

FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO
FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Synthetic Data Generation for Concept-based Explainable AI

Guilherme Sabino Vaz



Master's Degree in Artificial Intelligence

Supervisor: Prof. Luís F. Teixeira

Second Supervisor: Prof. Wilson Silva

July 31, 2026

Synthetic Data Generation for Concept-based Explainable AI

Guilherme Sabino Vaz

Master's Degree in Artificial Intelligence

Approved in oral examination by the committee:

President: Prof. Henrique Lopes Cardoso

Examiner: Prof. Cristiano Patricio

Member: Prof. Luís Filipe Teixeira

July 31, 2026

Resumo

Os modelos de *deep learning* para imagem médica alcançam um forte desempenho preditivo, mas permanecem opacos, e os métodos de explicabilidade baseada em conceitos que poderiam torná-los transparentes são limitados por dois obstáculos: a escassez de anotações ao nível do conceito e o entrelaçamento dos conceitos clínicos com correlações espúrias em dados observacionais. Esta dissertação investiga se modelos generativos podem fornecer dados sintéticos controlados por conceito para mitigar ambos os obstáculos.

Testamos um *pipeline* de injeção de conceito sem máscara construído sobre um modelo Stable Diffusion congelado, que combina Textual Inversion com Null-Text Inversion: um *concept token* aprendido é injetado na inversão de uma imagem saudável de modo a inserir um conceito patológico isolado, procurando preservar a anatomia circundante, sem fazer *fine-tuning* do *backbone* e sem máscara manual na inferência. Estudamos três estratégias para aprender o *concept token*: um *prompt* escrito à mão, um *token* treinado em *patches*, e um *token* treinado em imagens inteiras com uma *loss* restrita à máscara; e comparamo-los com *baselines* de *copy-paste* e de *inpainting* por difusão.

Para avaliar os dados sintéticos para além do realismo visual, desenvolvemos um *framework* de avaliação multi-eixo que mede a fidelidade perceptual (FID), a validade semântica (Target Class Accuracy), a aprendibilidade do conceito (*concept learnability*) relativamente a um limite superior obtido com dados reais, e a robustez a jusante a correlações espúrias através de regularização baseada em conceitos. O *pipeline* é avaliado em três modalidades de imagem: fotografia retiniana, dermoscopia e ecografia mamária.

Nos três conjuntos de dados, a injeção de conceito só é bem-sucedida quando a modalidade de imagem está próxima do *backbone* e o conceito-alvo é saliente. Na fotografia retiniana, o *token* de *loss* mascarada produz imagens conceito-positivas cuja aprendibilidade iguala a dos dados reais; na dermoscopia, uma edição mínima apenas iguala uma *baseline* de *copy-paste*; e na ecografia mamária a inversão degrada-se e a qualidade da imagem colapsa sob *guidance* forte. O *token* de *loss* mascarada é a mais forte das três estratégias onde o método funciona, e um ponto de operação de edição mínima é o *default* seguro. O teste de robustez a jusante revelou-se estatisticamente sem poder (*underpowered*) e é reportado como inconclusivo. No conjunto, o estudo delimita um envelope de operação candidato: uma inversão fiel em conjunto com um conceito saliente e separável, em vez de uma capacidade universal.

Abstract

Deep learning models for medical imaging achieve strong predictive performance yet remain opaque, and the concept-based explainability methods that could make them transparent are held back by two obstacles: the scarcity of concept-level annotations and the entanglement of clinical concepts with spurious correlations in observational data. This thesis investigates whether generative models can supply concept-controlled synthetic data to alleviate both obstacles.

We test a mask-free concept-injection pipeline built on a frozen Stable Diffusion model that combines Textual Inversion with Null-Text Inversion: a learned concept token is injected into the inversion of a healthy image so as to insert an isolated pathological concept while aiming to preserve the surrounding anatomy, without fine-tuning the backbone and without a manual mask at inference. We study three strategies for learning the concept token: a hand-written prompt, a patch-trained token, and a token trained on whole images with a mask-restricted loss, and benchmark them against copy-paste and diffusion-inpainting baselines.

To assess the synthetic data beyond visual realism, we develop a multi-axis evaluation framework that measures perceptual fidelity (FID), semantic validity (Target Class Accuracy), concept learnability relative to a real-data upper bound and downstream robustness to spurious correlations via concept-based regularisation. The pipeline is evaluated on three imaging modalities: retinal photography, dermoscopy, and breast ultrasound.

Across the three datasets, concept injection succeeds only where the imaging modality is close to the backbone and the target concept is salient. On retinal photography the masked-loss token produces concept-positive images whose learnability matches the real-data; on dermoscopy a minimal edit only matches a copy-paste baseline; and on breast ultrasound the inversion degrades and image quality collapses under strong guidance. The masked-loss token is the strongest of the three strategies where the method works, and a minimal-edit operating point is the safe default. The downstream-robustness probe proved statistically underpowered and is reported as inconclusive. Overall, the study delimits a candidate operating envelope: faithful inversion together with a salient, separable concept rather than a universal capability.

Contents

1	Introduction	1
1.1	Context	1
1.2	Motivations	1
1.3	Objectives	2
1.4	Contributions	3
1.5	Document Structure	3
2	Literature Review	4
2.1	From Attribution Maps to Concept-based Reasoning	4
2.1.1	The Limitations of Saliency Methods	4
2.1.2	Concept-based Explainability	6
2.1.3	The Annotation and Entanglement Bottlenecks	8
2.2	Generative Models for Medical Imaging	9
2.2.1	Representation Learning and Disentanglement (VAEs)	9
2.2.2	Adversarial Synthesis (GANs)	10
2.2.3	Diffusion Models: from DDPMs to Latent and Conditional Diffusion	11
2.2.4	Vision Language Models	12
2.2.5	Model Personalisation and Concept Learning	13
2.3	Synthetic Images for Medical XAI	14
2.3.1	Visual Counterfactual Explanations	14
2.3.2	Case-based and Prototype-based Reasoning	15
2.3.3	Concept-Conditional Generation and Editing	16
3	Methodology	19
3.1	Baseline Methods	19
3.1.1	Baseline 1: Lesion Copy-Paste	19
3.1.2	Baseline 2: Diffusion-Based Inpainting	20
3.2	Tested Method: Null-Text Inversion with Concept Injection	20
3.2.1	Overview	20
3.2.2	DDIM Inversion	22
3.2.3	Null-Text Optimisation	24
3.2.4	Concept Token Training	26
3.2.5	Reverse Diffusion with Concept Injection	28
4	Experimental Setup	31
4.1	Datasets	31
4.1.1	IDRiD - Diabetic Retinopathy (Fundus)	32
4.1.2	ISIC 2018 - Dermoscopy	32

4.1.3	BUSI - Breast Ultrasound	32
4.2	Evaluation Pipeline	33
4.2.1	Supporting Semantic Classifier	33
4.2.2	Perceptual Fidelity: Fréchet Inception Distance	34
4.2.3	Semantic Validity: Target Class Accuracy	34
4.2.4	Concept Learnability	35
4.2.5	Robustness via LCRReg	36
4.2.6	Reproducibility	39
4.3	Implementation Details	40
4.4	Concept-Based Regularisation Setup	41
5	Results and Discussion	42
5.1	Reporting Protocol	42
5.2	Per-Dataset Results	43
5.2.1	IDRiD	43
5.2.2	ISIC 2018	44
5.2.3	BUSI	45
5.3	The Guidance–Concept Trade-off	46
5.4	Perceptual Fidelity versus Concept Validity	47
5.5	Concept Alignment and the Limits of a Linear CAV	48
5.6	Edit Localisation	48
5.7	Downstream Robustness via LCRReg	49
5.8	Cross-Dataset Synthesis	50
5.9	Discussion	50
6	Conclusion and Future Work	52
6.1	Summary and main findings	52
6.2	Answering the research questions	53
6.3	Limitations	54
6.4	Future work	54
	References	56
A	Reproducibility Supplement	60
A.1	Source Code	60
A.2	Compute Environment	61
A.3	Run Commands	61
A.4	Datasets	64
A.5	Qualitative Gallery	64
A.6	Qualitative Gallery: Guidance Sweep	65

List of Figures

2.1	Violation of input invariance, after Kindermans et al. Two networks receive the same digit under a prediction-preserving input shift. The shift leaves the prediction unchanged, yet the Gradient $\times$ Input attribution maps differ markedly, reflecting an artefact of the input reference rather than the features driving the decision.	5
3.1	Overview of the proposed pipeline. Stage 1 (left): a concept token v^* is trained via Textual Inversion with masked-loss weighting, using real images that contain the target pathology. Stage 2 (right): a healthy patient image is inverted through DDIM Inversion and Null-Text Optimisation; the learned token is then injected into the reverse diffusion process to synthesise a concept-positive version of the same anatomy.	22
3.2	Null-Text Optimisation procedure[36]. The DDIM inversion trajectory $\{z_0^*, \dots, z_T^*\}$ is computed once and fixed. At each reverse timestep t , the null-text embedding θ_t is optimised so that the guided DDIM step lands on the corresponding inversion target z_{t-1}^* . After optimisation, this full per-timestep set of embeddings $\{\theta_t^*\}$ enables near-exact reconstruction of the original image under Classifier-Free Guidance.	25
3.3	The proposed pipeline on a fundus image. Left: the healthy input. Centre: the concept-injected output (M3), in which hard exudates appear on the retinal background while the vessels and optic disc are left largely unchanged. Right: the absolute difference between the two, showing that the change is concentrated in a few focal regions.	29
5.1	Qualitative comparison on IDRiD: the two baselines (copy-paste and inpainting), and the masked-loss token (M3) across the guidance sweep. As guidance grows the injected exudates become more pronounced while image fidelity degrades (FID 84.7 $\rightarrow$ 194.4), illustrating the guidance–concept trade-off of Table 5.1.	44
5.2	Perceptual fidelity degrades monotonically with the guidance scale on every dataset: stronger concept conditioning makes the synthetic images progressively less realistic. The effect is far steeper on BUSI, where the mean injection FID climbs from ~ 133 at $w = 1.5$ to ~ 363 at $w = 5$, well above the copy-paste realism floor of 127.5 whereas on IDRiD and ISIC, FID rises but stays bounded. Values are the mean over the three injection variants (M1–M3); per-method numbers are in Tables 5.1–5.3.	47
5.3	Edit localisation via difference maps: healthy input (left), the M3 concept-injected counterpart (centre), and their absolute difference (right). Top (IDRiD) : the edit is confined to a few focal regions, consistent with the localisation argument of Section 3.2.5. Bottom (BUSI) : the difference is nearly empty because the concept does not form on ultrasound (Section 5.2.3), not because of a clean localised edit.	49

A.1	IDRiD (example <code>IDRiD_143</code> , injection guidance $w = 3$): the healthy input, its null-text inversion reconstruction, the copy-paste and inpainting baselines, and the three injection variants (M1 prompt, M2 patches, M3 masked loss).	64
A.2	ISIC 2018 (example <code>ISIC_0000040</code> , injection guidance $w = 1.5$): panels as in Figure A.1. The minimal edit ($w = 1.5$) is the best operating point for this concept.	64
A.3	BUSI (example <code>normal_6</code> , injection guidance $w = 1.5$): panels as in Figure A.1. On this modality the concept does not form and stronger guidance drives the output out of distribution.	65
A.4	IDRiD, the masked-loss token (M3) across the guidance sweep. The hard exudates strengthen as guidance grows (recognisable from $w = 3$), while perceptual fidelity degrades (FID $84.7 \rightarrow 256.9$ over the sweep).	65
A.5	ISIC 2018, M3 across the guidance sweep. The faint pigment network is best preserved at the minimal edit ($w = 1.5$); larger guidance overwrites it with generic lesion content and the concept collapses.	65
A.6	BUSI, M3 across the guidance sweep. On this out-of-distribution modality the concept never forms; by $w = 5$ the output has collapsed into out-of-distribution texture (FID ~ 373).	65

List of Tables

4.1	The three evaluation datasets. Each spans a different imaging modality and target concept; class sizes are the staged <code>healthy/diseased</code> counts (for ISIC 2018 the class membership rule is given, as the counts depend on the area threshold).	31
4.2	Generation-pipeline hyperparameters. The backbone is a frozen Stable Diffusion 1.5; the only trainable parameter in M2/M3 is the concept-token embedding.	40
4.3	Evaluation hyperparameters: the supporting/robustness classifier and the LCR robustness protocol (Section 4.2.5).	41
5.1	IDRiD results at $w \in \{1.5, 3, 5\}$. FID lower is better; TCA, CL AUC, clinical AUC higher is better; d is the Cohen’s d of the CAV projection gap (sign indicates orientation); Δ LCRReg is the balanced LCRReg delta (mean $\pm$ std over 5 seeds). CL upper bound AUC = 1.000; real-image TCA 79.6%.	44
5.2	ISIC 2018 results at $w \in \{1.5, 3, 5\}$. CL upper bound AUC = 0.846 (the lowest of the three datasets); real-image TCA 85.0%. The <i>real</i> -row CL AUC (0.915) is the in-sample value and exceeds this leakage-free upper bound (see Section 5.1). The clinical AUC is anti-oriented even for the real concept (0.425) and is not used to rank methods (Section 5.5).	45
5.3	BUSI results at $w \in \{1.5, 3, 5\}$. CL upper bound AUC = 0.986; real-image TCA 98.9% (inflated by the 647:133 class imbalance).	46
5.4	Cross-dataset summary. “Concept” columns use the Concept Learnability upper bound and the real-concept clinical AUC. The two-prerequisite reading is an organising hypothesis ($n = 3$, confounded).	50
A.1	Mapping of thesis components to source files across the two repositories. The CFG injection shared by M1–M3 and the NTI / training math live in <code>expl/utls.py</code> of the generation repository.	61
A.2	Datasets used in this work, with the target concept and source.	64

List of Acronyms

AUC	Area Under the (ROC) Curve
CAV	Concept Activation Vector
CBM	Concept Bottleneck Model
CFG	Classifier-Free Guidance
CL	Concept Learnability
CT	Computed Tomography
DDIM	Denoising Diffusion Implicit Model
DDPM	Denoising Diffusion Probabilistic Model
DL	Deep Learning
FID	Fréchet Inception Distance
GAN	Generative Adversarial Network
LDM	Latent Diffusion Model
NTI	Null-Text Inversion
PSNR	Peak Signal-to-Noise Ratio
RRR	Right for the Right Reasons
SD	Stable Diffusion
SSIM	Structural Similarity Index Measure
SVM	Support Vector Machine
TCA	Target Class Accuracy
TCAV	Testing with Concept Activation Vectors
TI	Textual Inversion
VAE	Variational Autoencoder
VLM	Vision-Language Model
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

1.1 Context

Medical imaging modalities such as MRI, CT, and X-ray are central to modern diagnosis. Over the past decade, deep learning (DL) has been widely adopted in this field, producing predictive models that match or surpass human performance on several diagnostic benchmarks. Nevertheless, this performance comes at a cost: the inherent opacity of these “black box” models prevents the clinical validation necessary for real-world medical practice.

Traditional explainability methods, such as post-hoc saliency maps, have proven insufficient for high-stakes clinical decision-making. While these methods successfully identify where a model focuses its attention (anatomical locations), they fail to explain what semantic concepts motivated the decision. Furthermore, they cannot determine if a concept plays a causal role in the prediction or if the model relies on spurious statistical correlations present in the training data. This limitation makes it impossible to distinguish between genuine pathological features and confounding factors, undermining trust in automated diagnosis.

1.2 Motivations

To bridge the gap between model complexity and human understanding, **Concept-based Explainable AI (XAI)** has emerged as a promising solution. Approaches such as Concept Bottleneck Models (CBMs)[29] and Testing with Concept Activation Vectors (TCAV)[26] align machine decisions with human reasoning by decomposing predictions into understandable clinical concepts. This improves trust and, in the case of CBMs, lets a human intervene to correct the model’s reasoning.

Nevertheless, the practical implementation of these techniques faces a severe **Annotation Bottleneck**. Unlike general computer vision tasks, medical concepts require expert knowledge to

annotate, making the acquisition of dense, concept-level labels economically unviable and unscalable.

In addition, medical data suffers from the problem of *concept entanglement*. Pathological concepts are rarely statistically independent in observational data; the joint distribution of concepts often exhibits strong correlations. Discriminative models trained on such data tend to learn statistical shortcuts rather than disentangled causal representations. Consequently, there is a fundamental methodological gap: current techniques, ranging from simple data augmentation to earlier generative models like GANs and VAEs, fail to generate realistic images that manipulate specific clinical concepts while preserving the patient’s anatomical identity and biological coherence.

These two problems are distinct but compound each other: annotations alone do not remove concept entanglement, and removing entanglement does not make annotation cheaper.

1.3 Objectives

This thesis aims to overcome the scarcity of concept-annotated data by leveraging generative AI. The primary objective is to develop a pipeline for generating visual images where a specific target concept is inserted while the remaining anatomical context is preserved. These synthetic data are intended to support the training and evaluation of concept-based explainability models while substantially reducing the need for dense, concept-level manual annotation. The work is scoped to a binary, single-concept setting and to a frozen, publicly available generative backbone, so that the contribution of each design choice can be isolated and measured. As the experiments will show, the approach succeeds within a specific operating envelope rather than universally: a modality close to the backbone’s pre-training distribution, together with a salient target concept. The downstream-robustness axis is reported as an inconclusive, underpowered result.

These objectives can be framed as three research questions:

- RQ1.** Can a frozen, general-purpose text-to-image diffusion model insert an isolated pathological concept into a healthy medical image while preserving the surrounding anatomy?
- RQ2.** How should the concept token be learned? In particular, does restricting the Textual Inversion loss to the lesion region (M3) yield a more transferable, clinically aligned concept than training on cropped patches (M2) or relying on a hand-written prompt (M1)?
- RQ3.** Is the injected concept valid and useful for concept-based explainability: recognised as the target pathology, transferable to real data, and able to support a downstream model against spurious correlations?

Answering these questions yields both a concrete generation pipeline and a multi-axis picture of when its synthetic data can be trusted.

1.4 Contributions

The main contributions of this thesis are the following:

1. A mask-free concept-injection pipeline built on a frozen Stable Diffusion backbone, combining Textual Inversion and Null-Text Inversion, which inserts an isolated pathological concept into a healthy image without fine-tuning the backbone and without a manual mask at inference.
2. A controlled comparison of three concept-token training strategies under a shared pipeline: a hand-written prompt (M1), a token trained on cropped lesion patches (M2), and a token trained on whole images with the reconstruction loss restricted to the lesion mask (M3).
3. A multi-axis evaluation framework for the concept validity of synthetic medical images, combining perceptual fidelity (FID), semantic validity (Target Class Accuracy), concept learnability against a real-data upper bound, and a downstream-robustness probe based on concept-based regularisation (LCRReg), with explicit controls against data leakage.
4. An empirical study across three imaging modalities: retinal photography, dermoscopy, and breast ultrasound; selected to span a range of distances from the backbone’s pre-training distribution, together with copy-paste and diffusion-inpainting baselines as reference points.

1.5 Document Structure

The remainder of this document is organised as follows. Chapter 2 reviews the literature on concept-based explainability, generative models for medical imaging, and synthetic data for medical XAI, and identifies the methodological gap addressed by this work. Chapter 3 formalises the proposed method together with the two baselines. Chapter 4 describes the datasets, the evaluation pipeline, and the implementation details. Chapter 5 reports and discusses the experimental results. Chapter 6 concludes and outlines directions for future work.

Chapter 2

Literature Review

2.1 From Attribution Maps to Concept-based Reasoning

2.1.1 The Limitations of Saliency Methods

The deployment of AI in high-stakes clinical decision-making faces a critical barrier of trust due to the opacity of “black box” models. One common approach explains model behaviour with post-hoc visual explanations, specifically saliency or attribution maps. However, there are some studies showing that these pixel-level approaches can be insufficient and even misleading for clinical use.

A fundamental requirement for any reliable explanation method is robustness. Kindermans et al. [28] rigorously demonstrated that several widely used saliency methods fail basic reliability tests, specifically violating the axiom of input invariance. This axiom posits that if a transformation to the input creates a functionally equivalent representation that yields the exact same model prediction, the resulting explanation should remain unchanged. Formally, if a shift v does not affect the model’s prediction, then:

$$f(x) = f(x + v) \implies E(x) = E(x + v)$$

Kindermans et al. showed that methods such as Gradient $\times$ Input are highly sensitive to such shifts. By adding a constant vector to the input data, a change that, by construction, had zero impact on the model’s prediction, they produced drastically different saliency maps. The authors also demonstrated that these methods are susceptible to purposeful manipulation: they engineered an input shift containing a hidden pattern, an image of a cat, which, while ignored by the classifier, dominated the resulting saliency map. This shows that pixel-level attributions often reflect artefacts of the input data rather than the causal features driving the decision.

Adebayo et al. [1] identified a further problem: a lack of sensitivity to the model’s learned parameters. They introduced a series of “sanity checks”, most notably the model parameter randomisation test. The authors demonstrated that several popular attribution methods, such as Guided

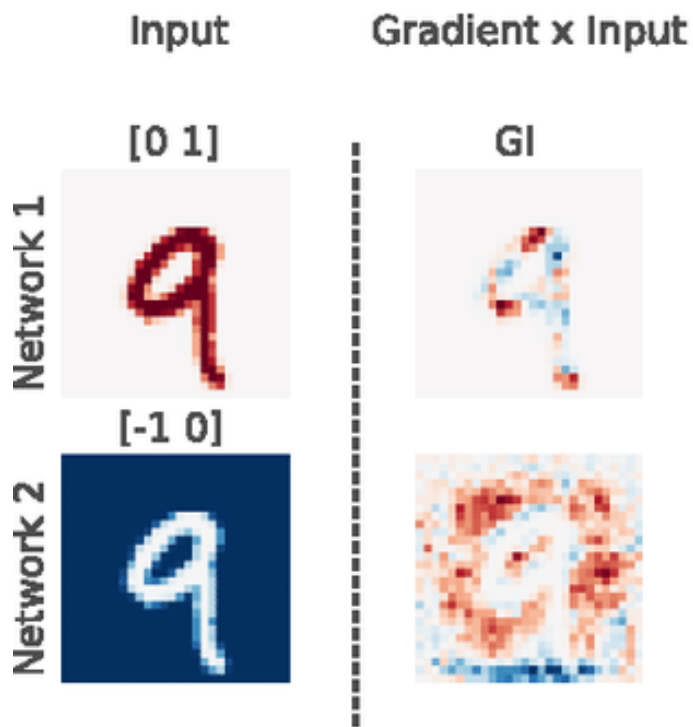


Figure 2.1: Violation of input invariance, after Kindermans et al. Two networks receive the same digit under a prediction-preserving input shift. The shift leaves the prediction unchanged, yet the $\text{Gradient} \times \text{Input}$ attribution maps differ markedly, reflecting an artefact of the input reference rather than the features driving the decision.

Back-propagation, produce highly similar and visually compelling explanations even when the weights of the neural network are progressively randomised. Instead of reflecting the model’s actual decision-making process, these methods often behave similarly to edge detectors, highlighting high-frequency structures in the input image regardless of the network’s trained state. In a clinical context, this implies that a physician could be presented with a highly detailed and convincing attribution map generated by a completely untrained random model.

Beyond technical fragility, there is a profound semantic limitation in using post-hoc explanations for critical tasks. Cynthia Rudin [45] argues that trying to explain black-box models in high-stakes scenarios is inherently problematic due to a lack of faithfulness. Saliency maps provide spatial attribution, where the model is “looking”, but fail to provide reasoning, i.e. how or why those features contribute to the decision.

This semantic gap creates dangerous ambiguity. Rudin illustrates this with an example in which a model could generate essentially identical saliency maps for two vastly different classes, as the visual edges activated are similar, yet the reasoning is distinct. In a medical context, this creates a false sense of security: a clinician might trust a diagnosis because the model highlights a correct anatomical region, unaware that the model might be relying on a confounding variable, such as

a specific X-ray tag or image artefact, rather than pathological tissue. As Rudin states, if an explanation were completely faithful to the complex black box, it would need to be as complex as the model itself.

Because pixel-level attribution cannot provide coherent “how and why” reasoning, the field has turned to methods that align model reasoning with human-understandable semantics, moving from pixels to concepts.

2.1.2 Concept-based Explainability

2.1.2.1 Post-hoc Concept-based Methods

Post-hoc concept-based methods were proposed to address the semantic limitations and lack of faithfulness inherent to pixel-based saliency methods. These approaches aim to translate the internal state of a deep neural network into human-understandable language, effectively mapping the model’s high-dimensional latent space (E_m) to a human concept space (E_h). Kim et al. [26] formalised this challenge with TCAV, a framework that moves beyond feature attribution to provide global explanations. TCAV grants the user the flexibility to define any concept of interest using a set of example images.

Mathematically, for a user-defined concept C , the method learns a linear boundary in the model’s activation space at layer l . The normal vector to this boundary is defined as the Concept Activation Vector (CAV), denoted as v_C^l . To quantify the importance of this concept for a specific prediction class k , TCAV first computes the conceptual sensitivity for a single input x . This is the directional derivative of the model’s logit $h_{l,k}$ (from layer l to class k) with respect to the activations $f_l(x)$ in the direction of the concept:

$$S_{C,k,l}(x) = \nabla h_{l,k}(f_l(x)) \cdot v_C^l$$

While this score measures the local sensitivity of a prediction to a concept, TCAV achieves global interpretability by aggregating these sensitivities across a dataset. The final TCAV score is defined as the fraction of inputs for class k where the conceptual sensitivity is positive. This metric allows clinicians to perform statistical hypothesis testing.

TCAV faces a significant limitation: it requires the user to know *a priori* which concepts to test. If a model relies on a feature unknown to the expert or a confounding variable not hypothesised by the user, TCAV will not detect it. To resolve this bottleneck, Ghorbani et al. [16] introduced Automated Concept-based Explanation (ACE). This method removes the need for human supervision in concept selection. Instead of manual definition, the algorithm automatically segments images into multi-resolution superpixels and clusters them based on their feature representations. These

clusters are treated as candidate concepts. ACE then applies the TCAV procedure to estimate the importance of each discovered concept, filtering out those that are not statistically significant.

While this automation solves the manual selection problem, it introduces a new dependency on the quality of segmentation and the training data: if the source dataset is noisy or the segmentation fails to isolate meaningful features, the resulting concepts may be ambiguous, redundant, or semantically incoherent, limiting their utility in clinical validation.

2.1.2.2 Ante-hoc Concept-based Models

In contrast to post-hoc methods, which attempt to explain a black-box model after training, ante-hoc approaches aim to structure the model architecture itself to ensure intrinsic transparency. The core principle of the Concept Bottleneck Model (CBM) [29] is to fundamentally alter the prediction flow, constraining the model to reason through a set of pre-defined human-understandable concepts. Formally, given an input x , the model is decomposed into two sequential sub-models: a concept predictor g mapping inputs to concepts ($c = g(x)$), and a label predictor h mapping concepts to the final target ($y = h(c)$).

This architecture also enables a form of human intervention: because the intermediate predictions correspond to interpretable concepts, a domain expert can inspect and potentially correct them during inference to observe how the final prediction changes. This “human-in-the-loop” capability fosters trust and allows for real-time error correction.

Nevertheless, training the concept predictor $g(x)$ requires explicit concept annotations, which are expensive and require expert labor, limiting the scalability of Concept Bottleneck Models (CBMs).

A second major limitation of standard CBMs is the accuracy–interpretability trade-off. Forcing all information to pass through a rigid, low-dimensional bottleneck can discard useful signals not captured by the pre-defined concepts, often degrading predictive performance compared to unconstrained black-box models. Addressing this, Corbetta et al. [11] introduced an in-hoc approach named LCRReg. Instead of treating concepts as a restrictive bottleneck ($x \rightarrow c \rightarrow y$), this method uses concept representations as a regularisation mechanism during training. The model retains a direct path from input to output features, preserving full predictive power, but is penalised via a concept-alignment loss if its internal representations do not align with concept activation vectors (CAVs). This encourages the model to rely on semantically meaningful features while mitigating spurious correlations, recovering concept-based interpretability without sacrificing predictive performance.

2.1.3 The Annotation and Entanglement Bottlenecks

While concept-based methods offer a theoretical path to interpretability, their deployment on real-world medical data is obstructed by two systemic phenomena: shortcut learning and the annotation bottleneck. Both are manifestations of the broader problem of concept entanglement introduced in Chapter 1: in observational medical data, clinical concepts are rarely independent of one another or of acquisition artefacts, so a model has no incentive to separate them. Here we focus on the concept–artefact form of this entanglement, in which a pathological feature is statistically tied to a non-pathological confounder.

Deep neural networks typically operate under the principle of least effort, optimising strictly for the training loss. Geirhos et al. [15] defined this behaviour as *shortcut learning*: the tendency of models to rely on superficial correlations that perform well on standard benchmarks but fail to generalise to out-of-distribution scenarios. In medical imaging, this issue is exacerbated because biological pathology is often inextricably entangled with other artefacts, creating spurious correlations that models exploit.

Badgeley et al. [2] reported a striking failure in hip-fracture detection: their deep learning models often acted as detectors of healthcare-process variables rather than of pathology. In their study, the model could predict the scanner model with near-perfect accuracy ($AUC = 1.00$) and whether an exam was marked as “priority” ($AUC = 0.79$) solely from the radiographs. While fracture itself was only moderately predictable from the image alone ($AUC = 0.78$), combining the image with patient and hospital-process metadata raised performance to $AUC = 0.91$. When the test set was re-sampled so that these confounders were matched across the fracture and non-fracture groups, the image-based detector performed no better than chance ($AUC = 0.52$, 95% CI 0.46–0.58). This demonstrates that the high apparent performance was largely driven by the correlation that sick patients are scanned on specific machines or marked as priority cases, rather than by the fracture itself.

Castro et al. [4] framed this as a failure of causal reasoning, characterising dataset shift and confounding through the lens of the causal relationships between image, acquisition process and label. A widely reported illustration of the phenomenon is the presence of visual confounders, such as skin markings, rulers, or treatment devices, that often appear alongside pathology: without causal intervention, models learn to associate these artefacts with the disease label. For instance, a ruler in a dermatoscopic image might become a stronger predictor of malignancy than the lesion’s texture, leading to catastrophic failures when the model encounters a ruler in a benign case. To mitigate this, Ross et al. [44] proposed “Right for the Right Reasons” (RRR), a regularisation technique designed to penalise the model for relying on irrelevant features. Mathematically, the RRR objective adds a penalty term to the standard loss function:

$$\mathcal{L}_{RRR} = \mathcal{L}_{task}(\hat{y}, y) + \lambda \sum_{k \in A} \left(\frac{\partial \log(p(y|x))}{\partial x_k} \right)^2$$

where A represents the binary mask of input features known to be irrelevant and the term inside the summation measures the sensitivity of the model’s (summed-class) log-likelihood to those features. By minimising this gradient, the method forces the model to be locally invariant to artefacts.

This solution introduces a severe annotation bottleneck, as it requires explicit annotations identifying the regions that should not influence the prediction. In the medical domain, where obtaining even basic diagnostic labels is expensive and time-consuming, requiring pixel-level masks for every “ruler”, “wire”, or “scanner tag” is practically infeasible.

These two facts reinforce each other: real-world medical data contains spurious correlations that are hard to disentangle, and the regularisation techniques that would correct them require annotations that cannot be obtained at scale. Synthetic data generation offers a way out: by controlling the generation process in a way that breaks the observational correlation that drives shortcut learning.

2.2 Generative Models for Medical Imaging

2.2.1 Representation Learning and Disentanglement (VAEs)

Unlike standard autoencoders that simply compress data, VAEs aim to learn a probabilistic latent space. A central objective in this domain is disentanglement: learning a representation where each dimension of the latent vector z controls a single, independent semantic factor of variation.

Higgins et al. [21] formalised this objective with the β -VAE. They proposed that by increasing the weight of the Kullback–Leibler (KL) divergence term in the evidence lower bound (ELBO), the model is forced to learn a more efficient and factorised latent distribution. The maximisation objective is defined as:

$$\mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \beta D_{KL}(q_{\phi}(z|x) || p(z))$$

where $\beta > 1$ is a hyperparameter. A higher β imposes a stronger constraint on the latent bottleneck, encouraging a more factorised and compressed latent representation. This comes at a cost: high β values often compromise reconstruction accuracy. Kim et al. [27] addressed this trade-off with FactorVAE. Instead of uniformly penalising the KL divergence, they explicitly penalise the total correlation (TC) of the latent distribution. The FactorVAE objective minimises a term based on $KL(q(z) || \prod_j q(z_j))$, effectively forcing the joint distribution of latent codes to be the product

of its marginals. This approach achieves better disentanglement without penalising the mutual information between data and latent codes as aggressively as β -VAE.

In the medical domain, Chatsias et al. [7] demonstrated the utility of this paradigm with the Spatial Decomposition Network (SDNet). Rather than relying on a single unstructured latent vector, SDNet explicitly factorises the image into two distinct representations: a spatial anatomical factor (s), which captures the structural segmentation of the organ, and a non-spatial modality factor (z), which encodes intensity characteristics. By recombining these factors ($x' = f(s, z)$), the model enables cross-modality synthesis: one can extract the anatomy (s) from an MRI and combine it with the modality code (z) of a CT scan, synthesising the patient's anatomy in a different imaging protocol while encouraging structural consistency.

Even with these advances, VAE-based disentanglement faces two critical challenges. First, a theoretical impossibility: Locatello et al. [33] proved that unsupervised disentanglement is fundamentally impossible without inductive biases. They demonstrated that for any dataset, there exist infinite generative models that fit the data distribution perfectly but possess entangled latent spaces. Without explicit supervision (labels) or architectural constraints (like SDNet's spatial bottleneck), there is no guarantee that a VAE will learn human-meaningful concepts rather than arbitrary basis functions. Second, limited image quality: due to the nature of the reconstruction loss and Gaussian assumptions, VAEs tend to produce blurry images with lost high-frequency details. In medical diagnostics, where texture micro-structures are vital, this lack of sharpness limits their clinical utility, paving the way for adversarial and diffusion models.

2.2.2 Adversarial Synthesis (GANs)

By leveraging an adversarial training scheme, a min-max game in which a generator attempts to fool a discriminator, GANs succeeded in synthesising images with photorealistic textures.

A major breakthrough in clinical applicability was the ability to perform synthesis without paired data (e.g., obtaining a CT and MRI of the same patient in the exact same pose is rare). Chatsias et al. [6] demonstrated the efficacy of this approach for unpaired multi-modal synthesis. Using a cycle-consistent adversarial architecture inspired by CycleGAN, they successfully translated cardiac CTs into MRIs without direct patient correspondence. The methodology leverages the cycle-consistency loss, which ensures that an image translated to the target domain and back returns to its original state. The full objective enforces this constraint in both directions:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)} [\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{data}(y)} [\|G(F(y)) - y\|_1]$$

where G translates source $\rightarrow$ target and F translates target $\rightarrow$ source. Chatsias et al. showed that using these synthetic MRIs as a form of sophisticated data augmentation significantly improved the performance of segmentation models, in line with the broader review by Yi et al. [51].

This focus on visual realism comes with a safety problem: a lack of semantic integrity. Cohen et al. [9] showed a failure mode in the distribution-matching objectives used by GANs. They demonstrated that these models have a tendency to “hallucinate” or actively remove clinical features to satisfy the statistical properties of the target dataset. Specifically, if the target domain is imbalanced, the GAN learns that “absence of tumours” is a defining characteristic of the target distribution $P_{target}(x)$. Consequently, when translating a diseased image, the discriminator penalises the presence of pathology as unrealistic, forcing the generator to erase a tumour or hallucinate a lesion solely to align the output with the target marginal distribution.

This phenomenon makes standard adversarial synthesis risky for concept-based XAI, particularly when the target distribution is imbalanced. In a framework where the goal is to generate a specific concept, a model that prioritises fooling a discriminator over preserving semantic constraints may inadvertently alter the diagnosis, severing the link between the explanation and the model’s true reasoning.

2.2.3 Diffusion Models: from DDPMs to Latent and Conditional Diffusion

Unlike GANs, which attempt to synthesise an image in a single, often unstable adversarial step, Denoising Diffusion Probabilistic Models (DDPMs) formulate generation as an iterative denoising process. Ho et al. [22] defined this as a parameterised Markov chain: a forward process progressively destroys the structure of an image x_0 by adding Gaussian noise until it becomes pure noise x_T ; a reverse process trains a neural network ϵ_θ to iteratively denoise the signal, recovering the original structure step-by-step. The training objective is stable in practice, simplifying to a reweighted mean-squared error between the added noise ϵ and the model’s prediction of that noise:

$$\mathcal{L}_{simple} = \mathbb{E}_{t, x_0, \epsilon} \left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\|^2 \right]$$

where t is the timestep and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ is the cumulative product of the per-step noise-schedule coefficients. While this iterative nature eliminates the mode collapse common in GANs, it initially imposed a prohibitive computational cost, requiring thousands of passes through the network to generate a single image in pixel space.

To address this limitation, Latent Diffusion Models (LDMs) were introduced, which move the diffusion process from pixel space to a compressed latent space. Building on this paradigm, Pinaya et al. [41] successfully applied latent diffusion to high-dimensional brain MRI synthesis. Instead of operating in the computationally expensive pixel space, they employed an autoencoder with a KL-regularised latent space, trained with a combination of L_1 , perceptual and patch-based adversarial losses, to compress images into a lower-dimensional, perceptually equivalent representation.

The diffusion process is then trained on these compact latent representations. This approach allowed for the synthesis of high-resolution, anatomically coherent brain scans with a fraction of the computational resources, effectively decoupling the training cost from the image resolution.

Two design choices in modern diffusion models are central to the method developed in this thesis. The first is deterministic, approximately invertible sampling. Song [49] introduced Denoising Diffusion Implicit Models (DDIM), a non-Markovian sampler that removes the stochastic term of the DDPM reverse step. Because the resulting mapping from noise to image is deterministic, the trajectory can in principle be run backwards to recover the noise code of a *real* image, the property that later enables real-image editing. The second is conditioning strength. Classifier-Free Guidance (CFG) [23] trains a single network in both a conditional and an unconditional mode and, at sampling time, extrapolates between the two predictions with a guidance scale that amplifies the influence of the conditioning signal. CFG is what allows a text prompt to steer generation as formalised in Chapter 3.

Beyond efficiency and control, diffusion models have shown strong semantic conditioning in the clinical domain. Chambon et al.[5] demonstrated this with RoentGen, a latent diffusion model fine-tuned on the MIMIC-CXR dataset. By adapting a domain-specific text encoder, RoentGen aligns medical language with visual features through cross-attention layers. The authors showed that the model could be prompted with complex clinical queries, generating synthetic chest X-rays that radiologists judged to reflect the requested pathology. This is encouraging for generative XAI, suggesting that, in at least some modalities, diffusion models can act as “rendering engines” for medical concepts.

2.2.4 Vision Language Models

Diffusion models can synthesise images of high quality, but their ability to generate a specific medical concept depends on how well they interpret the conditioning signal. Standard models rely on text encoders to translate prompts into vector representations that guide the denoising process, but encoders trained on general-domain data often fail to capture the nuances of medical terminology, creating a semantic gap between the clinician’s intent and the generated output.

To bridge this gap, the field has adopted Vision-Language Models (VLMs), pioneered by CLIP [43]. Unlike supervised classifiers trained on fixed labels, CLIP learns a joint embedding space where semantically related images and text descriptions are mapped close together. This allows the model to align visual features with their natural language descriptions. In the medical domain, Zhang et al. [52] scaled this approach with BiomedCLIP, pre-training on 15 million biomedical image-text pairs. They demonstrated that domain-specific alignment is crucial: a general VLM might associate “opacity” with transparency in design, whereas BiomedCLIP correctly maps it to pathological tissue density.

This vision-language alignment offers a partial solution to the annotation bottleneck discussed in Section 2.1.3. Patrício et al. [38] showed that, instead of relying on expensive pixel-wise segmentation masks to define a concept, concepts can be defined through textual prompts and retrieved using similarity in the VLM embedding space. By using textual descriptions of dermoscopic features as anchors, they demonstrated that the model could identify and retrieve concepts in a zero-shot or few-shot manner with performance comparable to dedicated concept-generation methods.

The text encoder of a latent diffusion model is itself such a VLM (CLIP, in the case of Stable Diffusion). This is what makes the conditioning embedding space the natural place to define a clinical concept. However, as the next subsection discusses, defining a fine-grained medical concept through a hand-written natural-language prompt is unreliable; this motivates learning the concept directly from images as a dedicated token within that same embedding space.

2.2.5 Model Personalisation and Concept Learning

The semantic gap above means that a general-domain text encoder rarely has a word for a specific clinical concept, and hand-written prompts (e.g. “a retina with yellow exudates”) produce, at best, a loose approximation of the intended pathology. *Personalisation* methods address exactly this problem: given a few example images of a concept, they teach a frozen text-to-image model a new way to refer to it. Two families exist. Weight-based personalisation, such as Dream-Booth [46], fine-tunes (part of) the diffusion backbone to bind the concept to a rare token, while low-rank adaptation (LoRA) [24] injects small trainable matrices into the attention layers. These are expressive but modify the model and risk overfitting or catastrophic forgetting in a few-shot regime.

Textual Inversion [14] takes the opposite, minimally invasive route. It freezes the entire model and learns only the embedding of a single new pseudo-word v^* in the text-encoder space, optimised with the standard diffusion reconstruction loss so that prompts containing v^* reproduce the concept. Because nothing but one embedding vector is trained, the method is extremely parameter-efficient, preserves the model’s prior, and is well suited to the small, expert-annotated datasets typical of medical imaging. This makes a learned concept token an attractive substitute for an unreliable natural-language prompt, and it is the personalisation strategy adopted by the method tested in this thesis (Chapter 3).

Finally, realistic synthesis raises the question of how synthetic images should be evaluated. The standard measure of generative realism is the Fréchet Inception Distance (FID) [20], which compares the feature statistics of real and synthetic image sets. However, distributional realism is not enough on its own here. An image can match the real diseased distribution closely yet still fail to render a valid clinical concept, and an image that looks imperfect can still be useful as concept

supervision. This gap between fidelity and validity is why this work adopts a multi-axis evaluation framework (Chapter 4).

2.3 Synthetic Images for Medical XAI

The previous section reviewed how images are generated; this section reviews how synthetic and generative manipulations have been used to *explain* medical models. The methods fall into three groups: counterfactual explanations, prototype-based reasoning, and concept-conditional editing.

2.3.1 Visual Counterfactual Explanations

One of the ways to explain a model’s decision is through Visual Counterfactual Explanations (VCEs). A VCE is a synthesised version of the patient’s image (x') that is minimally different from the original image (x) but sufficient to flip the model’s prediction to a target class y' . Formally, finding a counterfactual can be framed as an optimisation problem:

$$x' = \arg \min_{\hat{x}} (\mathcal{L}_{pred}(f(\hat{x}), y') + \lambda \|\hat{x} - x\|_p)$$

where the first term ensures the class change and the second term minimises the visual distance (usually L_1 for sparsity or a perceptual loss) to preserve the patient’s identity. The difference map $|x - x'|$ thus reveals the exact features the model utilised for its diagnosis.

Cohen et al. [10] introduced “Gifsplanation”. Using a method called Latent Shift, they iteratively perturbed the latent representation of an image within an autoencoder to maximise the probability of a target pathology. This allowed them to generate continuous animations (GIFs) showing the progressive onset of a disease. While clinically intuitive, the reliance on autoencoder-based latent representations often resulted in limited spatial fidelity and a loss of high-frequency structure.

Diffusion models alleviated this quality bottleneck. Fang et al. [13] proposed DiffExplainer, which trains a diffusion autoencoder with a denoising reconstruction objective and then distils the frozen black-box classifier into a lightweight surrogate “agent” model. Counterfactuals are generated by moving the encoded latent along the gradient of this surrogate and decoding the result, producing sharp, photorealistic images that flip the prediction label while focusing the intervention on the discriminative regions, avoiding the blurriness of VAE-based approaches.

Unconstrained modification of pixels can lead to biologically impossible images, but Sanchez et al. [47] addressed this with Diffusion Causal Models (Diff-SCM). They integrated structural causal models into the generation process, explicitly modelling the causal relationships between attributes. Their method performs counterfactual estimation via an abduction–action–prediction step: it first infers the exogenous noise from the original image (abduction) and then intervenes

on a causal parent node (e.g., $do(\text{class} = \text{True})$) during the reverse diffusion process. Although Diff-SCM was demonstrated on natural-image benchmarks (MNIST and ImageNet) rather than on medical data, the principle is directly relevant: conditioning generation on an explicit causal structure encourages counterfactuals that respect the dependencies between attributes, avoiding spurious or physically inconsistent modifications that might otherwise satisfy the classifier.

These approaches share a fundamental limitation: they are typically binary and holistic. The transition modifies the image from “class A” to “class B” as a whole, often altering multiple correlated attributes simultaneously. For example, transforming a “healthy” chest X-ray into “pneumonia” might simultaneously change the lung opacity and the heart border definition. This prevents the user from isolating specific features: observing a pathology appear is useful, but it does not clarify which specific characteristic drove the prediction.

2.3.2 Case-based and Prototype-based Reasoning

While counterfactual explanations modify the input image to reveal decision boundaries, prototype-based methods explain predictions by comparing a case with representative examples learned by the model.

Case-based reasoning (CBR) aligns closely with the cognitive process of medical diagnosis, where clinicians validate a decision by referencing similar past cases. To integrate this logic into deep learning, Li et al. [32] introduced ProtoPNet, a prototype-based neural network inspired by case-based reasoning. In this framework, the network learns a set of prototype vectors p_j in the latent space. The classification of a new image x is determined not by a black-box function, but by its distance to these prototypes.

To make these prototypes interpretable, the authors project the learned latent vectors p_j onto the nearest training patch, visualising them as actual image snippets.

Barnett et al. [3] demonstrated the clinical utility of this approach with IAIA-BL for mammography. Their model justifies a malignancy prediction by explicitly highlighting that the patient’s lesion is morphologically similar to a specific “prototype” extracted from the training dataset. However, relying on prototypes retrieved directly from real patient data introduces two critical limitations:

- **Privacy:** storing and displaying patches of real patient exams as explanations raises significant ethical and data-protection concerns, potentially revealing sensitive information.
- **Rigidity:** anatomical variability means a fixed prototype (e.g., a “standard” normal lung from patient A) rarely aligns perfectly with a new patient B’s unique anatomy, making pixel-wise comparisons visually incongruent.

More recent work makes prototype-based explanation more flexible and transferable. Keswani et al. [25] proposed Proto2Proto, a knowledge-distillation framework in which a compact “student” network is taught to reproduce the interpretable prototypes of a “teacher” through global (prototype-matching) and local (patch-prototype correspondence) losses, transferring case-based interpretability to lighter models. Shen et al. [48] extended prototype learning to segmentation with ProtoSAM-2D, which reformulates semantic segmentation as a post-hoc classification task: it fuses the encoder and decoder features of the Segment Anything Model (SAM) into mask-level embeddings that are classified by similarity to learned prototypes, so a new or few-shot class can be added by learning an additional prototype rather than retraining. Despite these advances, prototype methods still explain a decision by retrieval and comparison: they justify it by analogy to learned exemplars rather than by manipulating an isolated concept. Concept editing, examined next, instead explains by intervention.

2.3.3 Concept-Conditional Generation and Editing

While the counterfactual and prototype approaches above expose what a model has learned, concept-conditional editing refers to the capability to manipulate an isolated semantic concept within an image and analyse how that controlled intervention influences the model’s prediction. The relevant contrast is not visualisation versus manipulation, since counterfactuals already manipulate the image. It is between holistic, label-flipping editing (Section 2.3.1) and isolated, single-concept editing.

This capability aligns concept editing with the broader VCE framework: rather than flipping a global label, concept editing performs a structural causal intervention on a single semantic variable. Goyal et al. [17] and Pawlowski et al. [40] argue that true interpretability requires answering “what if?” questions about specific semantic variables. This transforms the generative model from a mere visualisation tool into an instrument for causal sensitivity analysis.

Lang et al. [31] presented StyleEx. Working with Generative Adversarial Networks (GANs), they demonstrated that standard latent spaces (W or $W+$) often contain entangled information. To resolve this, they exploited StyleSpace (S), a latent representation that exposes channel-wise style parameters and tends to be more semantically disentangled than the standard latent spaces. Formally, editing an image corresponds to identifying a specific channel index c in StyleSpace and manipulating its activation:

$$s_{new} = s + \alpha \cdot u_c$$

where u_c is the unit vector for the semantic concept and α is a scalar controlling the magnitude of the attribute. Lang et al. showed that manipulating these specific coordinates allows for the modification of fine-grained attributes without destroying the subject’s identity.

Following the paradigm shift to diffusion models, the focus moved toward integrating concept representations directly into the iterative denoising process. Prompt-to-Prompt [19] performs text-driven editing through *cross-attention control*, manipulating the attention maps that link textual tokens to image regions during denoising. RePaint [34], by contrast, is a mask-based inpainting method that conditions an *unconditional* diffusion model purely by resampling the known (unmasked) region at each reverse step, without any text guidance or cross-attention over textual tokens. Both illustrate that, in the natural-image domain, fine-grained editing typically relies either on text–region attention or on an explicit spatial mask a dependency that, as discussed below, is problematic in the medical setting. Where text guidance is used, the concept can be defined either through a natural-language prompt or, following Section 2.2.5, through a learned concept token, an idea related to concept-token architectures such as ViConEx-Med [39].

Translating text-guided and cross-attention editing to medical imaging presents unique challenges. Unlike natural images, medical scans lack distinct object boundaries, and clinically relevant features are often diffuse, subtle, or inextricably entangled with the surrounding anatomy. Consequently, many existing clinical approaches fall back on semantic inpainting, where a user manually masks a pathological region and instructs the model to regenerate it. While effective for localised interventions, this process is labour-intensive, prone to annotation bias, and completely destroys the original structural information within the masked region.

To reduce the dependency on pixel-level masks, Özbey et al. [37] proposed SynDiff, an adversarial diffusion framework for unpaired image-to-image translation in medical imaging. By combining the iterative refinement of diffusion with adversarial projections and cycle-consistency losses, the model learns to transform the global appearance of an image while preserving the underlying anatomical structure, without requiring paired data or manual masks. The edit it performs, however, is a global style transformation rather than the insertion of a single localised concept.

A complementary line of work edits *real* images by first inverting them into the diffusion model’s noise space. Mokady et al. [36] introduced Null-Text Inversion (NTI), which optimises the unconditional (null-text) embedding so that a real image can be reconstructed under Classifier-Free Guidance and subsequently edited by changing the prompt, leaving the model weights untouched. A related and computationally lighter strategy, SDEdit [35], adds a controlled amount of noise to the image and denoises it under the new condition, trading a small loss of fidelity for a large reduction in cost. These inversion-based editors are the technical basis for mask-free editing of real patient images, and they build directly on the DDIM inversion and CFG mechanisms introduced in Section 2.2.3.

Another recent approach is PRISM [30], which uses a language-guided Stable Diffusion model to synthesise high-resolution medical counterfactuals. The method fine-tunes the diffusion model to adapt the generative process to the target medical domain. Counterfactual attributes are specified

through hand-written textual prompts that guide the model towards clinically meaningful semantic modifications. By conditioning generation on these prompts, PRISM is focused on removing or attenuating features associated with specific diseases, biomarkers, or spurious medical artefacts without requiring explicit segmentation masks.

These advances still leave a gap in medical concept editing. Existing techniques fall into three groups, none of which supports mask-free, disentangled, localised concept *insertion*:

- **Inpainting:** requires a manual mask and overwrites the original pixel-level information within the masked region, which may introduce structural inconsistencies or discard clinically relevant detail.
- **Global translation:** alters the entire style or contrast of the image, which is overly aggressive when the goal is to adjust a single, specific pathological feature while keeping the rest of the clinical presentation fixed.
- **Backbone fine-tuning under data scarcity:** PRISM removes the manual-mask and global-translation limitations above and can target individual attributes, but it does so by fine-tuning the U-Net on a large labelled corpus (CheXpert, >200k images), and its disease edits are removal-oriented rather than insertions. A frozen backbone that learns only a concept embedding and inserts the concept into a healthy image remains unexplored.

Together these limitations point to one missing capability in concept-based medical XAI: mask-free, disentangled concept insertion on a frozen backbone, without fine-tuning the generator. Because such a method learns the concept from only a handful of annotated images, it could make concept-controlled data augmentation feasible even for the small datasets typical of this domain precisely the regime in which fine-tuning a full generative model is not an option.

Chapter 3

Methodology

3.1 Baseline Methods

To contextualise the contributions of the proposed approach, we define two baseline methods that serve as reference points for concept synthesis quality. The first is a deterministic copy-paste pipeline that transplants real lesion pixels directly onto healthy images, providing an explicit upper bound on pixel-level concept fidelity. The second replaces the copied patch with content generated by a pre-trained diffusion inpainting model, testing whether a general-purpose generative model guided solely by a spatial mask and a text prompt can produce a semantically valid concept representation without any form of concept-specific learning.

3.1.1 Baseline 1: Lesion Copy-Paste

3.1.1.1 Overview

The copy-paste baseline constructs synthetic concept-positive images by extracting lesion patches from annotated source images and pasting them onto healthy background images at the locations defined by the corresponding segmentation masks. No generative model is involved: the concept region of every synthetic image contains unmodified real tissue, making this method maximally faithful to the original pathology texture. As such, it constitutes a *pixel-level ceiling* on concept fidelity: any approach that operates in latent space and claims to capture the same concept must be measured against it.

3.1.1.2 Formalisation

Let $\mathcal{H}$ denote a set of healthy background images free of the target concept, and let $\mathcal{S}$ be a set of source images in which the pathology is present. Each source image $p \in \mathcal{S}$ is associated with a binary segmentation mask $m \in \{0, 1\}^{H \times W}$ that delineates the spatial extent of the lesion, where H and W are the image height and width, respectively.

Given a healthy background image $h_b \in \mathcal{H}$, a concept-positive synthetic sample x^+ is produced by extracting the lesion patch from a source image p at the region defined by m , and compositing it onto h_b at the same spatial location:

$$x^+ = \text{Paste}(h_b, \text{Patch}(p, m)) \quad (3.1)$$

where $\text{Patch}(p, m)$ denotes the sub-image of p within the mask region, and $\text{Paste}(\cdot)$ replaces the corresponding pixels of the background with the extracted patch. Outside the mask, the healthy background h_b is preserved exactly.

3.1.2 Baseline 2: Diffusion-Based Inpainting

3.1.2.1 Overview

The second baseline replaces the copied lesion patch with content synthesised by a pre-trained diffusion inpainting model conditioned on the mask region. This approach retains the spatial structure of the copy-paste pipeline, the mask defines where the concept should appear and the surrounding healthy tissue is preserved, but delegates the generation of the concept’s visual content to a generative model, removing the dependency on a fixed pool of source patches.

The motivation for this baseline is to test the hypothesis that a generative model, guided solely by a textual description of the pathology and a spatial mask, can produce a concept representation that is both visually realistic and semantically valid.

3.1.2.2 Formalisation

Let $\mathcal{H}$, $\mathcal{S}$, and m be defined as in Section 3.1.1. Given a healthy background image $h_b \in \mathcal{H}$, a binary mask m specifying the target region, and a pre-trained diffusion inpainting model $\mathcal{D}_\theta$ with parameters θ , the synthetic concept-positive image is produced as:

$$x^+ = \mathcal{D}_\theta(h_b, m, \tau) \quad (3.2)$$

where τ is a text prompt describing the target pathology. The model receives the unmasked region of h_b as conditioning context and generates plausible content for the masked region through iterative denoising, starting from Gaussian noise $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$ within the mask. Pixels outside m are held fixed and equal to h_b throughout the diffusion process, ensuring that only the lesion region is modified.

3.2 Tested Method: Null-Text Inversion with Concept Injection

3.2.1 Overview

The analysis in Section 2.3.3 revealed an underexplored research direction in concept-level image editing for medical XAI. Existing approaches are polarised between two extremes: *inpaint-*

ing methods require a manually drawn mask and overwrite all pixel-level information within the masked region, while *global translation* techniques alter the entire image appearance, making it impossible to attribute any change to a single clinical concept. Neither paradigm supports the kind of intervention required by concept-based regularisation frameworks: the controlled insertion of an isolated pathological concept into a healthy image while preserving the patient’s anatomical identity.

To address this gap, we explore a two-stage pipeline that combines Textual Inversion [14] with Null-Text Inversion (NTI) [36]. The pipeline, illustrated in Figure 3.1, operates as follows:

Stage 1 - Concept Token Training. Given a set of real images containing the target pathology together with their binary segmentation masks, we learn a dedicated concept token v^* using Textual Inversion. Unlike the standard formulation, in which the reconstruction loss is computed over the entire image, we weight the loss by the segmentation mask so that the gradient signal is concentrated on the lesion region. This forces the token to encode the visual signature of the concept rather than the surrounding healthy tissue. The mask is used *only during training* as a loss-weighting mechanism; it is not required at inference time.

Stage 2 - Inversion-Based Generation. To synthesise a concept-positive image for a new, healthy patient, we first carry the patient’s encoded healthy image to an intermediate noise level through a deterministic forward (inversion) trajectory, obtaining a pivot that encodes the patient’s anatomy. We then optimise a null-text embedding so that the Classifier-Free-Guided reverse trajectory reconstructs the original image with high fidelity. Finally, we replace the null prompt with the learned concept token v^* during the reverse diffusion process. Because the optimised null-text embedding anchors the trajectory to the patient’s anatomy, the principal degree of freedom left to the model is the concept injected through the prompt; this is intended to confine the edit to the concept-relevant region while preserving the anatomical background.

Method variants. To isolate the contribution of each component, we evaluate three concept-injection strategies that share the same evaluation protocol (Section 4.3):

- **M1 - NTI + Prompt:** the NTI generation pipeline is used with a hand-written text prompt describing the pathology; no concept token is trained.
- **M2 - NTI + Token (patches):** the concept token is trained with standard Textual Inversion on cropped lesion patches, following [14].
- **M3 - NTI + Token (masked loss):** the concept token is trained on whole images with the loss weighted by the segmentation mask, as described above.

The following subsections formalise each component of the pipeline: DDIM Inversion (Section 3.2.2), Null-Text Optimisation (Section 3.2.3), Concept Token Training (Section 3.2.4), and the Reverse Diffusion procedure with concept injection (Section 3.2.5).

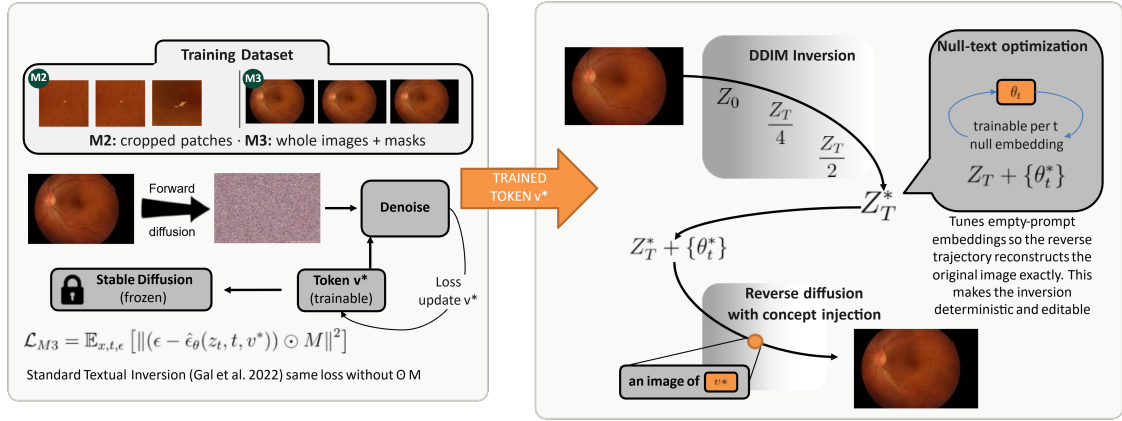


Figure 3.1: Overview of the proposed pipeline. **Stage 1** (left): a concept token v^* is trained via Textual Inversion with masked-loss weighting, using real images that contain the target pathology. **Stage 2** (right): a healthy patient image is inverted through DDIM Inversion and Null-Text Optimisation; the learned token is then injected into the reverse diffusion process to synthesise a concept-positive version of the same anatomy.

3.2.2 DDIM Inversion

The first step of the generation pipeline requires mapping a real patient image into the noise space of the diffusion model, obtaining a latent trajectory that can later be manipulated. This section formalises the inversion procedure based on Denoising Diffusion Implicit Models (DDIM) [49] and identifies the reconstruction error that motivates the Null-Text Optimisation described in Section 3.2.3.

3.2.2.1 Deterministic Sampling with DDIM

As discussed in Section 2.2.3, Denoising Diffusion Probabilistic Models learn to reverse a fixed forward process that progressively corrupts an image z_0 into Gaussian noise z_T . While the original DDPM formulation [22] defines a stochastic reverse process, Song et al. [49] showed that the same trained noise estimator ϵ_θ can be used in a deterministic sampling scheme by removing the stochastic component. The DDIM update rule from timestep t to $t-1$ is given by:

$$z_{t-1} = \underbrace{\sqrt{\bar{\alpha}_{t-1}} \frac{z_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(z_t, t)}{\sqrt{\bar{\alpha}_t}}}_{\text{predicted } z_0} + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta(z_t, t) \quad (3.3)$$

where $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ is the cumulative noise schedule. Because no random noise is injected, the mapping from z_T to z_0 is fully deterministic for a given initial noise vector: the same z_T always produces the same z_0 . This determinism is what makes the process invertible in principle.

3.2.2.2 Inverting the Sampling Trajectory

The deterministic nature of Equation (3.3) implies that, given a clean latent z_0 encoded from a real image, one can *reverse* the sampling direction to recover the corresponding noise map z_T . Rearranging Equation (3.3) to express z_t as a function of z_{t-1} yields the DDIM inversion step:

$$z_t = \sqrt{\bar{\alpha}_t} \frac{z_{t-1} - \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta(z_{t-1}, t)}{\sqrt{\bar{\alpha}_{t-1}}} + \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(z_{t-1}, t) \quad (3.4)$$

Iterating this expression from $t = 1$ to $t = T$ produces a sequence of increasingly noisy latents $\{z_0, z_1, \dots, z_T\}$, referred to as the *inversion trajectory*. The terminal latent z_T^* serves as the *noise pivot*: the starting point from which the reverse diffusion process will reconstruct the image. In practice, however, the inversion process is often truncated at an intermediate timestep $t_{\text{start}} < T$ rather than proceeding to pure noise. Following the intuition of SDEdit [35], this choice preserves a substantial portion of the encoded image structure in the pivot latent. The implications of this design choice are discussed in Section 3.2.3.

Equation (3.4) involves an approximation: it evaluates the noise estimator at z_{t-1} rather than at the true z_t , which is unknown at the time of computation. Under the unconditional model (no text guidance), this local linearisation error is small and accumulates only mildly over the trajectory, yielding reconstructions that are close to the original.

3.2.2.3 The Classifier-Free Guidance Problem

Modern text-conditioned diffusion models employ Classifier-Free Guidance (CFG) [23] to strengthen the influence of the conditioning signal. Under CFG, the noise estimate used during sampling is a linear extrapolation between the unconditional and conditional predictions:

$$\tilde{\epsilon}_\theta(z_t, t, \mathcal{C}) = \epsilon_\theta(z_t, t, \varnothing) + w [\epsilon_\theta(z_t, t, \mathcal{C}) - \epsilon_\theta(z_t, t, \varnothing)] \quad (3.5)$$

where $\mathcal{C}$ denotes the text conditioning, $\varnothing$ is the *null-text* embedding (the unconditional prompt), and w is the guidance scale. The term $\epsilon_\theta(z_t, t, \varnothing)$ acts as a baseline: it controls the direction from which the conditional signal $\mathcal{C}$ steers the generation.

When CFG is applied during inversion, the extrapolation amplifies the local approximation error introduced by Equation (3.4) at every timestep, since the guided noise estimate $\tilde{\epsilon}_\theta$ deviates more strongly from the true noise than the unconditional estimate alone. The result is a significant drift: applying the forward DDIM inversion followed by a guided reverse pass no longer reconstructs the original image. This drift is not merely a minor quality degradation; it can alter anatomical structures and introduce artefacts that invalidate any downstream concept-editing operation.

This reconstruction gap constitutes the core technical problem that Null-Text Optimisation (Section 3.2.3) is designed to solve.

3.2.3 Null-Text Optimisation

Rather than altering the inversion trajectory or abandoning guidance Mokady et al. [36] optimise the null-text embedding $\emptyset$ that appears in the unconditional branch of the CFG equation (Equation 3.5) on a *per-timestep* basis, so that the guided reverse process is forced to follow the precomputed inversion trajectory exactly. Figure 3.2

3.2.3.1 Intuition

Recall from Equation (3.5) that the guided noise estimate is an interpolation anchored by the unconditional prediction $\epsilon_\theta(z_t, t, \emptyset)$. In a standard pipeline, $\emptyset$ is the fixed embedding of an empty string and remains constant across all timesteps. The point to note is that $\emptyset$ need not be held fixed; it is simply the input to the text-conditioning branch of the U-Net. By allowing it to vary per timestep, one obtains a set of free parameters $\{\theta_t\}_{t=1}^T$ that can absorb the reconstruction error without modifying either the model weights or the inversion trajectory.

3.2.3.2 Optimisation Procedure

Let $\{z_0^*, z_1^*, \dots, z_T^*\}$ denote the inversion trajectory obtained by applying DDIM inversion (Equation 3.4) to the encoded real image. This trajectory is computed once and kept fixed throughout the optimisation.

For each timestep t from T down to 1, we seek a null-text embedding θ_t such that a single guided DDIM sampling step starting from the current latent $\bar{z}_t$ lands on the corresponding inversion target z_{t-1}^* . Concretely, define the guided DDIM step as:

$$\hat{z}_{t-1}(\theta_t) = \text{DDIM}(\bar{z}_t, t, \tilde{\epsilon}_\theta(\bar{z}_t, t, \mathcal{C}, \theta_t)) \quad (3.6)$$

where $\tilde{\epsilon}_\theta$ is the CFG noise estimate from Equation (3.5) with $\emptyset$ replaced by the learnable θ_t , and $\mathcal{C}$ is the original caption (or an empty conditioning signal, as the image has not yet been edited at this stage). The optimisation objective at timestep t is:

$$\theta_t^* = \arg \min_{\theta_t} \|\hat{z}_{t-1} - \hat{z}_{t-1}(\theta_t)\|_2^2 \quad (3.7)$$

This is a lightweight optimisation problem: the only trainable parameters are the embedding vectors θ_t , and convergence typically requires only a small number of gradient-descent iterations per timestep.

After θ_t^* is found, the latent is updated deterministically:

$$\bar{z}_{t-1} = \hat{z}_{t-1}(\theta_t^*) \quad (3.8)$$

and the procedure advances to timestep $t-1$. The process is initialised with $\bar{z}_T = z_T^*$ and repeated sequentially until $t = 1$. Upon completion, the set of optimised null-text embeddings $\{\theta_t^*\}_{t=1}^T$ together with the noise pivot z_T^* form an invertible representation of the input image: under this

canonical per-timestep optimisation, replaying the guided reverse process with these embeddings and the original conditioning $\mathcal{C}$ reproduces the input image with high fidelity.

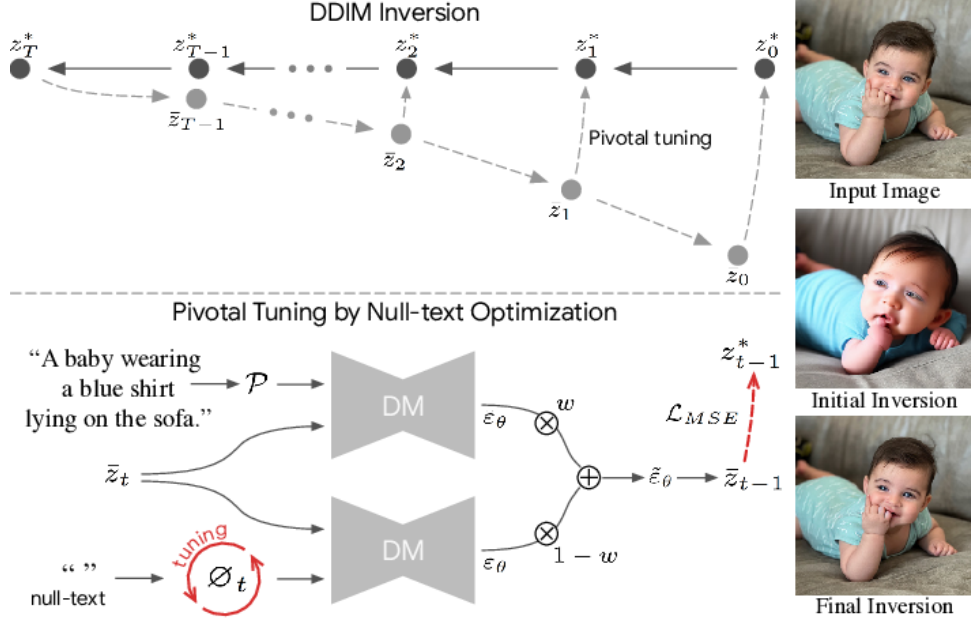


Figure 3.2: Null-Text Optimisation procedure[36]. The DDIM inversion trajectory $\{z_0^*, \dots, z_T^*\}$ is computed once and fixed. At each reverse timestep t , the null-text embedding θ_t is optimised so that the guided DDIM step lands on the corresponding inversion target z_{t-1}^* . After optimisation, this full per-timestep set of embeddings $\{\theta_t^*\}$ enables near-exact reconstruction of the original image under Classifier-Free Guidance.

3.2.3.3 Properties Relevant to Concept Editing

Two properties of this representation are critical for the downstream editing task:

1. **Almost exact reconstruction.** Because the null-text embedding is optimised to cancel the CFG drift, the guided reverse pass reconstructs the original image with substantially less of the blurring or structural distortion that would otherwise accumulate. In principle, this allows any difference between the reconstructed and the edited image to be attributed mostly to the change in the conditioning signal.
2. **Editability through prompt substitution.** The optimised embeddings $\{\theta_t^*\}$ encode the *anatomical identity* of the patient, while the conditioning signal $\mathcal{C}$ controls the *semantic content*. Replacing $\mathcal{C}$ with a new prompt, for instance one containing the learned concept token v^* (Section 3.2.4), steers the generation toward the target concept while the null-text embeddings keep the trajectory anchored to the patient’s anatomy.

The interplay between the fixed inversion trajectory and the optimised null-text embedding can be understood as a division of labour: the trajectory encodes *where* the anatomy is, while the prompt

controls *what* appears. Section 3.2.5 formalises how this division is exploited to inject a target clinical concept during the reverse diffusion process.

3.2.4 Concept Token Training

The Null-Text Optimisation procedure described in Section 3.2.3 provides a mechanism to reconstruct any real image from its noise pivot while leaving the conditioning prompt free to be modified. To exploit this freedom for concept editing, the prompt must contain a textual element that reliably steers the generation toward the target clinical concept. Hand-crafted descriptions are not enough: the alignment between general-domain text encoders and fine-grained medical concepts is weak, and the resulting prompts typically fail to produce the intended pathology (Section 2.2.4).

To overcome this semantic gap, we adopt the Textual Inversion (TI) framework of Gal et al. [14] and learn a dedicated concept token v^* directly from real images of the target pathology. This section formalises the training procedure and introduces the masked-loss variant we used.

3.2.4.1 Textual Inversion Background

Textual Inversion treats the embedding of a new token v^* as the only trainable parameter of a frozen text-to-image diffusion model. Given a small set of images $\{x^{(i)}\}$ depicting the same concept, the embedding is optimised so that the model, when conditioned on a prompt containing v^* , reconstructs the training images. Concretely, let $z = \mathcal{E}(x)$ be the latent encoding of a training image, and let $z_t = \sqrt{\alpha_t}z + \sqrt{1 - \alpha_t}\varepsilon$ be its noised version at timestep t with $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$. The standard Textual Inversion objective is the denoising loss restricted to the embedding parameter:

$$\mathcal{L}_{\text{TI}}(v^*) = \mathbb{E}_{x,t,\varepsilon} \left[\|\varepsilon - \hat{\varepsilon}_\theta(z_t, t, v^*)\|_2^2 \right] \quad (3.9)$$

where the conditioning is obtained by inserting v^* into a fixed prompt template (e.g. “an image of v^* ”), and $\hat{\varepsilon}_\theta$ denotes the noise prediction of the frozen U-Net. Crucially, the weights of the diffusion model and of the text encoder remain fixed; only the embedding vector v^* is updated. This makes TI extremely parameter-efficient.

3.2.4.2 Method Variants

We evaluate two strategies for selecting the training signal that defines what the token v^* should represent. Both share the objective of Equation (3.9) but differ in how the training images are prepared and in how the gradient is distributed across pixels.

M2 – Cropped patches. The most direct adaptation of natural-image TI to medical concept learning is to restrict the training set to image regions that *contain only the concept of interest*. Given a source image $p \in \mathcal{S}$ and its segmentation mask m , we crop a tight bounding box around the lesion and resize the resulting patch to the input resolution of the diffusion model. The token is then trained on this set of patches using the unmodified loss of Equation (3.9).

The appeal of this variant is its simplicity: the model never sees healthy tissue during training, so the gradient signal is, by construction, concentrated on the concept. Its limitation is equally direct: the token learns the concept *in isolation*, without any information about how the pathology integrates with the surrounding anatomy. We hypothesise that this divergence between the training and inference contexts limits concept transferability; this hypothesis is examined empirically in Chapter 5.

M3 – Whole images with masked loss. To preserve the contextual integration between the pathology and its anatomical surroundings, we train the token on *whole* source images while restricting the gradient signal to the lesion region through the segmentation mask. Formally, let M be the binary mask m resized to the spatial resolution of the latent representation. The masked Textual Inversion objective is:

$$\mathcal{L}_{M3}(v^*) = \mathbb{E}_{x,t,\varepsilon} \left[\left\| (\varepsilon - \hat{\varepsilon}_\theta(z_t, t, v^*)) \odot M \right\|_2^2 \right] \quad (3.10)$$

where $\odot$ denotes element-wise multiplication. The Equation (3.10) reduces to the standard TI objective (Equation 3.9) when M is everywhere one, and to no update when M is everywhere zero. In the intermediate regime, it performs a selective backpropagation: the model still receives the whole noised image as input (so its internal representations are contextualised by the surrounding anatomy), but only the pixels inside the mask contribute to the gradient that updates v^* .

This decoupling between *what the model sees* and *what the model is rewarded for predicting* is the defining feature of the M3 variant. It is intended to produce a token whose embedding captures the pathology in its natural anatomical context, rather than as an isolated patch. At generation time, this contextual grounding is expected to let the token inject the concept into a new patient’s anatomy with fewer of the stitching artefacts and contextual inconsistencies typical of patch-based methods. Whether the masked-loss token transfers better than its patch-based counterpart is evaluated directly in Chapter 5.

3.2.4.3 Training Setup

Both variants share the same overall training procedure:

Backbone a publicly available Stable Diffusion checkpoint, kept frozen. All weights of the U-Net, the VAE encoder/decoder, and the text encoder are detached from the computational graph.

Trainable parameters. a single embedding vector $v^* \in \mathbb{R}^d$, where d is the dimensionality of the text encoder’s token embedding space. The vector is initialised from the embedding of a semantically related natural-language token.

Prompt template. during training and inference, the token is wrapped in a fixed template of the form “an image of v^* ”.

Optimisation. the embedding is updated by stochastic gradient descent on the respective loss (Equation 3.9 for M2, Equation 3.10 for M3). Specific hyperparameters are reported in Chapter 4. After training, the learned token v^* is plugged into the generation pipeline described in Section 3.2.5, where it is combined with the optimised null-text embeddings of an unseen patient image to synthesise a concept-positive version of that patient’s anatomy.

3.2.5 Reverse Diffusion with Concept Injection

The components developed in Sections 3.2.2 through 3.2.4 provide everything required to synthesise a concept-positive image for an unseen patient. This section describes how they are assembled into the inference pipeline that produces, given a healthy patient image and a trained concept token v^* , a synthetic image in which the target pathology has been inserted.

3.2.5.1 Inputs and Pre-Processing

The inference procedure takes three inputs:

- A healthy patient image $x_h \in \mathcal{H}$, encoded into the latent space of the diffusion model as $z_0^{(h)} = \mathcal{E}(x_h)$;
- A trained concept token v^* , obtained from the procedure of Section 3.2.4;
- The frozen Stable Diffusion backbone (U-Net, VAE encoder/decoder and text encoder).

No segmentation mask, lesion patch, or pathological reference image is required at inference time. The only patient-specific information used by the pipeline is the healthy image itself.

3.2.5.2 Inversion and Null-Text Optimisation of the Target Image

The first stage applies the procedures of Sections 3.2.2 and 3.2.3 to the healthy image x_h . The inversion produces the pivot $z_{t_{\text{start}}}^{(h)}$ together with the inversion trajectory up to t_{start} , and the Null-Text Optimisation yields the embedding $\theta^{(h),*}$ that allows the guided reverse pass to reconstruct x_h with high fidelity.

At this point, the pipeline holds an invertible representation of the patient’s anatomy: replaying the guided reverse process with the original conditioning reconstructs x_h . The synthetic edit is obtained by replacing the conditioning prompt with one that contains the concept token.

3.2.5.3 Concept-Conditioned Reverse Pass

Let $\mathcal{C}_{v^*}$ denote the conditioning embedding obtained by inserting v^* into the same prompt template used during token training (Section 3.2.4.3), and let w be the Classifier-Free Guidance scale. Starting from the pivot $\bar{z}_{t_{\text{start}}} = z_{t_{\text{start}}}^{(h)}$, the reverse process iterates the guided DDIM update from $t = t_{\text{start}}$ down to $t = 0$:

$$\tilde{\epsilon}_t = \epsilon_\theta(\bar{z}_t, t, \theta^{(h),*}) + w \left[\epsilon_\theta(\bar{z}_t, t, \mathcal{C}_{v^*}) - \epsilon_\theta(\bar{z}_t, t, \theta^{(h),*}) \right] \quad (3.11)$$

The structure of Equation (3.11) mirrors the original Classifier-Free Guidance formulation (Equation 3.5), but with two key substitutions:

1. The unconditional branch is conditioned on the patient-specific null-text embedding $\theta^{(h),*}$ rather than on the generic empty embedding. This is what anchors the trajectory to the patient’s anatomy.
2. The conditional branch carries the concept embedding $\mathcal{C}_{v^*}$ instead of an arbitrary text prompt. The CFG extrapolation therefore moves the trajectory *away from the patient’s anatomy alone* and *toward the anatomy expressing the target concept*.

The final concept-positive image is recovered by decoding the last latent: $x^+ = \mathcal{D}(\bar{z}_0)$.

3.2.5.4 Why the Edit is Localised

The localisation of the edit follows from the asymmetric roles played by the two branches of Equation (3.11). The null-text embedding $\theta^{(h),*}$ is optimised to reproduce the patient’s anatomy with high fidelity: in the regions of the image where the concept does not apply, the conditional and unconditional predictions of the U-Net are expected to agree closely, so that the CFG term in Equation (3.11) is small and the trajectory stays near the reconstruction path. Only where the concept token introduces a semantic change should the two predictions diverge, producing a non-zero CFG contribution that locally modifies the latent. No explicit spatial constraint is imposed; the intended effect is that the interplay between the patient-anchored null embedding and the concept-anchored conditional embedding confines the modification to the semantically relevant regions. Whether the edit is in fact localised is not assumed but examined empirically, through difference maps between each healthy input and its concept-injected image. Figure 3.3

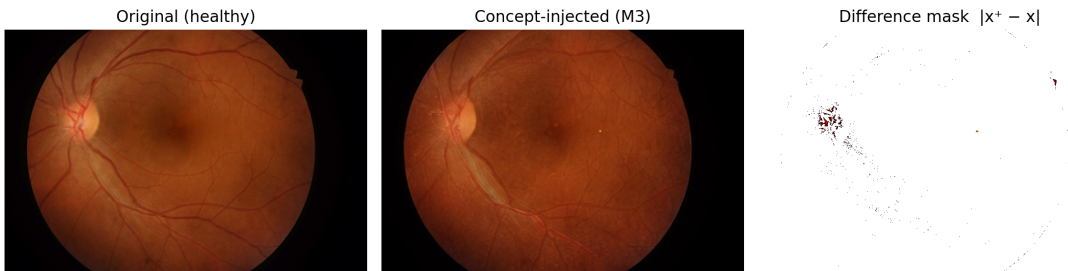


Figure 3.3: The proposed pipeline on a fundus image. Left: the healthy input. Centre: the concept-injected output (M3), in which hard exudates appear on the retinal background while the vessels and optic disc are left largely unchanged. Right: the absolute difference between the two, showing that the change is concentrated in a few focal regions.

3.2.5.5 Strategy Variants Revisited

The inference pipeline above defines the M3 method when the concept token is trained with the masked loss (Equation 3.10), and the M2 method when the token is trained on cropped patches with the standard loss (Equation 3.9). The M1 variant follows the same pipeline with one change: $\mathcal{C}_{v^*}$ is replaced by the embedding of a hand-crafted textual description of the pathology, and no token training is performed. All three methods share the same partial inversion, Null-Text Optimisation and CFG-based reverse pass; they differ only in the source of the conditioning embedding fed to the conditional branch of Equation (3.11). This design isolates the contribution of the concept token training strategy from the rest of the pipeline, allowing the experimental section to attribute performance differences directly to the learned concept representation.

Chapter 4

Experimental Setup

4.1 Datasets

The pipeline is validated on three publicly available medical-imaging datasets, each contributing a distinct imaging modality and a distinct target concept. The three were selected to span a range of *expected* distances between the imaging modality and the natural-image distribution on which Stable Diffusion 1.5 was pre-trained: colour fundus photography and dermoscopy are colour photographs and are therefore expected to lie closer to that distribution, whereas B-mode ultrasound: greyscale, speckled, and without a natural-image counterpart is expected to lie farther from it. This spread is a deliberate design choice; whether concept injection degrades with modality distance is treated as an empirical question (Chapter 5).

Across all three datasets the task is cast as a binary distinction between a `healthy` class (free of the target concept) and a `diseased` class (exhibiting it). Each dataset is reduced to three stem-aligned directories: `healthy/`, `diseased/` and `masks/`, by a dataset-specific staging script that creates symbolic links rather than copies, so that no image is duplicated on disk and the correspondence between an image and its lesion mask is given by a shared file stem. Only the `healthy` and `diseased` directories are used to train the supporting classifier; the masks serve the copy-paste, inpainting, crop-extraction and masked textual-inversion steps. Table 4.1 summarises the three datasets.

Table 4.1: The three evaluation datasets. Each spans a different imaging modality and target concept; class sizes are the staged `healthy/diseased` counts (for ISIC 2018 the class membership rule is given, as the counts depend on the area threshold).

Dataset	Modality	Target concept	Healthy	Diseased
IDRiD	Fundus photography	Hard exudates	168 (Grade 0)	54 (Task A)
ISIC 2018	Dermoscopy	Pigment network	906	1369
BUSI	Breast ultrasound	Tumour (benign + malignant)	133 (normal)	647 (437+210)

4.1.1 IDRiD - Diabetic Retinopathy (Fundus)

IDRiD [42] provides high-resolution colour fundus photographs (4288×2848 px). The diseased class is taken from the Lesion Segmentation sub-task (Task A): its 54 training images each carry a pixel-level binary mask for *hard exudates*, the target concept. Hard exudates were chosen over the other annotated lesions (microaneurysms, haemorrhages, soft exudates, optic disc) because they combine a comparatively large mask area with a visually distinctive appearance, bright yellow lipid deposits on the retinal background. The healthy class is drawn from the Disease Grading sub-task (Task B): images graded 0 (no diabetic retinopathy) that are not present in Task A are used as healthy (168 images). The overlap exclusion prevents a graded-0 image that also carries lesion annotations from contaminating the healthy class. Both classes share the same acquisition protocol and field of view, so any visual difference between them is attributable to pathology rather than acquisition.

4.1.2 ISIC 2018 - Dermoscopy

ISIC 2018 Task 2 [8] provides dermoscopic images annotated with five lesion-attribute masks: pigment network, negative network, streaks, milia-like cysts, and globules. The target concept is the *pigment network*, selected as the most frequent attribute with the largest typical mask area. An image is assigned to the diseased class when its pigment-network mask covers at least 0.5% of the image area, a threshold that discards marginal annotations; it is assigned to the healthy class only when *all five* attribute masks are empty. Images that carry some other attribute but not a pigment network are excluded from both classes, so that the healthy class is free of *any* annotated attribute and a synthetic edit can be attributed to the injected concept alone. An important caveat must accompany these results: “healthy” here does *not* mean healthy skin: every ISIC image depicts a skin lesion, and the binary task is therefore “does this lesion exhibit a pigment network?” rather than “lesion versus normal skin”.

4.1.3 BUSI - Breast Ultrasound

BUSI [12] contains breast ultrasound images in three categories: 133 normal, 437 benign, and 210 malignant. The healthy class is the 133 normal scans; the diseased class merges the benign and malignant tumours into a single 647-image pool, because the binary task asks whether a mass is present rather than which subtype it is, and merging makes all annotated tumours available for concept training. Both tumour categories ship with segmentation masks, which are linked into `masks/` by stem. The resulting 647:133 ($\approx 5:1$) class imbalance is carried through to the evaluation, where it motivates reporting balanced accuracy and AUC rather than raw accuracy (Section 4.2.5).

4.2 Evaluation Pipeline

Evaluating synthetic images for this use case cannot rely on image-quality metrics alone, because a synthetic image can be visually realistic and still fail to encode the clinical concept in a useful way. Conversely, a synthetic image can be perceptually imperfect and still serve as effective concept supervision.

To address this, the evaluation framework adopted in this work organises the analysis around four complementary questions:

1. How close is the synthetic distribution to the real distribution?
2. Are the synthetic images recognised as the target pathology by a model trained on real data?
3. Does the concept encoded in the synthetic samples transfer to real data?
4. Does the synthetic concept improve a downstream model under spurious correlations?

Each question is answered by a dedicated metric: Fréchet Inception Distance (FID), Target Class Accuracy (TCA), Concept Learnability (CL), and Robustness via LCR regularisation (ΔLCRReg). This section describes how each metric is computed, the supporting infrastructure required, and the reproducibility choices that make the comparison across synthesis methods meaningful.

4.2.1 Supporting Semantic Classifier

Three of the four evaluation metrics depend on a supporting classifier trained on real data. This classifier is not the final model under evaluation: it acts as the scoring model for TCA and as the feature extractor for Concept Learnability and for the LCR experiment. Centralising these roles in a single backbone removes a source of variability between metrics: the same representation space is used to classify synthetic images, to project them onto the CAV, and to extract the features regularised by LCR.

Architecture. The backbone is a ResNet-50 [18] initialised with ImageNet-pretrained weights. The final fully connected layer is replaced by a binary classification head:

$$\text{head}(z) = \text{Linear}(\text{Dropout}_{0.5}(z)), \quad (4.1)$$

where z is the 2048-dimensional penultimate feature. The same 2048-dimensional representation is later reused for all feature-based metrics.

Training. The classifier is trained on the real `healthy` and `diseased` image folders using cross-entropy loss and the Adam optimiser, with early stopping on validation loss. The train/validation split is deterministic, controlled by a fixed seed, and the resulting checkpoint is reused unchanged by every downstream metric. The specific hyperparameters are reported in Section 4.3.

4.2.2 Perceptual Fidelity: Fréchet Inception Distance

The first pillar quantifies the distributional similarity between the synthetic images and the real images of the target class. The Fréchet Inception Distance [20] models the activations of an Inception-v3 network as a multivariate Gaussian for each set and computes the Fréchet distance between the two Gaussians:

$$\text{FID} = \|\mu_r - \mu_s\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_s - 2(\Sigma_r \Sigma_s)^{1/2}\right), \quad (4.2)$$

where (μ_r, Σ_r) and (μ_s, Σ_s) are the mean and covariance of the 2048-dimensional Inception features for the real and synthetic sets, respectively. Lower FID indicates closer distributional alignment.

Implementation. FID is computed with the TORCHMETRICS implementation using the standard 2048-feature Inception block.

Choice of reference distribution. The real reference set used in Equation (4.2) is the real *diseased* folder rather than the full real dataset. Because the synthetic images are concept-positive by construction, comparing them against the real diseased distribution answers the question that matters for this work: “how close are the synthetic samples to the real population they are intended to represent?”.

Limits. FID measures distributional similarity in a generic feature space and does not test clinical correctness. Low FID is important but not sufficient: a model that reproduces the global appearance of the real diseased set without correctly forming the concept can still achieve a competitive FID. The remaining three pillars are designed to expose exactly this kind of failure.

4.2.3 Semantic Validity: Target Class Accuracy

The second pillar verifies that the synthetic images are recognised as the intended pathology by a classifier trained on real data. Whereas FID asks whether the synthetic set *looks* like the real diseased set in aggregate, TCA asks whether each individual synthetic image is classified as diseased.

Definition. Let f_{cls} denote the supporting classifier of Section 4.2.1 and let $p(\text{diseased} | x)$ be the softmax probability assigned to the target class. Given a threshold τ , Target Class Accuracy is defined as the fraction of synthetic samples whose target-class probability exceeds τ :

$$\text{TCA} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\left[p(\text{diseased} | x_{\text{syn}}^{(i)}) > \tau\right], \quad (4.3)$$

with $\tau = 0.5$ in our experiments. A high TCA indicates that the synthesis method does not merely reproduce the background texture of the diseased class but produces features that the classifier identifies as pathological.

Interpretation alongside FID. TCA and FID are complementary. A method can produce images with low FID (close to the real diseased distribution) but low TCA (the specific discriminative features are absent), or vice versa. The combination is informative: high FID with high TCA indicates concept generation without distribution matching, low FID with low TCA indicates style matching without semantic content.

4.2.4 Concept Learnability

Target Class Accuracy verifies that the synthetic images are classified as diseased *by a model trained on real data*, but it does not test whether the concept encoded in the synthetic samples generalises to real data when treated as concept supervision. The Concept Learnability metric directly addresses this gap.

Protocol. The metric trains a linear probe to distinguish the synthetic concept from real healthy tissue, and then evaluates the probe on real diseased versus real healthy images. Features are extracted from the penultimate layer of the supporting classifier of Section 4.2.1, yielding 2048-dimensional representations.

The real healthy set is split deterministically into two halves: one half is used at training time as the negative class, and the other half is kept exclusively for testing.

- **Training set:** synthetic positive images (label 1) and half of the real healthy images (label 0).
- **Test set:** all real diseased images (label 1) and the held-out half of the real healthy images (label 0).

The probe is a logistic regression classifier on standardised features, trained with balanced class weights and a fixed random state for reproducibility. The reported metric is the ROC–AUC on the real test set.

What this metric reveals. A high test ROC–AUC indicates that the synthetic positives occupy a region of feature space that aligns with the region occupied by real diseased samples, despite the probe never having seen a real diseased image. This is the strongest direct evidence that the synthesis method captures a clinically meaningful concept rather than a synthetic-vs-real domain shift. Conversely, a low ROC–AUC indicates that the synthetic concept is separable from healthy tissue in feature space, but along a direction that does not correspond to the real disease axis.

Real upper bound. To calibrate the synthetic scores, the same probe protocol is additionally run with the synthetic positives replaced by *real* diseased images: the probe is trained on one half of the real diseased set, together with the same training half of the real healthy images, and tested on the held-out halves of both classes. Because the diseased training and test halves are disjoint, this reference introduces no leakage. The resulting ROC–AUC is the upper bound attainable in this

feature space with genuine positives; the ratio between the synthetic ROC–AUC and this upper bound expresses how much of the achievable concept signal each synthesis method recovers.

4.2.5 Robustness via LCRReg

The final pillar evaluates the downstream utility of the synthetic concept as supervision for a concept-based regulariser. It is the most demanding test in the framework, in principle, because it places the synthetic data in a realistic role and asks whether it improves a model on a task where the model would otherwise rely on a spurious shortcut; its statistical power in practice is examined in Chapter 5.

4.2.5.1 Experimental Design

The robustness protocol compares two classifiers trained on the *same biased data*:

- a **baseline** model trained with cross-entropy alone;
- an **LCR-regularised** model trained with cross-entropy augmented by the LCRReg loss [11] computed from a Concept Activation Vector (CAV) derived from the synthetic images produced by each generation method.

Both models share the same architecture, optimisation setup, and biased training data. They differ only in the presence of the LCRReg term, and the LCR model further differs across synthesis methods only through the CAV used to compute the loss.

4.2.5.2 Bias Injection

The real dataset is first partitioned into training, validation and test splits. The split is deterministic and is controlled by a *dataset seed* held fixed across all robustness runs, ensuring that no image migrates between splits.

A label-correlated artificial bias is then injected into the training and validation samples. The injection follows a Bernoulli rule with a correlation parameter ρ : with probability ρ the bias is aligned with the label, and with probability $1 - \rho$ it is misaligned. A typical setting of $\rho = 0.95$ makes the bias an almost-perfect predictor during training and validation, so a model trained without regularisation is strongly encouraged to learn the shortcut.

The test set is left unbiased. As a consequence, a classifier that relies on the shortcut performs well during training and validation but collapses on the test set, while a classifier that learns the true clinical signal generalises.

4.2.5.3 CAV Computation

The CAV is computed from the same feature space as the supporting classifier (2048-dimensional ResNet-50 penultimate features [18]). The positive set consists of synthetic concept images produced by the synthesis method under evaluation, and the negative set consists of a held-out fraction

of the real healthy images that is excluded from the biased training, validation and test splits. The hold-out is essential: reusing healthy images that participate in the biased experiment would create leakage between the concept definition and the downstream evaluation.

The CAV is the unit-norm normal vector to the linear SVM decision boundary that separates the two sets and is computed once per robustness run, independently of the training seed used downstream. This SVM-derived CAV plays a different role from the logistic-regression probe of Section 4.2.4: the probe is a full linear classifier used to measure concept transfer, whereas the CAV is a single direction used to regularise the downstream model.

4.2.5.4 CAV Diagnostics

Before using the CAV in LCRReg, three diagnostics characterise the quality and the failure modes of the concept direction.

Linear separability. The accuracy of the SVM used to compute the CAV is recorded as a sanity check. Because synthetic positives and real healthy images typically differ along multiple directions, this value is almost always near 100% and is reported only to confirm that the SVM problem is well posed.

Clinical AUC. The most informative diagnostic projects real diseased and real healthy images onto the CAV and computes the ROC–AUC for separating the two classes. An AUC close to 1.0 means that the CAV captures a direction that is clinically meaningful in the real data; an AUC close to 0.5 means that the CAV primarily captures a synthetic-vs-real domain shift rather than the disease concept. Clinical AUC is the primary quality indicator for the CAV.

Bias alignment. The third diagnostic measures the degree to which the artificial bias is entangled with the CAV. For each sample, the code computes the projection on the CAV of the clean image, $p_{\text{clean},i}$, and of the biased version, $p_{\text{biased},i}$, and records the difference:

$$\Delta_i = p_{\text{biased},i} - p_{\text{clean},i}. \quad (4.4)$$

The mean of Δ_i is computed separately for healthy and diseased samples and summarised into an alignment score. The sign and magnitude of this score determine the regime in which LCRReg operates:

- near zero: the bias is approximately orthogonal to the concept, and LCRReg can in principle deflect the shortcut without interfering with the concept;
- positive: the bias pushes features in the CAV direction, meaning LCRReg may inadvertently reinforce the shortcut;
- negative: the bias pushes against the CAV, so LCRReg pulls away from the shortcut even more strongly.

Concept effect size. Alongside the clinical AUC, the magnitude of the separation between the two projected populations is summarised by Cohen’s d , the standardised mean difference between the CAV projections of the real diseased and real healthy images:

$$d = \frac{\bar{p}_{\text{dis}} - \bar{p}_{\text{health}}}{s_{\text{pooled}}}, \quad s_{\text{pooled}} = \sqrt{\frac{1}{2}(s_{\text{dis}}^2 + s_{\text{health}}^2)}, \quad (4.5)$$

where $\bar{p}_{\text{dis}}$ and $\bar{p}_{\text{health}}$ are the mean projections onto the CAV of the real diseased and real healthy sets and s_{dis} , s_{health} their standard deviations. A large positive d means the diseased images project well onto the concept-positive side of the CAV; a negative d indicates an anti-aligned direction. Whereas the clinical AUC measures the rank separation of the two projected sets, d quantifies its magnitude in pooled standard-deviation units; the two are reported together throughout Chapter 5.

4.2.5.5 Baseline and LCR Models

Both the baseline and the LCR-regularised model use the same backbone as the supporting classifier of Section 4.2.1, fine-tuned on the biased training data. The baseline is trained with cross-entropy alone:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{cls}}. \quad (4.6)$$

The LCR model adds the LCRReg penalty:

$$\mathcal{L}_{\text{LCR}} = \mathcal{L}_{\text{cls}} + \lambda \cdot \frac{1}{|P|} \sum_{i \in P} (1 - \cos(f_i, c)), \quad (4.7)$$

where f_i is the 2048-dimensional feature of training sample i , c is the CAV, P is the subset of the batch with positive label, and λ controls the strength of the regularisation. We adopt the *positive-only* mode: the cosine alignment is enforced only on diseased samples, which is the regime in which LCRReg is intended to pull the positive cluster towards the CAV direction [11].

Baseline caching. Because the baseline model depends only on the dataset content, the bias configuration, and the training hyperparameters, it is cached and reused across synthesis methods sharing those settings. The cache key is a hash of the relevant settings together with a fingerprint of the data. This is methodologically important: it removes any drift in baseline performance between synthesis methods, so the ΔLCRReg measured below reflects only the difference introduced by the synthetic-derived CAV.

4.2.5.6 Per-Seed Comparison

For each training seed, the baseline and the LCR model are evaluated on the same unbiased test set. Two deltas are recorded:

$$\Delta_{\text{raw}} = \text{Acc}_{\text{LCR}} - \text{Acc}_{\text{base}}, \quad \Delta_{\text{bal}} = \text{BalAcc}_{\text{LCR}} - \text{BalAcc}_{\text{base}}. \quad (4.8)$$

Raw accuracy is the percentage of correctly classified test samples; balanced accuracy is the mean of the per-class accuracies and is robust to class imbalance in the test set. Both are reported, but the discussion in Chapter 5 focuses on the balanced delta, which is the more conservative quantity.

4.2.5.7 Multi-Seed Aggregation

The unbiased test set is small, so a single training seed produces noisy delta estimates. To stabilise the comparison, the LCR model is re-trained for K training seeds, while the dataset seed remains fixed. This separation is deliberate: holding the dataset constant isolates training stochasticity from data stochasticity, so the resulting variance reflects the optimisation path of the LCR model rather than incidental properties of a particular split or bias placement.

For each delta in Equation (4.8), the pipeline reports the mean, standard deviation, minimum, maximum, and number of seeds. The mean and standard deviation are used as the main summary statistics in the results chapter.

Shortcut-filtered aggregation. The LCRReg loss can only deflect a shortcut that the baseline actually learned. In seeds where the baseline already generalises to the unbiased test set, Δ_{bal} is close to zero by construction, and including these seeds in the aggregation dilutes the signal that the experiment is designed to measure.

To address this, the pipeline computes a second summary restricted to the seeds where the baseline failed to generalise, operationalised as

$$\text{Acc}_{\text{base}} < \tau_{\text{shortcut}}, \quad (4.9)$$

with $\tau_{\text{shortcut}} = 0.85$ in our experiments. The shortcut-filtered summary is the one to interpret when discussing the *actual effect of LCRReg on shortcut learning*, while the all-seeds summary documents the overall stability of the method.

4.2.6 Reproducibility

Every stochastic component of the pipeline is governed by an explicit seeding policy. The Python, NumPy, and PyTorch (CPU and CUDA) random number generators are seeded from a single base seed declared in the configuration file, and the cuDNN backend is set to deterministic mode. Data loader workers receive deterministic seeds derived from the PyTorch initial seed, which is necessary because data augmentations may rely on the Python or NumPy RNGs internally.

For the robustness experiment, the seeding strategy distinguishes explicitly between the *dataset seed*, which governs the train/validation/test split and the bias placement, and the *training seed*, which governs only the stochastic optimisation path. The dataset seed is held fixed across all robustness runs; the training seed is swept across K values starting from the base seed. This design guarantees that, across synthesis methods, every model is trained on identical data with identical bias and evaluated on an identical test set, so the differences attributable to each method are not confounded by data variance.

4.3 Implementation Details

This section consolidates the requirements to reproduce the experiments. All generation steps use the same frozen Stable Diffusion 1.5 backbone (`runwayml/stable-diffusion-v1-5`); no weights of the U-Net, VAE or text encoder are updated at any stage. The generation code is maintained as a separate codebase from the evaluation pipeline described earlier in this chapter.

Generation pipeline. Table 4.2 lists the hyperparameters of the four generation stages: lesion-crop extraction, concept-token training (M2 and M3), inversion, and concept-conditioned injection. The forward trajectory is truncated at $t_{\text{start}} = 300$ of the 1000-step schedule, and a null-text embedding is optimised for up to 100 iterations per image. As a fidelity check, the inversion step records the PSNR and SSIM between each healthy image and its reconstruction. The number of inverted healthy images is capped at 200, which uses all available healthy images for IDRiD and BUSI and a 200-image subset for ISIC 2018; the copy-paste and inpainting baselines instead process every healthy image, so the number of synthetic samples differs across methods and is reported with the results.

Table 4.2: Generation-pipeline hyperparameters. The backbone is a frozen Stable Diffusion 1.5; the only trainable parameter in M2/M3 is the concept-token embedding.

Stage	Parameter	Value
Lesion-crop extraction	output resolution	512×512
	bounding-box margin	0.2
	min. component area	40 mask px
Concept token (M2, M3)	trainable parameters	1 token embedding
	initialiser token	<code>lesion</code>
	learning rate	5×10^{-4}
	training steps	3000
	resolution	512×512
Inversion	t_{start} / schedule	300 / 1000
	timestep interval	100
	null-text learning rate	0.1
	max. optimisation steps	100
	guidance scale (inversion)	1.5
	precision	fp32
Concept injection	guidance-scale sweep	$\{1.5, 3, 5, 7.5, 10\}$
	inverted healthy images	≤ 200

Supporting classifier and robustness models. The supporting classifier (Section 4.2.1) and the baseline and LCR-regularised models of the robustness experiment share the same ResNet-50 backbone. Their training hyperparameters, together with those of the LCR protocol, are given in Table 4.3.

Table 4.3: Evaluation hyperparameters: the supporting/robustness classifier and the LCR robustness protocol (Section 4.2.5).

Component	Parameter	Value
Classifier	backbone	ResNet-50
	classification head	Linear $\circ$ Dropout _{0.5}
	optimiser	Adam
	learning rate	1×10^{-3}
	batch size	32
	max. epochs	20 (early stop, patience 5)
	train/validation split	80/20
LCR robustness	bias type	colour tag
	bias correlation ρ	0.95
	CAV estimator	linear SVM
	healthy hold-out (CAV neg.)	0.15
	LCR weight λ	0.1
	LCR mode	positive-only
	number of training seeds K	5
	dataset seed	fixed across runs

4.4 Concept-Based Regularisation Setup

The concept-based regulariser used in the robustness evaluation (Section 4.2.5) is LCRReg [11]. Its formulation, a cosine-alignment penalty between the positive-class features and the synthetic-derived Concept Activation Vector, added to the cross-entropy loss, is given in Section 4.2.5.5 (Equation 4.7), and its hyperparameters are listed in Table 4.3. We adopt the *positive-only* variant with $\lambda = 0.1$: the alignment is enforced only on diseased samples, the regime in which the regulariser is intended to pull the positive cluster towards the concept direction. The CAV that parameterises the penalty is recomputed for each synthesis method, from that method’s synthetic positives and a fixed held-out set of real healthy images (Section 4.2.5.3), so that the only quantity that varies across the LCR comparison is the synthetic-derived concept direction. This isolates the effect of each synthesis method on downstream robustness from any variation in the regulariser itself.

Chapter 5

Results and Discussion

This chapter reports the evaluation of the proposed pipeline and the two baselines across the three datasets of Chapter 4, organised around the four evaluation axes (FID, TCA, Concept Learnability, and ΔLCRReg) together with two CAV diagnostics introduced by the robustness protocol (clinical AUC and the projection effect size). Section 5.1 fixes the reporting conventions; Section 5.2 presents the per-dataset results; Sections 5.3–5.7 analyse each axis across datasets; Section 5.8 synthesises the three datasets; and Section 5.9 discusses the findings.

5.1 Reporting Protocol

Every synthesis method is evaluated with an identical protocol. The three injection variants (M1–M3) are run over a guidance sweep $w \in \{1.5, 3, 5, 7.5, 10\}$; the two baselines and a *real* reference have no guidance. The *real* row provides the empirical ceiling for TCA, Concept Learnability and the robustness CAV. This *real* row is computed in-sample (all real diseased images used as the positive set) and is therefore distinct from the leakage-free Concept Learnability *upper bound* of Section 4.2.4 (a probe trained on one half of the real diseased set and tested on the held-out half), which is the value quoted as the upper bound in each table caption; the two coincide on IDRiD but diverge on BUSI and ISIC. For the same reason, the real-image TCA is an optimistic ceiling, since the supporting classifier was trained on real diseased images. Tables 5.1–5.3 report the operating points $w \in \{1.5, 3, 5\}$, which bracket the behaviour of interest; the higher scales $w \in \{7.5, 10\}$ are discussed in the text but omitted from the tables, as on IDRiD they only raise FID further while the concept metrics have already saturated, and on the other two datasets the concept has already collapsed by $w = 5$.

Three quantities require care in interpretation and are stated alongside each metric: (i) the *Concept Learnability upper bound*, the ROC–AUC of a probe trained on real diseased images instead of synthetic ones, against which the synthetic ROC–AUC is read; (ii) the *CAV clinical AUC*, the ROC–AUC of real diseased vs. real healthy images projected onto the CAV, which tests whether the concept *direction* aligns with the real-disease axis; and (iii) Cohen’s d of that same projection gap, an effect size for the alignment.

5.2 Per-Dataset Results

5.2.1 IDRiD

On IDRiD the supporting feature space separates real diseased from real healthy almost perfectly: the Concept Learnability upper bound is $AUC = 1.000$ on a held-out split, and the null-text inversion reconstructs the healthy images with high fidelity ($PSNR = 35.8$ dB, $SSIM = 0.89$). Two consequences follow. First, CL AUC *saturates*: every method at $w \geq 3$ sits at 0.98–1.00 (Table 5.1), so CL AUC barely discriminates methods and the more informative quantities are TCA, the CAV clinical AUC, and Cohen’s d . Second, the real-image TCA is only 79.6%: the classifier, trained on a healthy-majority split, recognises roughly four in five *real* diseased images. TCA is effectively capped near 80%. (Figure 5.1)

Within this regime, guidance *helps*: as w increases from 1.5 to 3 the concept methods move from weak/anti-aligned to strongly aligned. The masked-loss token (M3) is the strongest concept method: at $w = 1.5$ it already reaches $CL = 0.999$ at one of the lowest injection FIDs (84.7, essentially tied with M1’s 84.5 and comparable to the baselines), and at $w = 3$ it attains the highest TCA on IDRiD (88.8%) with a strongly aligned CAV (clinical $AUC = 0.985$, $d = +1.94$). This value exceeds the real-image TCA (79.6%). Since the *real* row is intended as the empirical ceiling, surpassing it most likely reflects the supporting classifier over-recognising an over-expressed synthetic concept rather than a genuinely better one, and should not be read as the synthetic concept being “better than real”. M3 reaches a recognisable, aligned concept at much lower guidance than M2 or M1 (which produce $\leq 9.7\%$ TCA at $w = 3$) that is the clearest evidence in favour of the masked-loss design. Above $w = 5$, FID continues to rise (256.9 for M3 at $w = 10$) with no further gain in CL AUC or clinical AUC and a decline in TCA, so the higher scales are dominated.

The two baselines expose the central caveat of CL. Copy-paste attains a CL AUC of 0.997, its CAV is strongly *anti-aligned* with real disease (clinical $AUC = 0.093$, $d = -2.69$): a probe trained on copy-paste positives separates real disease, but the maximum-margin CAV direction points real diseased images to the *healthy* side. Diffusion inpainting (baseline 2) is the realistic-but-empty counter-example: its FID (69.3) matches copy-paste, yet no inpainted image is classified as diseased (TCA = 0%) and its CAV is likewise anti-aligned (clinical $AUC = 0.045$). These two rows already show that perceptual fidelity and concept validity are distinct.

Table 5.1: IDRiD results at $w \in \{1.5, 3, 5\}$. FID lower is better; TCA, CL AUC, clinical AUC higher is better; d is the Cohen’s d of the CAV projection gap (sign indicates orientation); ΔLCRReg is the balanced LCRReg delta (mean $\pm$ std over 5 seeds). CL upper bound AUC = 1.000; real-image TCA 79.6%.

Method	w	FID	TCA (%)	CL AUC	clin AUC	d	ΔLCRReg
Baseline 1 (copy-paste)	–	68.9	81.3	0.997	0.093	−2.69	−0.6 $\pm$ 10.9
Baseline 2 (inpaint)	–	69.3	0.0	0.857	0.045	−3.33	−1.4 $\pm$ 8.6
M1 (prompt)	1.5	84.5	0.0	0.063	0.231	−0.82	−4.6 $\pm$ 8.5
M1 (prompt)	3	104.2	4.5	0.992	0.964	+2.20	+0.3 $\pm$ 7.1
M1 (prompt)	5	150.8	42.5	0.992	0.956	+1.84	−4.0 $\pm$ 5.2
M2 (patches)	1.5	99.2	0.0	0.876	0.470	−0.01	−2.8 $\pm$ 8.5
M2 (patches)	3	124.9	9.7	0.994	0.971	+2.28	−2.1 $\pm$ 3.2
M2 (patches)	5	173.7	47.8	0.980	0.973	+2.25	+2.8 $\pm$ 6.0
M3 (masked)	1.5	84.7	24.6	0.999	0.561	+0.43	−3.0 $\pm$ 7.3
M3 (masked)	3	148.4	88.8	0.999	0.985	+1.94	−5.1 $\pm$ 6.9
M3 (masked)	5	194.4	86.6	0.998	0.992	+3.68	+3.4 $\pm$ 7.3
<i>real</i>	–	–	79.6	1.000	0.999	+5.20	−0.6 $\pm$ 6.1

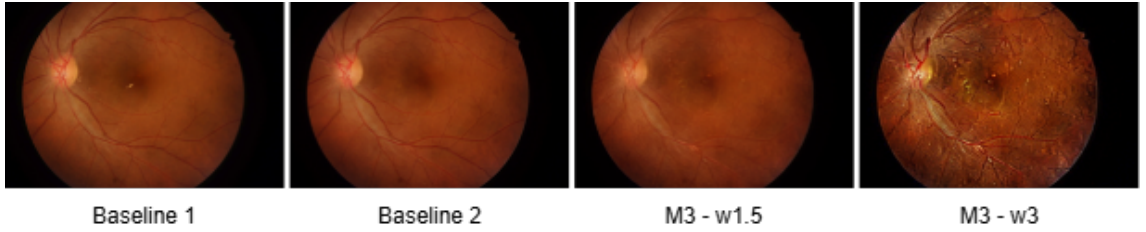


Figure 5.1: Qualitative comparison on IDRiD: the two baselines (copy-paste and inpainting), and the masked-loss token (M3) across the guidance sweep. As guidance grows the injected exudates become more pronounced while image fidelity degrades (FID 84.7 $\rightarrow$ 194.4), illustrating the guidance–concept trade-off of Table 5.1.

5.2.2 ISIC 2018

ISIC presents the lowest baseline FID of the study (62.7) and the inversion is faithful (PSNR = 32.9 dB). The obstacle is the *concept*: pigment network is the weakest concept of the three, with a Concept Learnability upper bound of only AUC = 0.846 (vs. 1.000 and 0.986), reflecting that the ISIC “healthy” class consists of attribute-empty *lesions* rather than normal skin.

On ISIC, guidance is monotonically harmful. Within *every* injection method, increasing w raises FID and lowers both CL and TCA: over the nine injection configurations $\rho(\text{FID}, \text{CL}) = -1.00$ and $\rho(\text{FID}, \text{TCA}) = -0.97$ (Table 5.2). These nine configurations are three methods $\times$

three guidance levels so the perfect coefficient reflects a near-deterministic *per-method* trend rather than nine independent samples, and no p -value is meaningful for the tied ranks. The reading is nonetheless unambiguous: the minimal edit ($w = 1.5$) is the best concept point for every method, the *opposite* of IDRiD. At $w = 1.5$ the masked token reaches $CL = 0.685$, matching copy-paste (0.691) at $\approx 81\%$ of the upper bound, with the lowest injection FID (125.3) and the highest TCA (46.5%); by $w = 5$ the concept has collapsed (CL 0.14–0.21, $TCA \leq 5\%$). A minimal edit thus *preserves* the concept at the level of pasting real pixels, but is not shown to *add* concept content beyond it.

The clinical AUC is uninformative for ranking on ISIC: even the real pigment-network CAV scores below chance ($AUC = 0.425$). This is an orientation effect: clinical AUC tracks the sign of the projection gap ($\rho = 0.74$), and the sign-corrected separation is 0.575 for the real concept and up to 0.716 for M3 at $w = 1.5$ so the concept is at best a *weak* linear direction, not necessarily an absent one. We cannot distinguish from these data whether the linear concept is genuinely absent or merely sign-flipped and weak; the orientation explanation is the more parsimonious. The practical consequence is that CL and TCA are the discriminating axes on ISIC, and both point clearly to $w = 1.5$.

Table 5.2: ISIC 2018 results at $w \in \{1.5, 3, 5\}$. CL upper bound $AUC = 0.846$ (the lowest of the three datasets); real-image TCA 85.0%. The *real*-row CL AUC (0.915) is the in-sample value and exceeds this leakage-free upper bound (see Section 5.1). The clinical AUC is anti-oriented even for the real concept (0.425) and is not used to rank methods (Section 5.5).

Method	w	FID	TCA (%)	CL AUC	clin AUC	d	$\Delta LCRR_{\text{reg}}$
Baseline 1 (copy-paste)	–	92.5	40.7	0.691	0.349	−0.62	+5.7 ± 6.5
Baseline 2 (inpaint)	–	62.7	24.7	0.683	0.393	−0.42	−1.4 ± 8.5
M1 (prompt)	1.5	132.3	39.5	0.534	0.350	−0.42	−0.3 ± 10.6
M1 (prompt)	3	159.1	22.0	0.377	0.617	+0.56	−3.7 ± 3.8
M1 (prompt)	5	206.2	5.0	0.214	0.481	+0.06	−2.2 ± 4.3
M2 (patches)	1.5	130.8	45.0	0.637	0.384	−0.63	+1.8 ± 11.3
M2 (patches)	3	179.1	14.5	0.235	0.530	+0.30	+3.1 ± 9.3
M2 (patches)	5	225.0	3.0	0.137	0.384	−0.55	+4.0 ± 7.7
M3 (masked)	1.5	125.3	46.5	0.685	0.284	−0.70	+5.6 ± 8.9
M3 (masked)	3	178.8	12.5	0.271	0.566	+0.41	+4.8 ± 10.5
M3 (masked)	5	218.2	1.5	0.146	0.405	−0.43	−4.7 ± 6.5
<i>real</i>	–	–	85.0	0.915	0.425	−0.35	−1.8 ± 8.5

5.2.3 BUSI

Here the inversion is less faithful than on IDRiD ($PSNR = 31.7$ dB, $SSIM = 0.85$); baseline FID is roughly double (127–152); and FID rises sharply at $w = 5$ (356–373 across the injection methods, Table 5.3). The concept never reaches the alignment seen on IDRiD: across all nine injection

configurations the clinical AUC stays ≤ 0.72 and Cohen’s $d \leq +0.72$, against the real concept’s $0.977 / +2.91$; at $w = 1.5$ the CAVs are in fact anti-aligned ($d \approx -1.1$). Concept Learnability separates the masked token from the others: M3 stays high across guidance (CL = 0.82–0.94, i.e. the concept remains *learnable* by a probe), whereas M1 and M2 are erratic and frequently below chance (CL = 0.36–0.46 at $w \geq 3$). On BUSI, then, the masked concept is learnable but not clinically aligned, while the patch and prompt concepts fail to form at all.

The only route to a high TCA on BUSI is destructive: TCA reaches 100% at $w = 5$ precisely where FID is 356–373 and the images are out-of-distribution texture. The diseased-majority class imbalance (647:133) further inflates TCA, the real-image TCA is 98.9%, so on BUSI a high TCA is not evidence of a valid concept. The masked-loss token is the strongest concept method here too (highest CL at every guidance, 0.82–0.94) but its CAV never aligns ($d \leq +0.50$): the learnability half of the masked-loss advantage partly survives the domain gap, but the alignment half does not. Copy-paste behaves differently from IDRiD: its CAV is domain-shift rather than anti-aligned (clinical AUC = 0.501, $d = +0.01$), confirming that the failure mode of a baseline is dataset-dependent.

Table 5.3: BUSI results at $w \in \{1.5, 3, 5\}$. CL upper bound AUC = 0.986; real-image TCA 98.9% (inflated by the 647:133 class imbalance).

Method	w	FID	TCA (%)	CL AUC	clin AUC	d	$\Delta\text{LCRR}_{\text{reg}}$
Baseline 1 (copy-paste)	–	127.5	85.7	0.921	0.501	+0.01	+1.0 $\pm$ 2.4
Baseline 2 (inpaint)	–	151.7	67.7	0.680	0.243	−0.86	+6.5 $\pm$ 5.7
M1 (prompt)	1.5	135.1	78.9	0.820	0.182	−1.12	+4.3 $\pm$ 9.5
M1 (prompt)	3	175.7	63.2	0.400	0.604	+0.35	−2.6 $\pm$ 10.7
M1 (prompt)	5	356.0	99.2	0.441	0.614	+0.41	+1.7 $\pm$ 4.6
M2 (patches)	1.5	131.8	52.6	0.734	0.220	−1.05	+0.1 $\pm$ 0.5
M2 (patches)	3	206.3	72.9	0.360	0.595	+0.33	+6.7 $\pm$ 12.1
M2 (patches)	5	358.4	100.0	0.455	0.718	+0.72	+3.3 $\pm$ 13.3
M3 (masked)	1.5	132.8	75.9	0.866	0.182	−1.15	+2.0 $\pm$ 8.1
M3 (masked)	3	204.1	75.9	0.821	0.595	+0.33	+3.4 $\pm$ 10.9
M3 (masked)	5	373.5	100.0	0.940	0.644	+0.50	−1.7 $\pm$ 11.9
<i>real</i>	–	–	98.9	0.989	0.977	+2.91	+0.4 $\pm$ 8.4

5.3 The Guidance–Concept Trade-off

Guidance controls the strength of the concept conditioning against adherence to the inverted healthy image. Its effect on the concept is *opposite* on IDRiD and ISIC, and this contrast is one of the more informative observations of our experiments. On IDRiD, raising w from 1.5 to 3 turns the concept on: TCA and clinical AUC rise sharply and the concept becomes recognisable and aligned. On ISIC, any increase in w degrades the concept monotonically (CL and TCA fall

in lockstep, $\rho(\text{TCA}, \text{CL}) = +0.97$). FID, by contrast, rises monotonically with w on *all* datasets (Figure 5.2).

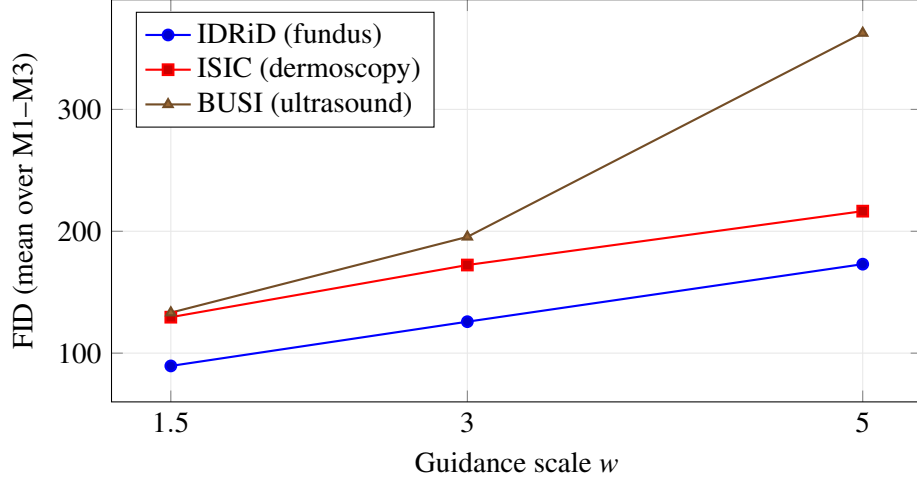


Figure 5.2: Perceptual fidelity degrades monotonically with the guidance scale on every dataset: stronger concept conditioning makes the synthetic images progressively less realistic. The effect is far steeper on BUSI, where the mean injection FID climbs from ~ 133 at $w = 1.5$ to ~ 363 at $w = 5$, well above the copy-paste realism floor of 127.5 whereas on IDRiD and ISIC, FID rises but stays bounded. Values are the mean over the three injection variants (M1–M3); per-method numbers are in Tables 5.1–5.3.

A plausible account of the sign flip is that the optimal guidance depends on the salience of the target concept. The observed pattern is consistent with this: stronger conditioning renders the bold, focal hard exudates more clearly but overwrites the faint reticular pigment network with generic lesion content. We did not test the mechanism directly, and concept type is confounded with modality across our three datasets, so we offer the salience account as a hypothesis rather than a conclusion. If this holds, the operating point should be chosen per concept rather than fixed, and minimal guidance is the safe default precisely where the concept is subtle. The result is also a direct demonstration of the minimal-counterfactual goal of Chapter 3: on ISIC the barely-perceptible edit carries as much concept as is available, and any larger edit only destroys it.

5.4 Perceptual Fidelity versus Concept Validity

Across all three datasets FID and the concept metrics tell different stories, the divergence that motivates a multi-axis evaluation. Diffusion inpainting is the clearest case: it produces the most realistic images on ISIC (FID = 62.7) and matches copy-paste on IDRiD, yet it carries little or no recognisable concept (TCA = 0% on IDRiD, 24.7% on ISIC) and an anti-aligned CAV throughout. Conversely, the inversion-based methods reach their best concept scores at the FID-cheapest operating point and are penalised by FID as guidance grows. A model selection based on FID alone would therefore prefer the baselines and the minimal edits, and would discard exactly the configurations that, on IDRiD, carry the strongest concept. For concept-based explainability the relevant

question is whether the concept is learnable and aligned, not whether the image is photorealistic; FID is necessary as a sanity check but is insufficient and, taken alone, misleading.

5.5 Concept Alignment and the Limits of a Linear CAV

The clinical AUC and Cohen’s d test whether the concept *direction*, not merely the concept’s learnability by a full probe, matches the real disease axis, and they expose failures that CL alone hides. On IDRiD the contrast is stark: copy-paste reaches the CL ceiling (0.997) but its CAV is anti-aligned ($d = -2.69$), whereas the diffusion concepts at $w \geq 3$ are genuinely aligned ($d = +1.84$ to $+3.68$, clinical AUC 0.96–0.99). Showing that a learned, diffusion-injected concept yields a more clinically aligned direction than a copy-paste of *real* lesion pixels is a non-obvious result in the method’s favour, and it holds only from $w = 3$ upward: at the minimal edit the alignment is still weak (M3 $w = 1.5$: clinical AUC 0.561, $d = +0.43$). The diagnostic also distinguishes “learnable” from “aligned”: at $w = 1.5$ M3 has CL = 0.999 yet $d = +0.43$, i.e. a probe finds the concept before the single CAV direction does.

Two cautions limit how far the alignment axis can be pushed. First, the CAV is a single linear direction and is fragile: on ISIC even the real concept is anti-oriented (clinical AUC 0.425), so the axis cannot rank methods there and the sign-corrected separations are confounded by a synthetic-vs-real direction. Second, the per-method differences in d on a single feature space carry no confidence interval, so only large gaps (the copy-paste anti-alignment, the IDRiD vs. BUSI/ISIC contrast) should be read as real.

5.6 Edit Localisation

Chapter 3 argued that the interplay between the patient-anchored null-text embedding and the concept-conditioned branch should confine the edit to the concept-relevant region, with no explicit spatial constraint imposed. We examine this here through difference maps between each healthy input and its concept-injected counterpart, $|x_h - x^+|$ (Figure 5.3). On IDRiD the change is concentrated in a few focal retinal regions, consistent with a localised edit, and the surrounding vessels and optic disc are left close to unchanged. On BUSI the difference map is nearly empty, but this reflects the failure of the concept to form on this out-of-distribution modality (Section 5.2.3) rather than a clean, localised insertion.

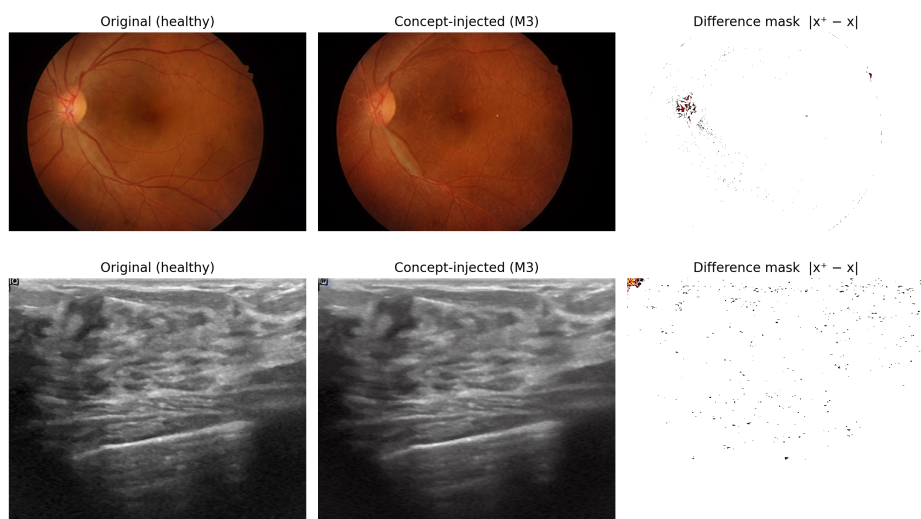


Figure 5.3: Edit localisation via difference maps: healthy input (left), the M3 concept-injected counterpart (centre), and their absolute difference (right). **Top (IDRiD)**: the edit is confined to a few focal regions, consistent with the localisation argument of Section 3.2.5. **Bottom (BUSI)**: the difference is nearly empty because the concept does not form on ultrasound (Section 5.2.3), not because of a clean localised edit.

5.7 Downstream Robustness via LCRReg

The robustness experiment asks whether the synthetic-derived CAV can deflect an artificial colour-tag shortcut. Two findings hold across datasets.

First, the experiment is *statistically underpowered*. On IDRiD only one of five seeds learned the shortcut (the biased baseline generalised in the other four), and on BUSI and ISIC only three of five seeds did, with the other two collapsing to trivial single-class predictors; the unbiased test sets are small (33, 152 and 428 images). Across five seeds, no ΔLCRReg reaches significance at $\alpha = 0.05$ on a one-sample t -test against zero; the closest to significance, baseline 2 on BUSI ($+6.5 \pm 5.7$, $t_4 \approx 2.5$, $p \approx 0.06$), is only borderline. Even the *real* concept CAV produces a near-zero delta on every dataset. The artificial bias barely perturbs any CAV, a normalised bias–CAV alignment score below 0.02 for every method, so there is little shortcut signal along the concept direction for the regulariser to remove.

Second, where any trend exists it runs *against* clinical validity. On both BUSI and ISIC the least clinically aligned CAVs deflect the shortcut best, while the real concept barely helps: $\rho(\text{clinical AUC}, \Delta\text{LCRReg}) = -0.34$ on BUSI and -0.51 on ISIC (neither significant, $p \approx 0.06$ for ISIC). This is the geometric independence anticipated in Chapter 4: a direction can interfere with a shortcut without being the disease direction. We therefore report ΔLCRReg but do not draw method-quality conclusions from it; a larger unbiased test set and a bias that is not orthogonal to the concept would be required to make this axis informative.

5.8 Cross-Dataset Synthesis

Table 5.4 summarises the three datasets along the axes that separate them. The pattern suggests that the method’s success depends on two conditions that the three datasets happen to dissociate: the imaging modality must be close enough to the generative backbone for the inversion to be faithful, and the target concept must be salient and separable enough to be rendered by the injection and captured as a direction. IDRiD satisfies both and the method works; BUSI fails the first (the inversion degrades and FID explodes); ISIC fails the second (the concept is weak, and only a minimal edit preserves it). We present this two-prerequisite reading as an *organising hypothesis* rather than a demonstrated result: with a single dataset per condition, modality and concept type are fully confounded and the independence of the two conditions cannot be established from these experiments.

Table 5.4: Cross-dataset summary. “Concept” columns use the Concept Learnability upper bound and the real-concept clinical AUC. The two-prerequisite reading is an organising hypothesis ($n = 3$, confounded).

	IDRiD (fundus)	ISIC (dermoscopy)	BUSI (ultrasound)
Domain gap (baseline FID)	small (~ 69)	small (~ 63 – 92)	large (127 – 152)
Inversion PSNR / SSIM	35.8 / 0.89	32.9 / 0.85	31.7 / 0.85
Concept upper bound (AUC)	1.000	0.846	0.986
Real-concept clinical AUC	0.999	0.425 (weak)	0.977
Guidance effect on CL	helps (to $w = 3$)	monotone decline	erratic (M1/M2)
Best operating point	M3, $w = 3$	M3, $w = 1.5$	none usable
Apparent limiting factor	works	weak concept	domain gap

5.9 Discussion

Three conclusions are supported by the results. **(i) Concept validity and perceptual fidelity are distinct, and the evaluation must measure both.** The inpainting baseline is realistic yet carries no concept; copy-paste is maximally learnable yet clinically anti-aligned; the diffusion concepts trade FID for concept validity. A single-axis evaluation would mis-rank every one of these cases, which is the justification for the four-axis framework. **(ii) The masked-loss token is the strongest concept-injection variant where the method works.** On IDRiD it reaches a recognisable, clinically aligned concept at far lower guidance than the patch- or prompt-based variants, and on ISIC it is the best concept method at the minimal edit; its advantage is real but is confined to the favourable regimes and does not survive the domain gap on BUSI. **(iii) The minimal-counterfactual operating point is the safe default, and is essential where the concept is subtle.** The ISIC sweep shows that strong guidance can destroy a faint concept; the IDRiD sweep shows that a faint concept can also be brought out by guidance when it is bold. The operating point should therefore be selected per concept, with minimal guidance as the conservative choice.

For concept-based explainability the practical implication is that inversion-based concept injection on a *frozen* backbone can supply concept-positive images whose features transfer to real data: on IDRiD the masked-loss concept matches the real-image ceiling on Concept Learnability. As argued in Section 5.8, this holds only where both prerequisites are met; where they are not, like ultrasound, or the weak pigment-network direction, the synthetic concept either fails to form or only matches a pixel-copy baseline. Inversion fidelity and the real-concept clinical AUC are therefore candidate diagnostics to check before synthesis.

Chapter 6

Conclusion and Future Work

This thesis asked whether a frozen, general-purpose diffusion model can be turned into a source of concept-controlled synthetic data for medical explainability, without the dense annotations and the entangled correlations that make real data difficult to use for this purpose. The pipeline proposed in Chapter 3 combines a Textual Inversion concept token with Null-Text Inversion, and injects the token into the inversion of a healthy image through classifier-free guidance. It uses no fine-tuning of the backbone and no manual mask at inference. To judge the output, Chapter 4 introduced an evaluation framework with four axes: perceptual fidelity (FID), semantic validity (Target Class Accuracy), concept learnability against a real-data upper bound, and a downstream-robustness probe built on LCRReg. The pipeline and two baselines were evaluated on three modalities chosen to sit at different distances from the backbone’s pre-training distribution.

6.1 Summary and main findings

The evaluation framework earned its place. A single-axis judgement would have ranked the methods incorrectly: the inpainting baseline was realistic yet carried almost no recognisable concept, copy-paste was maximally learnable yet its concept direction was anti-aligned with real disease, and the diffusion-based concepts traded perceptual fidelity for semantic validity. Measuring fidelity, validity, transfer and robustness separately was what made these distinctions visible.

Among the three concept-injection strategies, the masked-loss token (M3) was the strongest wherever the method worked. On retinal photography it reached a recognisable, clinically aligned concept at far lower guidance than the patch-based or prompt-based variants, and on dermoscopy it was the best of the three at the minimal-edit operating point. This supports the central design choice of training the token on whole images while restricting the gradient to the lesion region, although the advantage did not survive the domain gap on ultrasound.

The clearest empirical result is that the method works within a delimited operating envelope rather than universally. On IDRiD, where the hard exudates are a salient concept, concept injection produced positive images whose learnability matched the real-data upper bound. On ISIC the limiting factor was the concept itself: pigment network is a weak linear direction, and a minimal

edit only matched a copy-paste of real pixels rather than adding concept content beyond it. On BUSI the inversion degraded and image quality collapsed under strong guidance, so no usable operating point was found. Across the three datasets, success co-varied with two conditions: faithful inversion and a salient, separable concept. As section 5.8 notes, this is an organising hypothesis rather than a tested interaction. In both failure cases the binding constraint is the same frozen, general-purpose backbone.

Two practical points follow. First, the minimal-edit operating point is the safe default, and is essential where the concept is subtle: stronger guidance helped a bold concept on IDRiD but destroyed a faint one on ISIC. Second, the downstream-robustness experiment was statistically underpowered. No $\Delta\text{LCRRReg}$ reached significance, even the real-concept CAV produced a near-zero effect, and the artificial bias was nearly orthogonal to the concept direction, so the axis is reported as an inconclusive result rather than as evidence for or against the method.

Taken together, the contribution is threefold: a mask-free concept-injection pipeline on a frozen backbone, a controlled comparison of three token-training strategies, and a multi-axis evaluation framework that exposes the gap between looking realistic and carrying a valid concept. The empirical study is best read as a map of where this kind of synthetic data is trustworthy, not as a claim that it is trustworthy everywhere.

6.2 Answering the research questions

The three research questions posed in the introduction can now be answered directly; the supporting evidence is detailed in Chapter 5 and only the verdicts are restated here.

RQ1: inserting an isolated concept while preserving anatomy. Conditionally yes. Within the favourable regime the pipeline inserts the target pathology while leaving the surrounding anatomy largely intact, and the difference maps of Section 5.6 show that the edit is spatially confined even though no mask is imposed at inference. Outside that regime the answer is negative: on dermoscopy a minimal edit only matches a copy-paste of real pixels, and on ultrasound the concept does not form. The capability is therefore real but bounded, not universal.

RQ2: how the concept token should be learned. Restricting the Textual Inversion loss to the lesion region (M3) is the best of the three strategies wherever the method works, confirming the masked-loss design over patch training (M2) and a hand-written prompt (M1). As reported above and in Chapter 5, M3 reaches a recognisable, clinically aligned concept at far lower guidance than M2 or M1 on IDRiD, and is the strongest concept method at the minimal-edit operating point on ISIC.

RQ3: validity and usefulness of the injected concept. Answered affirmatively on recognition and transfer, and left open on downstream robustness. The injected concept is recognised as the

target pathology and transfers to real data: on IDRiD its Concept Learnability matches the real-data upper bound, the strongest single result of the study. The remaining sub-question, whether the concept can support a downstream model against spurious correlations, is not established: the LCRReg probe proved statistically underpowered, no Δ LCRReg reached significance, and the injected bias turned out nearly orthogonal to the concept direction. RQ3 is therefore a qualified “yes” on validity, and the inconclusive robustness result is taken up among the limitations that follow.

6.3 Limitations

The three datasets confound imaging modality with concept type, so the two-prerequisite reading cannot be established as an interaction from these experiments. The robustness experiment is underpowered, with small unbiased test sets and several seeds in which the baseline did not learn the shortcut. The concept-alignment diagnostics rest on a single linear direction in one feature space and degenerate to an orientation artefact on ISIC. Edit localisation is shown only qualitatively, through difference maps, without a quantitative score. Finally, the reduction in annotation effort is argued indirectly, through concept learnability against a real-data ceiling, rather than by training a downstream concept-based model on the synthetic data.

A more fundamental limitation is intrinsic to the lightweight design itself. Freezing a general-purpose Stable Diffusion backbone and learning only a concept token is what makes the method cheap enough for the small, expert-annotated datasets it targets, but it also makes the backbone’s pre-training prior the binding constraint: classifier-free guidance can only amplify representations the backbone already has. Where those are usable and uncontested the edit lands (IDRiD); where the concept is too subtle to form a separable direction the guidance instead recruits a generic-lesion prior (ISIC). The ultrasound case is the most telling, and is worth noting because it cuts against the simple reading that failure tracks raw distance from the pre-training data: the backbone does not lack breast priors, it arguably has more semantic context for a breast mass than for a retinal exudate, but that context is photographic, so under strong guidance the output drifts off the ultrasound manifold towards a literal breast rather than a B-mode tumour, and fidelity collapses (Figure 5.2). This hints that the failure is the prior pointing the wrong way rather than the prior being absent, an observation rather than a controlled result. Either way, a frozen general-purpose backbone trades capability for cheapness, and these experiments delimit where that trade pays off.

6.4 Future work

Several directions follow directly from these limitations.

Make the robustness axis informative. The LCRReg experiment needs more statistical power: larger unbiased test sets, more training seeds, and, above all, an artificial bias that is not nearly

orthogonal to the concept direction. Only then can $\Delta\text{LCRR}_{\text{reg}}$ measure whether the synthetic concept deflects a shortcut rather than merely some unrelated direction.

Quantify localisation. The difference maps of Section 5.6 should be replaced by a measured score, such as the fraction of the edit energy that falls inside the lesion mask, or the background-preservation SSIM outside it. This would turn the localisation claim that motivates the whole design into something that can be tested rather than inspected.

Compare against external generators. A direct comparison with PRISM and other counterfactual generators, under the same four-axis protocol, would place the method in the literature more firmly than the internal baselines alone.

Start from a medical backbone. The most promising route past this limitation is not to fine-tune the backbone to each target dataset, the regime this work addresses rarely has enough images for that, but to replace the frozen *general-purpose* base with one carrying broad *medical* priors. A diffusion model pre-trained once on heterogeneous medical imagery, used frozen and shared across downstream tasks, would already represent modalities such as ultrasound, so the same lightweight token injection could express a concept without the guidance drifting towards a natural-image prior keeping the mask-free, no-per-dataset-fine-tuning property that motivates the method.

The broader takeaway is a measured one. A frozen diffusion backbone can act as a concept-injection engine for medical explainability, but only inside a specific operating envelope set by its own pre-training prior, a modality it has seen and a concept compatible with that prior, and the usefulness of the resulting data has to be judged along several axes rather than by image realism alone.

References

- [1] Julius Adebayo et al. “Sanity checks for saliency maps”. *Advances in neural information processing systems* 31 (2018).
- [2] Marcus A Badgeley et al. “Deep learning predicts hip fracture using confounding patient and healthcare variables”. *NPJ digital medicine* 2.1 (2019), p. 31.
- [3] Alina Jade Barnett et al. “A case-based interpretable deep learning model for classification of mass lesions in digital mammography”. *Nature Machine Intelligence* 3.12 (2021), pp. 1061–1070.
- [4] Daniel C Castro, Ian Walker, and Ben Glocker. “Causality matters in medical imaging”. *Nature Communications* 11.1 (2020), p. 3673.
- [5] Pierre Chambon et al. “Roentgen: vision-language foundation model for chest x-ray generation”. *arXiv preprint arXiv:2211.12737* (2022).
- [6] Agisilaos Chatsias et al. “Adversarial image synthesis for unpaired multi-modal cardiac data”. In: *International workshop on simulation and synthesis in medical imaging*. Springer. 2017, pp. 3–13.
- [7] Agisilaos Chatsias et al. “Disentangled representation learning in cardiac image analysis”. *Medical image analysis* 58 (2019), p. 101535.
- [8] Noel Codella et al. “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)”. *arXiv preprint arXiv:1902.03368* (2019).
- [9] Joseph Paul Cohen, Margaux Luck, and Sina Honari. “Distribution matching losses can hallucinate features in medical image translation”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2018, pp. 529–536.
- [10] Joseph Paul Cohen et al. “Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays”. In: *Medical Imaging with Deep Learning*. PMLR. 2021, pp. 74–104.
- [11] Valentina Corbetta et al. “In-hoc Concept Representations to Regularise Deep Learning in Medical Imaging”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, pp. 7312–7321.
- [12] Walid Al-Dhabyani et al. “Dataset of breast ultrasound images”. *Data in Brief* 28 (2020), p. 104863.
- [13] Yingying Fang et al. “Diffexplainer: Unveiling black box models via counterfactual generation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 208–218.
- [14] Rinon Gal et al. “An image is worth one word: Personalizing text-to-image generation using textual inversion”. *arXiv preprint arXiv:2208.01618* (2022).

- [15] Robert Geirhos et al. “Shortcut learning in deep neural networks”. *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.
- [16] Amirata Ghorbani et al. “Towards automatic concept-based explanations”. *Advances in neural information processing systems* 32 (2019).
- [17] Yash Goyal et al. “Counterfactual visual explanations”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2376–2384.
- [18] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [19] Amir Hertz et al. “Prompt-to-prompt image editing with cross attention control”. *arXiv preprint arXiv:2208.01626* (2022).
- [20] Martin Heusel et al. “Gans trained by a two time-scale update rule converge to a local nash equilibrium”. *Advances in neural information processing systems* 30 (2017).
- [21] Irina Higgins et al. “beta-vae: Learning basic visual concepts with a constrained variational framework”. In: *International conference on learning representations*. 2017.
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.
- [23] Jonathan Ho and Tim Salimans. “Classifier-free diffusion guidance”. *arXiv preprint arXiv:2207.12598* (2022).
- [24] Edward J Hu et al. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *International Conference on Learning Representations*. 2022.
- [25] Monish Keswani et al. “Proto2Proto: Can you recognize the car, the way I do?” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10233–10243.
- [26] Been Kim et al. “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)”. In: *International conference on machine learning*. PMLR. 2018, pp. 2668–2677.
- [27] Hyunjik Kim and Andriy Mnih. “Disentangling by factorising”. In: *International conference on machine learning*. PMLR. 2018, pp. 2649–2658.
- [28] Pieter-Jan Kindermans et al. “The (un) reliability of saliency methods”. In: *Explainable AI: Interpreting, explaining and visualising deep learning*. Springer, 2019, pp. 267–280.
- [29] Pang Wei Koh et al. “Concept bottleneck models”. In: *International conference on machine learning*. PMLR. 2020, pp. 5338–5348.
- [30] Amar Kumar et al. *PRISM: High-Resolution Precise Counterfactual Medical Image Generation using Language-guided Stable Diffusion*. 2025. arXiv: 2503.00196 [cs.CV]. URL: <https://arxiv.org/abs/2503.00196>.
- [31] Oran Lang et al. “Explaining in style: training a GAN to explain a classifier in stylespace”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 693–702.
- [32] Oscar Li et al. “Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.
- [33] Francesco Locatello et al. “Challenging common assumptions in the unsupervised learning of disentangled representations”. In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.

- [34] Andreas Lugmayr et al. “Repaint: Inpainting using denoising diffusion probabilistic models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11461–11471.
- [35] Chenlin Meng et al. “SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations”. In: *International Conference on Learning Representations*. 2022.
- [36] Ron Mokady et al. “Null-text inversion for editing real images using guided diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 6038–6047.
- [37] Muzaffer Özbey et al. “Unsupervised medical image translation with adversarial diffusion models”. *IEEE Transactions on Medical Imaging* 42.12 (2023), pp. 3524–3539.
- [38] Cristiano Patrício, Luis F Teixeira, and João C Neves. “Towards concept-based interpretability of skin lesion diagnosis using vision-language models”. In: *2024 IEEE international symposium on biomedical imaging (ISBI)*. IEEE. 2024, pp. 1–5.
- [39] Cristiano Patrício, Luís F Teixeira, and João C Neves. “ViConEx-Med: Visual Concept Explainability via Multi-Concept Token Transformer for Medical Image Analysis”. *arXiv preprint arXiv:2510.10174* (2025).
- [40] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. “Deep structural causal models for tractable counterfactual inference”. *Advances in neural information processing systems* 33 (2020), pp. 857–869.
- [41] Walter HL Pinaya et al. “Brain imaging generation with latent diffusion models”. In: *MIC-CAI workshop on deep generative models*. Springer. 2022, pp. 117–126.
- [42] Prasanna Porwal et al. “Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research”. *Data* 3.3 (2018), p. 25.
- [43] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.
- [44] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. “Right for the right reasons: Training differentiable models by constraining their explanations”. *arXiv preprint arXiv:1703.03717* (2017).
- [45] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. *Nature machine intelligence* 1.5 (2019), pp. 206–215.
- [46] Nataniel Ruiz et al. “Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 22500–22510.
- [47] Pedro Sanchez and Sotirios A Tsaftaris. “Diffusion causal models for counterfactual estimation”. *arXiv preprint arXiv:2202.10166* (2022).
- [48] Yiqing Shen et al. “ProtoSAM-2D: 2D semantic Segment Anything Model with mask-level prototype-learning and distillation”. In: *Medical Imaging 2025: Image Processing*. Vol. 13406. SPIE. 2025, pp. 18–22.
- [49] Jiaming Song, Chenlin Meng, and Stefano Ermon. “Denoising diffusion implicit models”. *arXiv preprint arXiv:2010.02502* (2020).
- [50] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions”. *Scientific Data* 5.1 (2018), pp. 1–9.

- [51] Xin Yi, Ekta Walia, and Paul Babyn. “Generative adversarial network in medical imaging: A review”. *Medical image analysis* 58 (2019), p. 101552.
- [52] Sheng Zhang et al. “Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs”. *arXiv preprint arXiv:2303.00915* (2023).

Appendix A

Reproducibility Supplement

Most reproducibility details (seeds, hyperparameters, the train/validation split and the bias settings) are given in Chapter 4 (Tables 4.2 and 4.3). This appendix records only what is not already in the main text: the compute environment, the exact run commands, and the datasets’ access and licensing.

A.1 Source Code

The implementation is split into two public repositories, mirroring the division between Chapter 3 (the generation method) and Chapter 4 (the evaluation framework):

- **Generation / method:** <https://github.com/guilhermevaz8/experiments>
- **Evaluation framework:** https://github.com/guilhermevaz8/eval_pipeline

The generation repository takes a healthy image, performs Null-Text Inversion (NTI) of it, and injects the disease concept through Classifier-Free Guidance, emitting a directory of synthetic images per method (prompt-only, patch-trained token, masked-loss token) alongside the two baselines. The evaluation repository consumes exactly those synthetic-image directories: it trains a supporting ResNet-50 and then scores each method on the four axes (FID, TCA, Concept Learnability and Robustness/ Δ LCRRReg) plus the CAV diagnostics. In short, one repository *generates* the images and the other *scores* them; the handoff is the set of generated image folders.

Table A.1 maps each thesis component to its source file. Paths in the generation repository are relative to its root (`exp1/`, `baseline1/`, `baseline2/`); paths in the evaluation repository are relative to the `eval_pipeline/` package.

Thesis component	Repository	Source file
NTI inversion	generation	exp1/run_sanity_check.py
M1 prompt injection	generation	exp1/run_prompt_injection.py
M2 patch token	generation	exp1/train_concept.py
M3 masked-loss token	generation	exp1/train_concept_masked.py
Lesion-crop extraction (M2)	generation	exp1/extract_lesion_crops.py
Trained-token injection (M2/M3)	generation	exp1/run_concept_injection.py
Copy-paste baseline	generation	baseline1/baseline1_patcher.py
Inpainting baseline	generation	baseline2/baseline2_diffusion_inpainting.py
FID	evaluation	eval_pipeline/metrics/fid.py
TCA	evaluation	eval_pipeline/metrics/tca.py
Concept Learnability	evaluation	eval_pipeline/metrics/concept_learnability.py
Robustness / LCRReg	evaluation	eval_pipeline/metrics/robustness.py
CAV diagnostics	evaluation	eval_pipeline/utils/cav.py

Table A.1: Mapping of thesis components to source files across the two repositories. The CFG injection shared by M1–M3 and the NTI / training math live in `exp1/utils.py` of the generation repository.

A.2 Compute Environment

Generation uses a frozen Stable Diffusion 1.5 checkpoint (`runwayml/stable-diffusion-v1-5`); the inversion and null-text optimisation run in single precision (`fp32`). The evaluation pipeline uses a ResNet-50 backbone (ImageNet-initialised).

Both repositories manage their dependencies with Poetry. The generation environment uses Python 3.9–3.12 with `torch 2.2.0+cu118`, `diffusers 0.27`, `transformers 4.36` and `peft 0.7.1`; the evaluation environment uses Python ≥ 3.10 with `torch 2.2+cu121`, `torchmetrics` and `scikit-learn`.

A.3 Run Commands

The two pipelines are run in sequence: generation produces the synthetic-image directories, and evaluation scores them. Both are Poetry-managed Python packages. In the listings below `<ds>` stands for one of `idrid`, `isic2018` or `busi`; absolute cluster paths have been abbreviated to `<staging>`, `<gen>` and `<ckpt>`.

Generation (method repository)

The commands below are the per-script equivalents of `scripts/run_<ds>_pipeline.sh` orchestrator, in dependency order. All `exp1/` scripts assume the working directory is `exp1/` (they use sibling-module imports such as `import utils`). The IDRiD prompts and tokens are shown; substitute the per-dataset values for ISIC2018 / BUSI.

```
# Environment (once)
```

```
poetry install --no-root

# Download + stage a dataset (IDRiD example)
bash scripts/prepare_idrid_staging.sh \
    "$RAW_IDRID" "$STAGING_IDRID" hard_exudates

# Inversion: NTI of the healthy images + reconstruction check
cd exp1
python run_sanity_check.py \
    --dataset-dir <staging>/healthy \
    --output-dir <gen>/inversions \
    --limit 200 --dtype fp32 \
    --prompt "a retinal fundus photograph"

# M1: prompt-only injection sweep (no learned token)
python run_prompt_injection.py \
    --inversion-dir <gen>/inversions \
    --output-dir <gen>/inject_prompt \
    --disease-prompt "fundus photograph with hard exudates \
and microaneurysms, diabetic retinopathy" \
    --guidance-scales 1.5,3,5,7.5,10 --limit 0

# M2 step 1: extract lesion crops
python extract_lesion_crops.py \
    --diseased-dir <staging>/diseased \
    --masks-dir <staging>/masks \
    --output-dir <gen>/lesion_crops

# M2 step 2: train the patch token (standard loss)
python train_concept.py \
    --crops-dir <gen>/lesion_crops/crops \
    --output <gen>/concept_token \
    --token "<IDRiD_lesion>" --steps 3000 \
    --prompt-template "a fundus photograph with {tok}"

# M3: train the masked-loss token on full images
python train_concept_masked.py \
    --images-dir <staging>/diseased \
    --masks-dir <staging>/masks \
    --output <gen>/concept_token_masked \
    --token "<IDRiD_lesion_masked>" --steps 3000 \
    --prompt-template "a fundus photograph with {tok}"

# Inject a trained concept (M2 shown; swap paths for M3)
python run_concept_injection.py \
    --inversion-dir <gen>/inversions \
    --output-dir <gen>/inject_concept \
    --concept <gen>/concept_token/concept_IDRiD_lesion.pt \
    --prompt-template "a fundus photograph with {tok}" \
```

```

--guidance-scales 1.5,3,5,7.5,10 --limit 0

# Baseline 1: copy-paste of real lesion pixels
cd ..
python baseline1/baseline1_patcher.py \
  --healthy <staging>/healthy \
  --diseased <staging>/diseased \
  --masks <staging>/masks \
  --output <gen>/baseline1 --seed 42

# Baseline 2: SD inpainting (--diseased takes the MASK dir;
# pass --model with an SD-1.5 *inpainting* checkpoint)
python baseline2/baseline2_diffusion_inpainting.py \
  --healthy <staging>/healthy \
  --diseased <staging>/masks \
  --output <gen>/baseline2 \
  --prompt "hard exudates, yellow deposits, \
diabetic retinopathy, fundus lesions" \
  --steps 50 --guidance 7.5

```

Evaluation (evaluation repository)

All four axes are driven through one CLI (`python -m eval_pipeline.main`) with a config file per dataset (`config_<ds>.yaml`). The supporting ResNet-50 must be trained first: every axis except FID depends on `best_model.pth`. The LLM judge is off by default and is skipped cleanly when no real API key is set.

```

# Train the supporting ResNet-50 (once per dataset)
python -m eval_pipeline.main \
  --config config_<ds>.yaml --train

# FID + TCA (LLM judge auto-skipped)
python -m eval_pipeline.main \
  --config config_<ds>.yaml --evaluate \
  --output-json results/<ds>_eval.json

# Concept Learnability (CL AUC vs real upper bound)
python -m eval_pipeline.main \
  --config config_<ds>.yaml --concept-learnability \
  --output-json results/<ds>_cl.json

# Robustness (delta-LCRReg + CAV diagnostics, multi-seed)
python -m eval_pipeline.main \
  --config config_<ds>.yaml --robustness \
  --output-json results/<ds>_rob.json

# Comparison report: master table + figures over all runs
python -m eval_pipeline.scripts.aggregate_results \
  --results-dir results \

```

```
--out-dir      results/analysis \
--dataset <ds>
```

A.4 Datasets

Table A.2: Datasets used in this work, with the target concept and source.

Dataset	Modality	Target concept	Source
IDRiD	Fundus photography	Hard exudates	[42]
ISIC 2018	Dermoscopy	Pigment network	[8, 50]
BUSI	Breast ultrasound	Tumour	[12]

All three datasets are publicly available for research. IDRiD is distributed through IEEE Data-Port [42]; the ISIC 2018 Task 2 data is available from the ISIC Archive [8]; and BUSI is available from its original authors [12].

A.5 Qualitative Gallery

For one representative image per dataset, the panels below show the healthy input, its null-text inversion reconstruction, the two baselines, and the three injection variants at the dataset’s representative guidance scale.

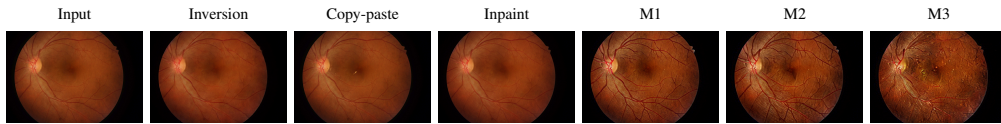


Figure A.1: IDRiD (example `IDRiD_143`, injection guidance $w = 3$): the healthy input, its null-text inversion reconstruction, the copy-paste and inpainting baselines, and the three injection variants (M1 prompt, M2 patches, M3 masked loss).

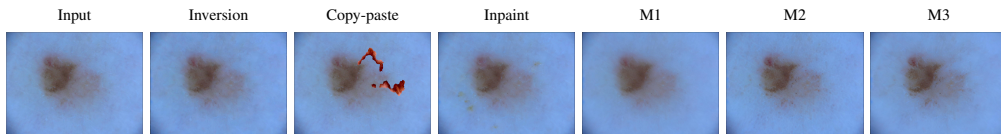


Figure A.2: ISIC 2018 (example `ISIC_0000040`, injection guidance $w = 1.5$): panels as in Figure A.1. The minimal edit ($w = 1.5$) is the best operating point for this concept.

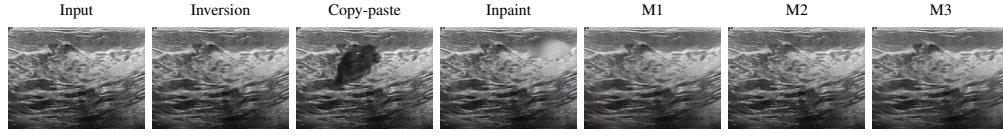


Figure A.3: BUSI (example `normal_6`, injection guidance $w = 1.5$): panels as in Figure A.1. On this modality the concept does not form and stronger guidance drives the output out of distribution.

A.6 Qualitative Gallery: Guidance Sweep

The figures below show the masked-loss token (M3) for one representative image per dataset across the guidance sweep, next to the healthy input. They make the guidance–fidelity trade-off of Section 5.3 visible: as w grows the injected concept becomes more pronounced but the image becomes progressively less realistic. The same drop-in convention applies.

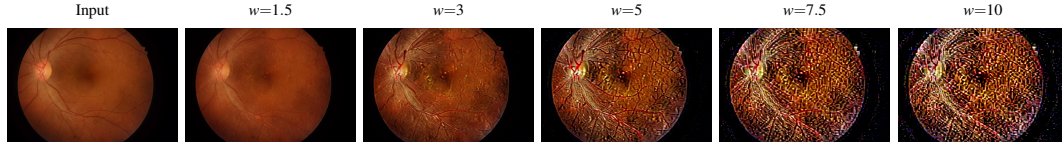


Figure A.4: IDRiD, the masked-loss token (M3) across the guidance sweep. The hard exudates strengthen as guidance grows (recognisable from $w = 3$), while perceptual fidelity degrades (FID $84.7 \rightarrow 256.9$ over the sweep).

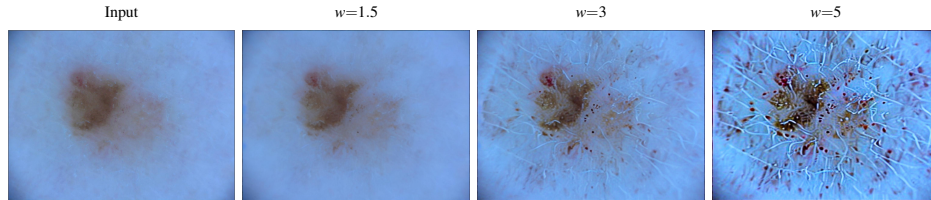


Figure A.5: ISIC 2018, M3 across the guidance sweep. The faint pigment network is best preserved at the minimal edit ($w = 1.5$); larger guidance overwrites it with generic lesion content and the concept collapses.

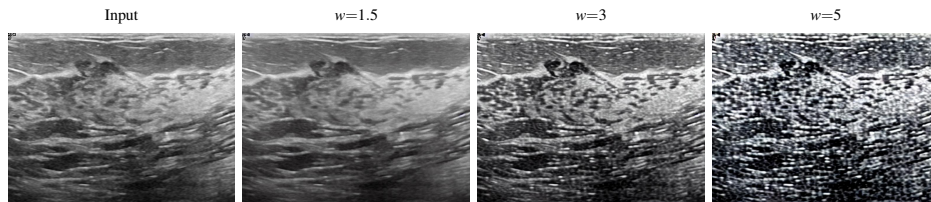


Figure A.6: BUSI, M3 across the guidance sweep. On this out-of-distribution modality the concept never forms; by $w = 5$ the output has collapsed into out-of-distribution texture (FID ~ 373).