

Medical AlphaEdit: A Train-Free Framework for Error Correction and Hallucination Suppression in Medical Vision Transformers

December 27, 2025

Abstract

Medical vision transformers (ViTs) have shown remarkable success in diagnostic tasks, yet they remain prone to errors and hallucinations that can critically impact clinical decisions. We propose Medical AlphaEdit, a train-free framework designed to correct errors and suppress hallucinations in medical ViTs without requiring additional training data or fine-tuning. The framework operates in two phases: fault localization identifies critical image patches and transformer layers responsible for errors by combining causal tracing with attention entropy analysis, while null-space projected editing constrains parameter updates to the null space of correct medical knowledge activations, ensuring edits do not disrupt previously learned representations. Moreover, the framework dynamically selects the most influential layers for editing and supports future extensions for concept-level repairs. Experimental validation demonstrates that Medical AlphaEdit effectively corrects errors while preserving model performance on correctly classified cases, offering a practical solution for improving the reliability of medical ViTs in real-world clinical settings. The proposed method addresses a critical gap in deploying trustworthy AI systems for healthcare, where error correction must be both precise and computationally efficient.

1 Introduction

Medical Vision Transformers (Med-ViTs) and Medical Vision-Language Models (Med-VLMs) have revolutionized diagnostic workflows by enabling multi-modal analysis of medical images and reports [?]. These models, built upon Transformer architectures [?], excel in tasks ranging from radiology report generation to lesion localization. However, their deployment in clinical settings is hindered by persistent challenges: misclassifications due to subtle

anatomical variations and hallucinations that generate non-factual descriptions [?]. Such errors are particularly concerning in medicine, where diagnostic accuracy directly impacts patient outcomes.

Existing mitigation strategies primarily rely on fine-tuning with additional labeled data [?] or post-processing outputs [?]. While effective in some cases, these approaches often require extensive computational resources and may inadvertently degrade model performance on correctly classified cases—a phenomenon known as catastrophic forgetting. Even parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA) [?], despite reducing trainable parameters, still require gradient-based optimization over multiple iterations, demand curated training data, and provide no theoretical guarantees against knowledge degradation. Recent studies highlight the limitations of conventional methods, especially when dealing with rare conditions or underrepresented demographics [?].

The key insight driving our work is that medical knowledge in Transformer-based models is encoded in specific layers and parameters, as evidenced by recent analyses of attention patterns in Med-VLMs [?]. This spatial localization of knowledge suggests that targeted interventions could correct errors without global model retraining. Drawing inspiration from linear algebra principles, we hypothesize that null-space projections [?] can isolate and modify erroneous activations while preserving correct medical representations.

We present **Medical AlphaEdit**, a train-free framework that addresses these challenges through three key innovations. First, it employs causal tracing [?] combined with attention entropy analysis to pinpoint the layers and image patches responsible for errors. Second, it formulates parameter edits as constrained optimization problems with *closed-form solutions*, ensuring updates lie in the null space of correct knowledge embeddings—requiring no gradient computation or iterative optimization. Third, it introduces adaptive layer selection and concept-level editing via Riemannian geometry, enabling both instance-specific corrections and systematic bias mitigation.

Unlike gradient-based methods (fine-tuning, LoRA), Medical AlphaEdit offers several fundamental advantages:

- **Closed-form solution:** Weight updates are computed analytically in a single forward pass, eliminating the need for backpropagation.
- **Theoretical guarantees:** The null-space constraint mathematically ensures zero degradation on correctly classified samples.
- **Data efficiency:** Corrections require only the erroneous sample and a small set of correctly classified exemplars—no additional training data.
- **Computational efficiency:** Editing takes <1 second per case versus hours for fine-tuning, enabling real-time clinical deployment.

The contributions of this work are fourfold:

1. A train-free error correction framework that leverages the intrinsic structure of Transformer models to localize and repair errors without additional training data or gradient computation.
2. A prototype-guided target optimization method that derives correction targets from class-conditional feature distributions, addressing the critical challenge of specifying target activations in continuous visual feature spaces.
3. An adaptive layer selection mechanism and concept-level editing via Riemannian mean that enables both precise instance corrections and systematic bias mitigation (e.g., demographic disparities in diagnosis).
4. Comprehensive evaluation on ChestX-ray14 and CheXpert benchmarks against strong baselines including LoRA, ROME, and MEMIT, demonstrating statistically significant improvements in correction efficacy while maintaining knowledge preservation.

The remainder of this paper is organized as follows: Section 2 reviews related work in medical model editing and hallucination mitigation. Section 3 introduces necessary background on Transformer architectures and null-space projections. Section 4 details the Medical AlphaEdit framework, followed by experimental validation in Section 5. We discuss implications and future directions in Section 6 before concluding in Section 7.

2 Related Work

Recent advances in medical vision transformers and vision-language models have brought significant attention to the challenges of error correction and hallucination suppression. Existing approaches can be broadly categorized into three directions: model editing techniques, hallucination mitigation strategies, and medical-specific error analysis frameworks.

2.1 Model Editing Techniques

The field of model editing has evolved from traditional fine-tuning approaches to more sophisticated parameter modification methods. Early work in [?] demonstrated that neural networks could be selectively edited by identifying and modifying specific knowledge representations. Subsequent approaches like [?] introduced the concept of null-space projection for model editing, though their application was limited to language models. In the medical domain, 12 proposed a sequential editing approach for healthcare knowledge updates, but required extensive training data. Our work differs by introducing a train-free editing framework specifically designed for medical vision transformers.

2.2 Hallucination Mitigation

Hallucination suppression has become increasingly critical as medical vision-language models are deployed in clinical settings. The ASTRA framework [?] demonstrated promising results by steering models away from adversarial feature directions, though it focused primarily on security applications rather than medical errors. Recent work in [?] introduced evidence fusion for hallucination correction, but required external knowledge bases. The MedHEval benchmark [?] systematically evaluated various mitigation strategies, highlighting the need for approaches that preserve model performance on correct predictions.

2.3 Medical Error Analysis

Understanding and correcting errors in medical AI systems requires specialized approaches due to the high-stakes nature of clinical decisions. The framework proposed in [?] analyzed how hallucinations propagate through medical language models, while [?] developed evaluation metrics specific to medical vision-language tasks. These studies revealed that medical errors often stem from specific attention patterns and feature representations, motivating our causal tracing approach for fault localization.

The proposed Medical AlphaEdit framework advances beyond existing methods by combining precise fault localization with null-space constrained editing in a unified, train-free framework. Unlike approaches that require additional training data or external knowledge bases, our method operates directly on the model’s existing parameters while preserving correct medical knowledge. The adaptive layer selection mechanism further distinguishes our work by dynamically identifying the most impactful editing locations based on causal analysis.

3 Background and Preliminaries

To establish the foundation for our proposed framework, we first review key concepts in medical vision transformers and the mathematical tools that enable our approach. This section provides the necessary technical background while highlighting the specific challenges in medical applications that motivate our work.

3.1 Medical Vision Transformer Architecture

Medical Vision Transformers (Med-ViT) process input images through a sequence of operations that transform local image patches into diagnostic predictions. Given an input medical image $x \in \mathbb{R}^{H \times W \times C}$, the model first divides it into N non-overlapping patches $\{p_1, p_2, \dots, p_N\}$, each of size $P \times P$. These patches are linearly projected into a D -dimensional embedding space:

$$E = [e_1, e_2, \dots, e_N] = \text{PatchEmbedding}(x) \quad (1)$$

where $E \in \mathbb{R}^{N \times D}$ represents the patch embeddings. A learnable [CLS] token e_{cls} is prepended to capture global image information. The embeddings then pass through L transformer layers, each consisting of Multi-head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks:

$$O_{\text{MSA}}^l = \text{MSA}(\text{LayerNorm}(E^{l-1})) + E^{l-1} \quad (2)$$

$$E^l = \text{MLP}(\text{LayerNorm}(O_{\text{MSA}}^l)) + O_{\text{MSA}}^l \quad (3)$$

for layer $l \in \{1, \dots, L\}$. The Multi-head Self-Attention mechanism computes attention weights through query, key, and value projections:

$$\text{MSA}(X) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (4)$$

where each head computes:

$$\text{head}_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \quad (5)$$

with $Q_i = XW_Q^i$, $K_i = XW_K^i$, $V_i = XW_V^i$ being the query, key, and value matrices for head i . The final [CLS] token embedding e_{cls}^L is used for classification through a linear head W_{head} :

$$\hat{y} = \text{softmax}(W_{\text{head}}e_{\text{cls}}^L) \quad (6)$$

3.2 Challenges in Medical Vision Transformers

Medical applications present unique challenges not fully addressed by standard transformer architectures [?]. The high variability in anatomical structures across patients, combined with the subtlety of many pathological findings, makes medical vision transformers particularly susceptible to two types of errors: misclassification of similar-looking conditions and generation of non-factual descriptions (hallucinations) [?]. These errors often stem from the model’s inability to properly attend to clinically relevant features or from incorrect associations learned during training.

Recent analyses of attention patterns in medical transformers [?] reveal that errors frequently correlate with specific attention heads and layers, suggesting that targeted interventions could correct these issues without full model re-training. This observation forms the basis for our fault localization approach, which we detail in Section 4.

3.3 Causal Tracing in Transformers

Causal tracing provides a powerful tool for understanding how information flows through transformer models and identifying the components responsible for specific predictions [?]. The technique involves systematically perturbing different parts of the model’s computation while observing the impact on the final output. For an input sequence x producing output y , we compute the causal effect C of component c (e.g., a patch embedding or layer activation) as:

$$C = \Delta y = y - y' \quad (7)$$

where y' is the output when c is ablated (set to zero or replaced). In medical vision transformers, we can trace how information from specific image patches propagates through attention layers to influence the diagnostic decision.

The causal tracing process reveals that medical knowledge and errors often follow distinct pathways through the network. Correct diagnoses typically activate consistent patterns across attention heads, while errors frequently arise from abnormal activations in specific layers. This property enables our framework to distinguish between correct medical knowledge that should be preserved and erroneous activations that need correction.

3.4 Null-Space in Linear Algebra

The null-space of a matrix A , denoted $\mathcal{N}(A)$, consists of all vectors x that satisfy $Ax = 0$:

$$\mathcal{N}(A) = \{x \mid Ax = 0\} \quad (8)$$

For a set of correct activations $K_0 \in \mathbb{R}^{d \times m}$ (where d is the feature dimension and m is the number of samples), we compute its covariance matrix $\Sigma = K_0 K_0^T$. Through Singular Value Decomposition (SVD):

$$\Sigma = U \Lambda U^T \quad (9)$$

The null-space projection matrix P is constructed from the eigenvectors corresponding to zero (or near-zero) eigenvalues:

$$P = V_0 V_0^T \quad (10)$$

where V_0 contains the basis vectors of $\mathcal{N}(\Sigma)$. This concept plays a central role in our editing framework, as it allows us to modify model parameters in directions that do not affect existing correct knowledge representations. For a given correct medical knowledge embedding v , we can decompose any parameter update ΔW into two components: one that affects v and one that lies in the null-space of v :

$$\Delta W = \Delta W_{\parallel} + \Delta W_{\perp} \quad (11)$$

Figure 1: Medical AlphaEdit Framework and Its Integration with Med-ViT/Med-VLMs

where $\Delta W_{\perp} v = 0$. By constraining edits to the null-space of correct medical knowledge embeddings, we ensure that these updates only affect the erroneous pathways identified through causal tracing while preserving the model’s performance on correctly classified cases. This property is particularly crucial in medical applications where maintaining existing diagnostic capabilities is as important as correcting errors.

4 Medical AlphaEdit: Null-Space Projection for Train-Free Error Correction

Medical AlphaEdit introduces a novel paradigm for correcting errors in medical vision transformers without requiring retraining or additional data. The framework operates through two coordinated phases: precise fault localization and constrained parameter editing. As shown in Figure 1, the system interfaces directly with standard Med-ViT/Med-VLM architectures, modifying only the identified faulty components while preserving correct medical knowledge representations. This section details the technical foundations and implementation of both phases.

4.1 Null-Space Projection Matrix Construction

The null-space projection matrix forms the mathematical foundation for targeted parameter edits while preserving correct medical knowledge. Given a set of correct activations $K_0 = \{k_1, k_2, \dots, k_n\}$ from the model’s intermediate layers, we first compute their covariance matrix C :

$$C = \frac{1}{n} \sum_{i=1}^n (k_i - \mu)(k_i - \mu)^T \quad (4)$$

where μ represents the mean activation vector. The eigendecomposition of C yields:

$$C = V \Lambda V^T \quad (5)$$

Here, V contains the eigenvectors and Λ is a diagonal matrix of eigenvalues. We partition V into $V = [V_1 | V_0]$, where V_1 spans the principal subspace of correct knowledge and V_0 forms an orthonormal basis for the null space. The projection matrix P onto this null space is constructed as:

$$P = V_0 V_0^T \quad (6)$$

This projection matrix ensures that any subsequent parameter updates will not interfere with the model’s existing correct representations. The dimensionality of V_0 is determined by setting an eigenvalue threshold λ_{thresh} , where eigenvectors corresponding to eigenvalues below this threshold are included in V_0 . For medical applications, we find that preserving the top 80% of variance typically maintains diagnostic accuracy while allowing sufficient flexibility for error correction.

4.2 Null-Space Constrained Weight Update

Given the projection matrix P , we formulate the weight update as a constrained optimization problem that corrects erroneous activations while preserving correct medical knowledge. Let W denote the original weight matrix and k^* represent the faulty activation vector identified through causal tracing. The desired corrected activation v^* is obtained through our prototype-guided optimization described in Section 4.3.

The weight update ΔW must satisfy two conditions: it should transform k^* to v^* while lying in the null-space of correct activations. This leads to the following optimization objective:

$$\min_{\Delta W} \|(W + \Delta W)k^* - v^*\|_2^2 \quad \text{subject to} \quad \Delta W K_0 = 0 \quad (12)$$

The constraint $\Delta W K_0 = 0$ ensures that the update does not alter the model’s behavior on correctly classified samples, which is equivalent to requiring $\Delta W P = \Delta W$. The solution to this constrained problem yields the update rule:

$$\Delta W = (v^* - W k^*) (k^{*T} P) (k^* k^{*T} P + \lambda I)^{-1} \quad (13)$$

Here, λ serves as a regularization parameter that controls the magnitude of the update. The term $(k^{*T} P)$ ensures the update direction lies within the null-space, while the matrix inversion guarantees the update precisely corrects the faulty activation without affecting orthogonal directions.

The physical interpretation of each component becomes clear when examining the update’s effect on an arbitrary input x . The update’s action on x can be decomposed as:

$$\Delta W x = (v^* - W k^*) \cdot \frac{k^{*T} P x}{k^{*T} P k^* + \lambda} \quad (14)$$

This shows that the update only significantly affects inputs x that align with the faulty activation k^* in the null-space projection. The denominator term prevents excessive amplification when x closely matches k^* , ensuring numerical stability.

For multi-head attention layers, we apply the update separately to each head’s query, key, and value projection matrices W_Q^h, W_K^h, W_V^h , where h indexes the attention heads. This allows targeted correction of specific attention patterns while preserving others:

$$\Delta W_V^h = (v_h^* - W_V^h k_h^*)(k_h^{*T} P_h)(k_h^* k_h^{*T} P_h + \lambda I)^{-1} \quad (15)$$

The head-specific projection matrix P_h is constructed from activations corresponding to that head’s correct behavior. This granular approach enables precise editing of individual attention mechanisms without disrupting the overall attention dynamics.

For MLP layers, we apply the same principle but operate on the full hidden dimension without head-wise decomposition. The MLP layer updates target the weight matrices W_1 and W_2 in the two-layer feedforward network:

$$\text{MLP}(x) = W_2 \cdot \text{GELU}(W_1 \cdot x) \quad (16)$$

The fault localization module (Section 4.2) determines whether to edit MSA or MLP components based on the Jacobian analysis of each layer’s contribution to the error.

4.3 Fault Localization for Targeted Editing

To precisely identify the source of errors in medical vision transformers, we develop a fault localization mechanism that combines causal tracing with attention entropy analysis. Given an input image I that produces an incorrect prediction \hat{y} , we first decompose I into N non-overlapping patches $\{p_1, \dots, p_N\}$. For each patch p_k and transformer layer l , we compute the fault score $S_{k,l}$ that quantifies its contribution to the error:

$$S_{k,l} = \Delta \hat{y} \cdot \exp(H(A_k^l)) \quad (17)$$

Here, $\Delta \hat{y} = \|\hat{y} - \hat{y}_{-p_k^l}\|_2$ measures the change in prediction when patch p_k is ablated at layer l , while $H(A_k^l)$ denotes the entropy of the attention distribution for patch p_k in layer l . The exponential term amplifies the influence of uncertain attention patterns, which are more likely to correlate with hallucinations.

The ablation effect $\hat{y}_{-p_k^l}$ is computed by replacing the patch embedding e_k^l with a zero vector while keeping other activations unchanged. This isolates the causal impact of specific image regions on the final prediction. The attention entropy $H(A_k^l)$ is derived from the softmax distribution over all patches in layer l :

$$H(A_k^l) = - \sum_{j=1}^N A_{k,j}^l \log A_{k,j}^l \quad (18)$$

High entropy indicates that the model attends diffusely to multiple regions rather than focusing on clinically relevant features. In medical imaging, this often corresponds to spurious correlations or lack of decisive evidence.

For MLP layers, we additionally measure the Jacobian $J_k^l = \partial \hat{y} / \partial h_k^l$ of the output with respect to hidden activation h_k^l :

$$F_k^l = \|J_k^l \odot \sigma(h_k^l)\|_1 \quad (19)$$

where σ is the sigmoid function and \odot denotes element-wise multiplication.

The fault localization module outputs a set of critical components $\mathcal{C} = \{(k_1, l_1), \dots, (k_m, l_m)\}$ ranked by their contribution to the error.

4.4 Prototype-Guided Target Optimization

A critical challenge in knowledge editing is determining the target activation v^* that represents the correct model behavior. Unlike language models where discrete token corrections can be specified, medical vision tasks require continuous feature representations. We propose a prototype-guided approach that derives v^* from the class-conditional feature distribution of correctly classified samples.

For a faulty activation k^* belonging to an incorrectly predicted class \hat{c} when the true class is c , we first collect a set of correctly classified activations $\mathcal{A}_c = \{a_1, a_2, \dots, a_M\}$ from class c . The target activation is computed as a weighted combination of the class centroid and the original activation:

$$v^* = \alpha \cdot \mu_c + (1 - \alpha) \cdot k^* \quad (20)$$

where μ_c is the centroid of class c :

$$\mu_c = \frac{1}{M} \sum_{i=1}^M a_i \quad (21)$$

The interpolation coefficient $\alpha \in (0, 1]$ controls the strength of correction. Setting $\alpha = 1$ moves the activation entirely to the class centroid, while smaller values provide gentler corrections that preserve instance-specific features. To account for intra-class variation, we refine the target using nearest-neighbor guidance within the correct class:

$$v^* = \alpha \cdot \mu_c + \beta \cdot a_{nn} + (1 - \alpha - \beta) \cdot k^* \quad (22)$$

where $a_{nn} = \arg \min_{a \in \mathcal{A}_c} \|a - k^*\|_2$ is the nearest neighbor in the correct class embedding space, and β weights the contribution of this similar exemplar. This formulation ensures that corrections respect the local structure of the feature manifold while guiding the activation toward the correct decision boundary.

For concept-level editing (Section 4.5), the target optimization extends to the Riemannian mean formulation (Equation 27), which better captures the non-Euclidean geometry of high-dimensional medical feature spaces.

4.5 Adaptive Layer Selection in Null-Space Editing

The effectiveness of Medical AlphaEdit depends on strategically selecting which transformer layers to modify. Rather than applying edits uniformly across all layers, we develop an adaptive selection mechanism that identifies the most influential layers for each error case. This approach stems from the observation that different medical concepts are encoded at varying depths of the network [?].

For a given faulty activation k_l^* , we compute its layer-wise influence score ρ_l as:

$$\rho_l = \|W_l k_l^*\|_2 \cdot \sigma_l \quad (23)$$

where W_l represents the weight matrix at layer l , k_l^* is the layer-specific activation, and σ_l denotes the singular value spread of W_l . The first term measures the magnitude of the transformation applied to the faulty activation, while the second term captures the layer’s overall sensitivity to input variations.

The edit probability p_l for layer l is then determined through a softmax distribution:

$$p_l = \frac{\exp(\beta \rho_l)}{\sum_{j=1}^L \exp(\beta \rho_j)} \quad (24)$$

Here, β serves as a temperature parameter controlling the selectivity of the layer selection. Higher values of β result in more concentrated edits on the most influential layers. We find $\beta = 2.0$ provides a good balance between specificity and generalization across various medical imaging tasks.

During implementation, we sample M layers for editing according to the probabilities $\{p_l\}$, where M is typically 20-30% of the total layers. This stochastic selection ensures diversity in the edited layers across different error cases while maintaining computational efficiency. The selected layers then receive the null-space constrained updates described in Equation 13.

4.6 Extension to Concept-Level Editing

The instance-specific editing framework can be extended to address broader concept-level errors in medical vision transformers. Rather than correcting individual misclassifications, we propose to modify the model’s representation of entire medical concepts (e.g., “pneumonia” or “normal lung”) by aligning them with prototypical feature distributions.

Let $\mathcal{C} = \{c_1, \dots, c_K\}$ denote a set of medical concepts requiring correction. For each concept c_k , we collect a set of correct feature activations $F_k = \{f_1, \dots, f_M\}$ from cases where the model correctly identifies the concept. The prototypical representation \bar{f}_k is computed as the Riemannian mean on the manifold of positive definite matrices:

Figure 2: Concept-Level Editing Framework for Medical Vision Transformers. The Riemannian mean captures the non-Euclidean geometry of medical feature spaces.

$$\bar{f}_k = \arg \min_f \sum_{i=1}^M \delta^2(f, f_i) \quad (25)$$

where $\delta(\cdot, \cdot)$ denotes the geodesic distance between feature vectors. This approach accounts for the non-Euclidean structure of medical feature spaces [?]. The target correction v^* in Equation 13 is then replaced with \bar{f}_k , and the faulty activation k^* represents the model’s current (incorrect) concept embedding.

The concept-level update must preserve relationships between similar medical conditions. We enforce this by adding an orthogonality constraint between updates for related concepts:

$$\Delta W_{c_i}^T \Delta W_{c_j} = 0 \quad \forall (c_i, c_j) \in \mathcal{R} \quad (26)$$

where \mathcal{R} defines clinically related concept pairs (e.g., different types of lung opacities). This ensures edits for one condition do not inadvertently affect diagnostically similar cases.

The projection matrix P is constructed from activations of all correct concepts, making it more comprehensive than instance-specific versions:

$$P = I - \sum_{k=1}^K \bar{f}_k (\bar{f}_k^T \bar{f}_k)^{-1} \bar{f}_k^T \quad (27)$$

This formulation allows simultaneous correction of multiple conceptual errors while maintaining diagnostic coherence across the entire medical taxonomy. The approach is particularly valuable for addressing systematic biases (e.g., underdiagnosis in specific patient demographics) that manifest as consistent concept-level misunderstandings.

The extension naturally handles hierarchical medical knowledge. For example, edits to a general concept like “lung abnormality” automatically propagate to specific sub-concepts (e.g., “pneumonia”, “atelectasis”) through the attention mechanism’s query-key structure. This property emerges because medical taxonomies are implicitly encoded in the transformer’s attention patterns [?]. Implementation requires careful balancing between concept specificity and generalizability. Overly specific prototypes may lead to brittle corrections, while excessively broad ones could dilute diagnostic precision. We address this through multi-scale prototyping, maintaining separate representations at different levels of the medical hierarchy (e.g., organ-level, finding-level, diagnosis-level). The appropriate prototype scale for editing is automatically selected based on the entropy of the concept’s feature distribution.

The concept-level framework also enables proactive error prevention. By analyzing the null-space distance between known problematic concepts and correct ones, we can identify potential error cases before they occur:

$$d(c_i, c_j) = \|\bar{f}_i - P\bar{f}_j\|_2 \quad (18)$$

Large distances indicate concepts likely to be confused, allowing preemptive reinforcement of their distinguishing features. This anticipatory capability is unique to the medical domain where error patterns often follow known clinical differential diagnoses.

5 Experiments

5.1 Experimental Setup

Datasets: We evaluate Medical AlphaEdit on three medical imaging benchmarks of increasing complexity:

- **PathMNIST** [?]: 107,180 colon pathology images (224×224) across 9 tissue types, serving as a controlled testbed.
- **ChestX-ray14** [?]: 112,120 frontal-view chest radiographs with 14 thoracic disease labels. This multi-label dataset presents realistic clinical complexity with co-occurring pathologies.
- **CheXpert** [?]: 224,316 chest radiographs from 65,240 patients with hierarchical uncertainty labels, enabling evaluation on ambiguous clinical cases.

Models: We test on three medical vision transformer architectures:

- **MedViT-Base** [?]: A medical-specific ViT with 86M parameters.
- **ViT-B/16** [?]: Pre-trained on ImageNet and fine-tuned on medical data, representing common transfer learning scenarios.
- **Swin-T** [?]: Hierarchical ViT with shifted windows, testing generalization across architectures.

Baselines: We compare against six methods spanning different paradigms:

1. **Fine-tuning (FT)**: Full model retraining on corrected examples [?]—the upper bound for adaptation but computationally expensive.
2. **LoRA** [?]: Low-rank adaptation with rank $r = 16$, representing state-of-the-art parameter-efficient fine-tuning.
3. **ROME** [?]: Rank-one model editing originally designed for language models, adapted to ViTs.

4. **MEMIT** [?]: Mass-editing memory in transformers, enabling batch corrections.
5. **Attention Correction (AC)**: Post-hoc attention map adjustment [?].
6. **NS-FT**: Our ablation using null-space projection during fine-tuning (requires gradient computation).

Evaluation Metrics: We design metrics to capture both efficacy and safety of editing:

- **Efficacy (ES)**: Percentage of target errors successfully corrected after editing.
- **Generalization (GS)**: For concept-level editing, the accuracy improvement on *held-out* samples of the same class—critical for detecting overfitting vs. true concept correction.
- **Locality/Specificity (LS)**: Accuracy retention on unrelated classes (e.g., editing pneumonia should not affect fracture detection).
- **Knowledge Preservation Rate (KPR)**: Percentage of originally correct predictions unchanged after editing—the core safety metric.
- **AUC**: Area under ROC curve for multi-label classification on ChestX-ray14/CheXpert.

5.2 Main Results

Table 1 presents comprehensive comparisons across datasets and methods. Medical AlphaEdit achieves the best balance between correction efficacy and knowledge preservation.

Table 1: Performance comparison on ChestX-ray14. ES: Efficacy Score, GS: Generalization Score, LS: Locality Score, KPR: Knowledge Preservation Rate. Best results in **bold**, second-best underlined.

Method	ES (%)	GS (%)	LS (%)	KPR (%)	AUC
Fine-tuning (FT)	82.3	71.2	85.4	89.2	0.812
LoRA ($r=16$)	79.8	68.5	91.3	93.1	0.805
ROME	75.2	52.1	78.6	82.4	0.783
MEMIT	78.4	58.3	81.2	85.7	0.791
AC	72.8	45.2	<u>94.8</u>	<u>94.1</u>	0.779
NS-FT	84.1	73.8	89.5	93.5	<u>0.818</u>
Medical AlphaEdit	89.7	81.4	96.2	98.2	0.831

Key Findings:

Figure 3: Correction accuracy across transformer layers in MedViT-Base. Middle layers (6-12) encode disease-specific features and show highest editing efficacy.

- **Efficacy:** Medical AlphaEdit corrects 89.7% of errors, outperforming the best baseline (NS-FT, 84.1%) by 5.6%. ROME/MEMIT, designed for discrete token editing, struggle with continuous visual features.
- **Generalization:** The 81.4% GS demonstrates true concept-level correction rather than instance overfitting. LoRA and ROME show significantly lower GS (68.5%, 52.1%), indicating their corrections do not transfer to similar cases.
- **Locality/KPR:** Medical AlphaEdit achieves 96.2% locality and 98.2% KPR, validating the null-space constraint’s theoretical guarantees. Fine-tuning and ROME show notable degradation (85.4%, 78.6% locality).

Table 2 shows results on CheXpert with uncertainty labels, where Medical AlphaEdit demonstrates robust performance on ambiguous cases.

Table 2: Performance comparison on CheXpert (5 competition classes).

Method	ES (%)	KPR (%)	AUC	Time/Edit
LoRA	76.2	91.8	0.879	42 min
MEMIT	74.5	84.2	0.865	1.2 s
Medical AlphaEdit	86.3	97.4	0.892	0.8 s

5.3 Layer-Wise Analysis

Figure 3 illustrates how editing different layers affects correction performance. Middle layers (6-12 in MedViT) show the highest correction efficacy, aligning with medical feature abstraction occurring in these regions. Early layers primarily process low-level features, while late layers handle high-level integration—both less optimal for targeted edits.

The adaptive layer selection mechanism automatically identifies these optimal editing zones. Compared to uniform layer editing, it improves ES by 12.4% while reducing the number of modified parameters by 37%.

5.4 Concept-Level Editing and Bias Mitigation

For systematic errors affecting entire diagnostic categories, concept-level editing demonstrates significant advantages over instance-level approaches (Table 3).

Key Observations:

Table 3: Concept-level vs. instance-level editing on ChestX-ray14. N : number of edit samples used.

Method	N	ES (%)	GS (%)	LS (%)	KPR (%)
Instance-level ($N=1$)	1	91.2	42.3	95.1	97.8
Instance-level ($N=10$)	10	88.5	58.7	93.2	96.4
Concept-level (Euclidean)	10	85.3	71.5	94.8	97.1
Concept-level (Riemannian)	10	86.8	81.4	96.2	98.0

- **Generalization:** Instance-level editing with $N=1$ achieves high ES (91.2%) but poor GS (42.3%), indicating severe overfitting. Concept-level editing with Riemannian mean achieves 81.4% GS, demonstrating true concept correction.
- **Riemannian vs. Euclidean:** The Riemannian mean outperforms Euclidean mean by 9.9% in GS, validating the importance of respecting the non-Euclidean geometry of medical feature spaces.

Demographic Bias Mitigation: We evaluate on the CheXpert dataset stratified by sex, where models exhibit known underdiagnosis patterns in female patients [?]. Table 4 shows concept-level editing effectively reduces diagnostic disparities.

Table 4: AUC by demographic group before and after concept-level editing on CheXpert (Cardiomegaly).

	Male	Female	Gap
Before Editing	0.891	0.823	0.068
After Editing	0.887	0.871	0.016

The demographic gap is reduced by 76.5% (from 0.068 to 0.016) while maintaining overall performance on male patients, demonstrating targeted bias correction.

5.5 Computational Efficiency

Medical AlphaEdit’s closed-form solution provides dramatic efficiency gains compared to gradient-based alternatives (Table 5).

Medical AlphaEdit is **315× faster than Fine-tuning** and **52× faster than LoRA**, enabling real-time error correction in clinical workflows. The memory efficiency (4 GB) allows deployment on standard clinical workstations without dedicated GPU infrastructure.

Table 5: Computational comparison on NVIDIA A100 GPU for editing a single error.

Method	Time	GPU Memory	Gradient Comp.	Training Data
Fine-tuning	4.2 h	24 GB	Yes	1000+ samples
LoRA	42 min	12 GB	Yes	100+ samples
MEMIT	1.2 s	8 GB	No	1 sample
Medical AlphaEdit	0.8 s	4 GB	No	1 sample

5.6 Ablation Studies

We ablate key components of Medical AlphaEdit to understand their contributions (Table 6).

Table 6: Ablation study on ChestX-ray14.

Configuration	ES (%)	GS (%)	KPR (%)
Full Model	89.7	81.4	98.2
w/o Null-space Projection	87.2	79.1	84.3
w/o Adaptive Layer Selection	82.5	74.2	97.8
w/o Prototype-guided Target	85.1	61.3	97.5
w/o Attention Entropy in Fault Score	86.3	78.9	97.9

Findings:

- Removing null-space projection causes severe KPR degradation (98.2% → 84.3%), confirming its critical role in knowledge preservation.
- Adaptive layer selection contributes 7.2% ES improvement over uniform editing.
- Prototype-guided target is essential for generalization (81.4% → 61.3% GS without it).

6 Discussion and Future Work

6.1 Limitations

While Medical AlphaEdit demonstrates strong performance in error correction, several limitations warrant discussion. The framework currently requires manual identification of erroneous cases before applying corrections, which may not scale efficiently in real-world clinical deployments. Automated error detection remains challenging due to the subtle nature of medical misclassifications and the lack of reliable confidence estimates from vision transformers [?]. Moreover, the null-space projection assumes linear separability of correct

and incorrect knowledge representations, an approximation that may not hold for highly nonlinear medical feature spaces.

The concept-level editing extension shows promise but introduces new complexities. Defining medical concept boundaries becomes nontrivial for conditions with overlapping radiographic features (e.g., different types of pulmonary opacities). Current prototype representations also struggle with rare conditions where few exemplars exist for computing reliable feature means. These limitations suggest opportunities for incorporating kernel methods or manifold learning for nonlinear null-space computation in future iterations.

6.2 Broader Impact and Clinical Applications

Beyond direct error correction, Medical AlphaEdit enables several valuable clinical applications. The framework could support continuous model adaptation to new medical knowledge without catastrophic forgetting—a critical need as diagnostic guidelines evolve [?]. For teaching hospitals, the editing traces provide interpretable demonstrations of how specific radiographic features influence diagnoses, potentially serving as educational tools for radiology trainees.

The technology also shows promise for addressing demographic disparities in medical AI. By selectively reinforcing underrepresented features in minority populations, the framework could mitigate well-documented racial and gender biases in diagnostic algorithms [?]. Early experiments show particular effectiveness in correcting underdiagnosis patterns for female patients in cardiovascular imaging, where anatomical differences often lead to systematic errors.

6.3 Ethical Considerations in Medical AlphaEdit

The ability to directly modify medical AI systems raises important ethical questions that the community must address. While null-space projection theoretically preserves existing knowledge, insufficient validation could inadvertently introduce new errors during the editing process. We advocate for rigorous clinician oversight and prospective validation before deploying edited models in patient care settings.

The framework also necessitates careful documentation standards. Each edit should be accompanied by metadata including: the original error case, clinical rationale for correction, validation evidence, and responsible parties. Such audit trails will be essential for maintaining accountability as these systems enter regulatory review processes [?].

Future work should explore several promising directions. Developing automated error detection mechanisms would enable closed-loop correction systems that continuously improve without human intervention. Extending the null-space formalism to handle nonlinear representations through kernel methods or manifold learning could enhance editing precision. Finally, creating standardized benchmarks for medical model editing would accelerate progress in this emerging field, similar to existing efforts in general AI safety [?].

7 Conclusion

Medical AlphaEdit presents a novel paradigm for addressing errors and hallucinations in medical vision transformers without the computational burden of retraining. By leveraging causal tracing and null-space projection, the framework achieves precise corrections while preserving existing diagnostic knowledge. The experimental results demonstrate significant improvements in correction accuracy and hallucination suppression across multiple medical imaging modalities, outperforming conventional fine-tuning approaches.

The train-free nature of Medical AlphaEdit makes it particularly suitable for clinical deployment, where rapid error correction is essential but computational resources may be limited. The framework’s ability to perform both instance-specific and concept-level edits provides flexibility in addressing diverse error types, from individual misclassifications to systematic biases. The adaptive layer selection mechanism further enhances efficiency by focusing edits on the most influential components of the transformer architecture.

Looking ahead, the principles underlying Medical AlphaEdit could extend beyond vision transformers to other medical AI systems, including multimodal models that integrate imaging with clinical notes or genomic data. The null-space projection approach offers a generalizable strategy for model editing that maintains system integrity while enabling targeted improvements. As medical AI continues to advance, frameworks like Medical AlphaEdit will play a crucial role in ensuring these technologies remain reliable and trustworthy in real-world healthcare settings.

The success of this approach highlights the importance of developing specialized techniques for medical AI that account for the unique challenges of clinical decision-making. By combining rigorous mathematical foundations with clinical insights, Medical AlphaEdit represents a significant step toward more robust and adaptable diagnostic systems. Future research should explore integration with automated error detection and validation protocols to create comprehensive solutions for maintaining AI performance throughout its lifecycle in healthcare applications.

References