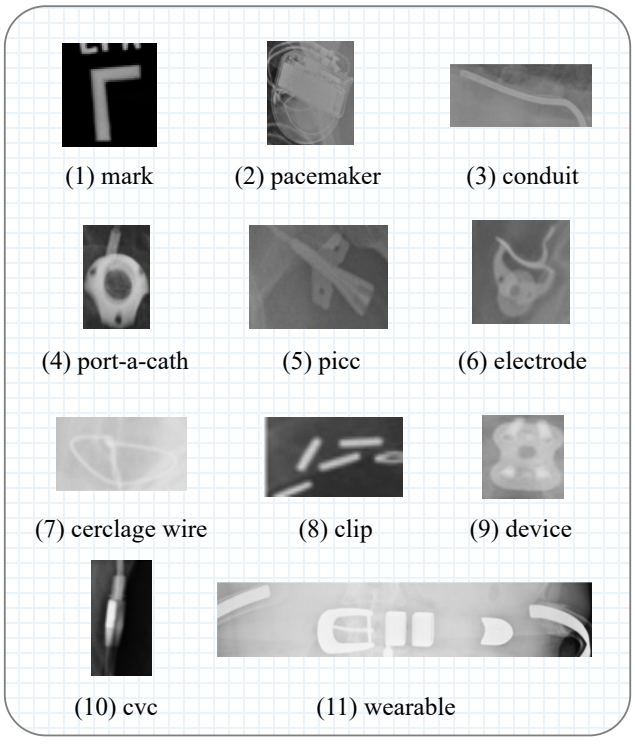


Figure S2. Visual examples of non-pathological visual confounders in CheXconf.

Class	Definition and Scope
Mark	External radiopaque markers placed on the patient's skin to indicate specific locations.
Pacemaker	Implanted electronic device to regulate heartbeat, typically visible in the upper left chest. The annotation includes the device, leads, and generator.
Conduit	Tubes or pipes used for drainage or access, such as chest tubes.
Port-a-Cath	Implanted venous access device, usually in the upper chest, consisting of a portal and a catheter.
PICC	Peripherally Inserted Central Catheter; a long, thin tube inserted into a peripheral vein and advanced to a central vein.
Electrode	Monitoring electrodes (e.g., for ECG) attached to the skin, including pads and lead wires visible on the X-ray.
Cerclage Wire	Wires used to hold fractured bones together, commonly seen after sternotomy or rib fracture repair.
Clip	Surgical clips used for hemostasis or marking, typically appearing as small, dense metallic objects.
Device	A general category for other implanted medical devices not covered by the other classes (e.g., valve replacements, orthopedic hardware).
CVC	Central Venous Catheter; a catheter placed into a large vein, distinct from PICCs by its insertion site (e.g., subclavian, jugular).
Wearable	Items worn by the patient that are visible on the X-ray, such as jewelry, clothing components (e.g., zippers), or external medical sensors.

Table S1. Definitions of the 11 non-pathological confounder classes in CheXconf.

C.2. Annotation Protocol and Quality Control

To ensure high-quality and consistent annotations, we established a rigorous protocol.

C.2.1. Annotation Process

The annotations were performed by a team of five annotators who underwent dedicated training led by our consulting radiologists. The process for each image was as follows: **(i) Initial Identification.** Annotators first scanned the image to identify all potential confounder instances based on the definitions in Table S1. **(ii) Bounding Box Localization.** A rectangular bounding box was initially drawn to loosely enclose each identified instance. **(iii) Pixel-level Delineation.** Using a semi-automated segmentation tool, annotators care-

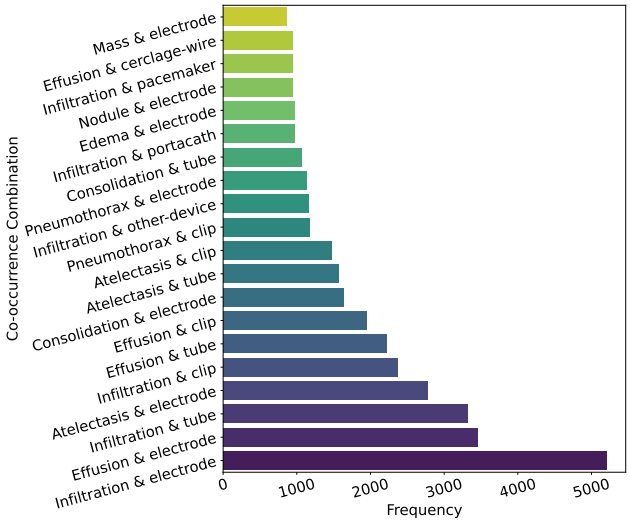


Figure S3. Frequency distribution of the top-20 most common disease-confounder pairs. The high frequency of certain pairs underscores the prevalence of potential shortcut learning opportunities in the dataset.

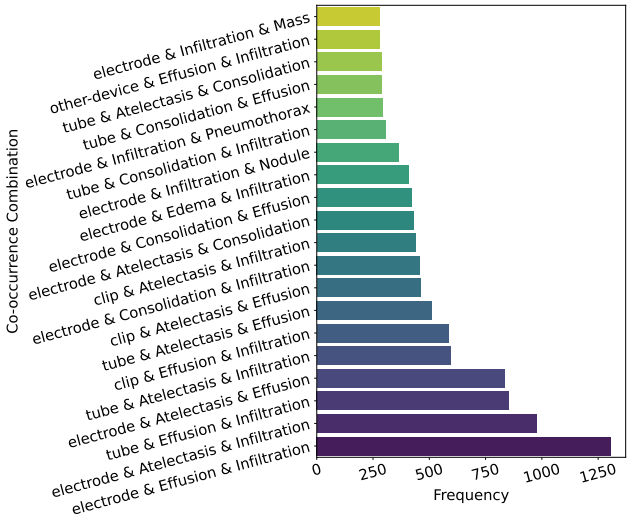


Figure S4. Frequency distribution of the top-20 most frequent confounding triplets. The existence of these complex, intertwined statistical relationships highlights the need for a comprehensive debiasing approach.

fully traced the precise contour of each instance to generate a pixel-level mask. Special attention was paid to boundaries, especially for objects like catheters and wires.

C.2.2. Quality Control

We implemented a two-round cross-review mechanism to maximize annotation accuracy: **(i) Peer Review.** Each annotated image was reviewed by a second annotator. Any disagreements were discussed and resolved by that pair. **(ii) Radiologist Audit.** A random sample of 10% of the annotated data from each batch was audited by one of the senior radiologists. If a batch's average Intersection over Union (IoU) with the radiologist's annotations fell below a threshold of 0.90, the entire batch was returned for re-annotation. This iterative process ensured the final dataset achieved a high degree of clinical accuracy.

C.3. Statistical Analysis

The final CheXconf dataset contains 40,213 annotated confounder instances across 10,000 CXR images sourced from the ChestX-ray14 dataset. Our statistical analysis focuses on quantifying the complex interplay between diseases, non-pathological confounders, and co-occurring pathologies, which forms the empirical basis for our dual adjustment strategy.

C.3.1. Disease-Confounder Pair Statistics

A primary motivation for this work is the spurious correlation between non-pathological confounders and disease labels. To quantify this, we analyzed the co-occurrence of disease-confounder pairs. Figure S3 displays the frequencies of the top-20 most common disease-confounder pairs found in our dataset. A prominent example is the pair (Infiltration, electrode), which appears with high frequency. This strong statistical link can mislead a model into learning the electrode's visual features as a proxy for diagnosing Infiltration, demonstrating a clear shortcut learning pathway. Similarly, pairs like (Effusion, Cerclage Wire) highlight correlations arising from post-surgical contexts.

C.3.2. Analysis of Confounding Triplets

The challenges in CXR classification are often more complex than simple pairwise correlations. Spurious associations can be amplified by the presence of a third, co-occurring pathology. To investigate this, we analyzed the statistics of "confounding triplets," defined as (Target Disease, Co-occurring Pathology, Confounder). These triplets represent scenarios where both of the confounding mechanisms addressed by DARC—pathological co-occurrence and non-pathological confounders—are present simultaneously.

Figure S4 presents the distribution of the top-20 most frequent confounding triplets. For instance, the triplet (Effusion, Infiltration, electrode) is highly prevalent. In this scenario, a model might not only use the electrode as a shortcut for Effusion but could also incorrectly associate the electrode with Infiltration due to the strong co-occurrence between the two diseases. This complex entanglement underscores the necessity of a framework like DARC, which can decouple both types of confounding sources simultaneously—a task that models addressing only a single type of confounder cannot achieve. This analysis validates the

design of our dual-stream architecture.

D. Additional Experiments

D.1. Hyperparameter Sensitivity Study

To validate the robustness of our DARC framework and to provide justification for our parameter choices, we conducted a sensitivity analysis on several key hyperparameters that directly influence the causal decoupling mechanisms. We investigated the impact of the local image patch size in the Local Stream, and the dimensionality of the confounder embeddings in the Global Stream.

D.1.1. Local Image Patch Size

The Local Stream's ability to approximate a counterfactual intervention, $P(Y|X, \text{do}(Z = z_0))$, hinges on its capacity to extract pure pathological features from localized regions, effectively "masking" the influence of co-occurring pathologies. The "Patch Size" hyperparameter directly controls the granularity of this spatial decoupling. An excessively small patch might fail to capture the complete morphology of a pathological feature, leading to an incomplete representation. Conversely, an overly large patch risks reintroducing the very co-occurrence bias we aim to eliminate, as it might encompass features from both the target and co-occurring pathologies, thereby violating the principle of the intervention.

To identify the optimal trade-off, we evaluated the model's performance (mAUC) on the ChestX-ray14 dataset while varying the Patch Size across a range of $\{16, 32, 64, 128\}$, keeping all other parameters fixed. The results, plotted in Figure S5(a), demonstrated a clear trend. The performance peaked at a patch size of 64, achieving an mAUC of 0.857. Both smaller and larger patch sizes resulted in a noticeable degradation in performance. This empirically validates our hypothesis that a moderately sized receptive field is optimal, providing sufficient context to identify local pathologies while minimizing the risk of capturing confounding signals from adjacent anatomical regions.

D.1.2. Confounder Embedding Dimension

The efficacy of the back-door adjustment in the Global Stream is contingent upon the representational capacity of the confounder embedding vector, $\mathbf{e}_{\text{conf\_agg}}$. This vector must encode sufficient information about the presence and nature of non-pathological confounders (Z') to effectively modulate the global feature vector via the FiLM layer. The "Confounder Embedding Dimension", which defines the dimensionality of this embedding, is therefore a critical hyperparameter. An insufficient dimension might lead to information bottlenecks, preventing the model from fully capturing the complex visual variations of confounders. Conversely, an excessively high dimension could lead to overparameterization, making the model susceptible to fitting spurious noise present in the confounder regions.

We analyzed the model's performance as a function of the "Confounder Embedding Dimension", testing values in the set $\{64, 128, 256, 512, 1024\}$. As illustrated in Figure S5(b), performance steadily improved as the dimension increased from 64 to 256, where it reached its peak mAUC of 0.857. This suggests that a 256-dimensional space provides adequate capacity to represent the 11 defined confounder classes and their visual characteristics. Beyond this point, increasing the dimension to 512 and 1024 yielded no further performance gains and introduced a slight decline, likely due to the model beginning to overfit on the confounder features. This analysis justifies our choice of 256 as an effective and efficient embedding dimension for confounder decoupling.

D.1.3. Causal Loss Weights

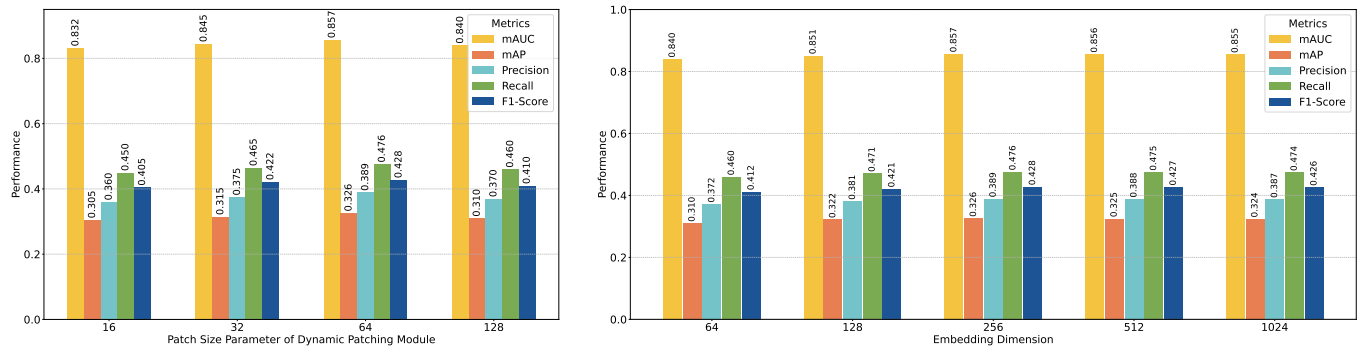
To investigate the impact of the two causal regularization loss terms, we fixed the main loss weight $w_1 = 1.0$ and conducted experiments over a range of values for w_2 and w_3. As shown in Figure S6, introducing these two loss terms substantially improved model performance. Performance peaked around $w_2 = 0.2$ and $w_3 = 0.05$. However, excessively large weights could interfere with the optimization of the main classification task, leading to a decline in performance. This analysis not only validates the effectiveness of the two causal loss terms but also provides a basis for selecting the optimal hyperparameters, demonstrating the importance of balancing the main classification objective and causal regularization.

D.2. Additional Qualitative Analysis with Grad-CAM

To further demonstrate the improved interpretability and causal fidelity of our DARC framework, we present additional qualitative results using Grad-CAM [3]. Following the analysis in Section 4.4 of the main paper, Figure S7 showcases a wider range of challenging clinical cases, comparing the visual attention of the baseline model against our DARC model.

Across diverse pathologies, a consistent pattern emerged: the baseline model was frequently distracted by spurious visual cues, whereas DARC consistently localized the true underlying pathology.

- For **Effusion**, which manifests as blunting of the costophrenic angles, the baseline model's attention was often diffuse, incorrectly highlighting unrelated areas such as the upper lung zones or the cardiac silhouette. In contrast, DARC produced a highly concentrated heatmap precisely over the fluid accumulation in the lung base, demonstrating superior localization.
- A particularly challenging case is **Pleural Thickening**, where the findings can be subtle. The baseline model



(a) Impact of Local Patch Size

(b) Impact of Confounder Embedding Dimension

Figure S5. Hyperparameter sensitivity analysis on the ChestX-ray14 dataset. (a) shows the effect of varying the Patch Size in the Local Stream on the final metrics. (b) shows the effect of varying the Confounder Embedding Dimension in the Global Stream's back-door adjustment module.

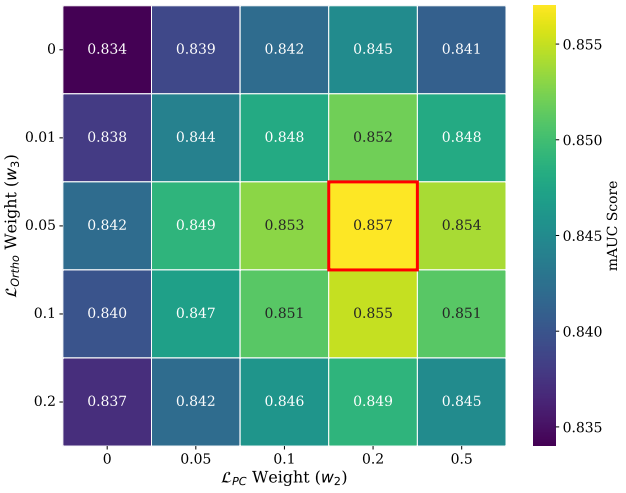


Figure S6. Impact of Causal Loss Weights.

often failed to generate any significant activation, effectively misclassifying the sample as normal. DARC, however, successfully identified the affected pleural regions, and its heatmap provided more comprehensive coverage of the thickened areas, indicating a higher sensitivity to fine-grained pathological features.

- When diagnosing **Edema**, characterized by widespread interstitial opacities, DARC's activation map correctly covered the full extent of the pathology across both lungs. The baseline model, while sometimes activating on parts of the edema, often produced a fragmented or incomplete heatmap, failing to capture the global nature of the condition.
- For **Hernia**, our analysis revealed DARC's strong robustness to confounders. The baseline model was significantly distracted by a wearable item on the patient's neck, producing anomalous activation in the head and neck region, completely irrelevant to the actual diaphragmatic hernia. DARC entirely ignored this confounder and correctly focused its attention on the herniated structure in the lower thoracic cavity.

Collectively, these additional examples reinforce our claim that DARC learns to ground its predictions in genuine pathological evidence, systematically ignoring both co-occurring pathologies and non-pathological artifacts. This leads to a more robust and trustworthy model for clinical application.

D.3. Extended Robustness Evaluation with t-SNE

To supplement the t-SNE analysis in the main paper, we conducted our confounder attack experiment on four additional, clinically relevant spurious correlations. This comprehensive evaluation served to validate the generalizability of DARC's robustness against a diverse range of non-pathological confounders. The experimental setup remains identical for each case: we compare the feature distributions of True Positive (TP), clean True Negative (TN), and Confounded Negative (CN) samples.

The four selected cases represent distinct and challenging types of confounding mechanisms:

1. **Surgical History Confounding (cerclage-wire → Atelectasis):** This case tests the model's ability to avoid shortcuts based on a patient's surgical history. As shown in Figure S8(a), the baseline model's feature space showed a significant overlap between CN samples (clean images with a synthetic wire) and TP samples (images with Atelectasis), confirming it has learned to associate surgical hardware with post-operative complications. DARC's feature space, however, showed a clear separation, with CN samples clustering tightly with TN samples, demonstrating its immunity to this bias.
2. **ICU Scenario Confounding (CVC → Infiltration):** This scenario mimics the intensive care unit (ICU) set-

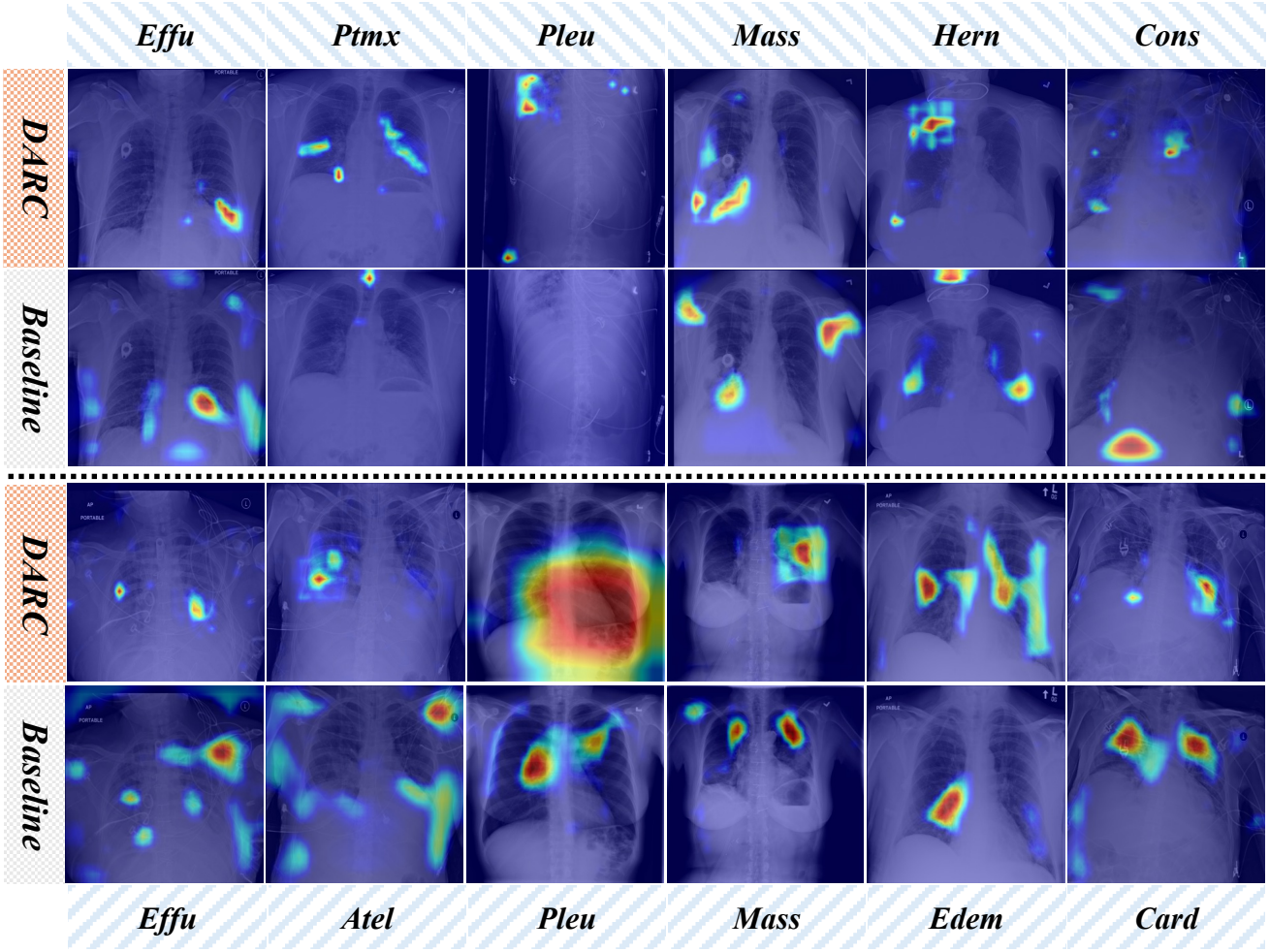


Figure S7. Additional Grad-CAM visualizations for various pathologies. For each case, we compare the attention maps of the baseline model with our DARC model. The results consistently show DARC's superior ability to focus on true pathological regions while ignoring spurious correlations.

ting, where medical devices are prevalent. Central Venous Catheters (CVCs) are common in critically ill patients, who are also at high risk for Infiltration. The baseline model, as seen in Figure S8(b), exploited this strong statistical link, confusing images with a CVC for those with actual infiltration. DARC again successfully disentangled these signals, focusing on the lung parenchyma instead of the device.

3. **External Artifact Confounding (mark → Nodule):** This evaluates the model's robustness against external objects that can be mistaken for internal pathologies. Figure S8(c) illustrates the model's response to skin markers. The baseline model was again confused, with its CN features drifting towards the TP cluster, indicating it misinterprets the external marker as a potential internal nodule. DARC successfully avoided this pitfall, grouping the CN and TN samples together and far from the TP cluster.
4. **Morphological Similarity Confounding (Port-a-Cath → Mass):** This is a particularly challenging case due to the high visual similarity between a port-a-cath's reservoir and a small mass. The baseline model failed spectacularly (Figure S8(d)), with its CN and TP feature clusters being almost indistinguishable. In stark contrast, DARC maintained a well-structured feature space where the confounder's presence did not mislead the classification, proving its advanced capability to learn intrinsic, rather than superficial, features.

In summary, these extended t-SNE analyses, covering a wide spectrum of confounding types, provide compelling evidence that DARC's robustness is not an isolated phenomenon but a systemic property of the framework. It can effectively mitigate shortcut learning, thereby enhancing its reliability for real-world deployment.

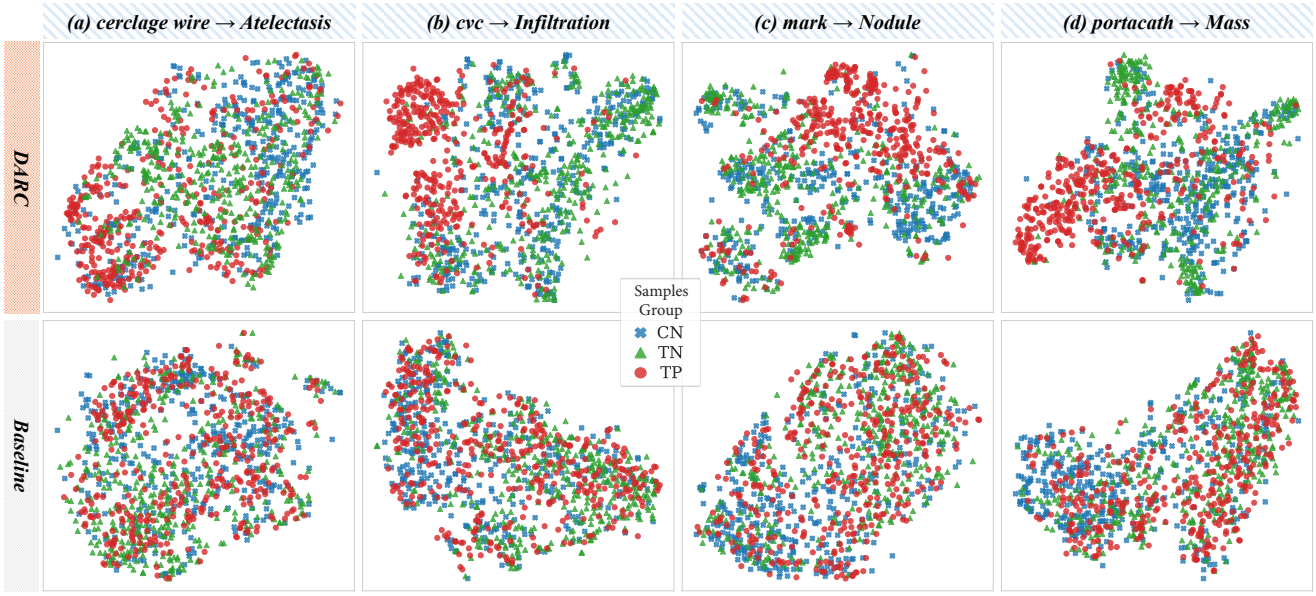


Figure S8. Extended t-SNE visualization of feature embeddings under four different confounder attacks: (a) cerclage-wire → Atelectasis, (b) CVC → Infiltration, (c) mark → Nodule, and (d) Port-a-Cath → Mass. In all scenarios, the baseline model is confused by the confounder (CN samples, in blue, overlap with TP samples, in red), while DARC successfully groups CN samples with TN samples (in green), demonstrating robust feature disentanglement.

References

[1] J. Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc., 2001. 1

[2] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018. 1

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6