Beyond Explanation: Debugging Medical Imaging Models via Concept Intervention

Anonymous Author(s)

Abstract. Medical imaging models often operate as black boxes, limiting interpretability and systematic debugging. We introduce an easy-to-use, plug-and-play framework for concept-based interpretation and model refinement. By aligning a single-modality encoder [6] to BioMedCLIP [14], we construct a Concept Bottleneck Model (CBM) that enables concept-level interventions. These interventions allow us to isolate causal versus spuriously correlated concepts, validate insights with domain experts, and generate counterfactual samples for targeted fine-tuning. We evaluate our framework on a Mayo Clinic ultrasound dataset and the CheXpert 5x200 chest X-ray dataset [4]. Results demonstrate that concept intervention enables reliable model diagnosis while maintaining, and occasionally improving predictive performance via guided fine-tuning. Our findings highlight the practical value of this framework for controlled, interpretable refinement of clinical deep learning models.

1 Introduction

Deep learning models achieve strong performance across diverse medical imaging tasks, including disease classification, segmentation, and prognosis. Model Interpretability is key in establishing trust between the practicing clinitian and AI models. We need models that can describe what they segment or classify in understandable text. Most modern image encoders operate as black boxes, producing predictions without explicit links to clinically meaningful abstractions, which complicates trust, validation, and systematic error analysis by domain experts in clinical environments.

Traditional post-hoc interpretability approaches [9,3] predominantly rely on visualization techniques such as saliency maps or gradient-based attribution. While these methods highlight regions of interest, they lack semantic grounding, exhibit high sensitivity to perturbations, and fail to align with clinical reasoning [1]. Clinicians typically interpret medical images through structured, human-understandable concepts, such as radiographic findings, rather than pixel-level activations. Bridging this gap remains an open challenge.

Concept Bottleneck Models (CBMs) [5] offer a promising alternative by explicitly reasoning over predefined concepts. However, most CBM approaches require dense concept-level supervision and end-to-end training [8], limiting their applicability to pre-trained, frozen, or deployed models. Concurrently, vision–language (VL) foundation models like CLIP [10] demonstrate strong semantic alignment between images and text, yet their clinical utility has largely

remained confined to passive zero-shot classification or retrieval, rather than enabling fine-grained debugging of task-specific models.

In this work, we propose a plug-and-play, concept-based fine-tuning framework that leverages large vision–language models as a shared semantic representation space to interpret and debug single-modality medical imaging models. Following recent insights that a linear mapping is sufficient to align disparate representation spaces [6,7], we project a task-specific image encoder into the BioMedCLIP [14] embedding space. This projection enables semantic grounding without requiring intrinsic textual supervision or end-to-end retraining. On top of this aligned representation, we construct a CBM using concepts derived from radiology reports or structured annotations. Crucially, we leverage gradient-guided concept-level interventions to turn interpretability into an active mechanism for model optimization, allowing us to systematically isolate missing, weak, or spurious concepts.

Our contributions are summarized as follows: (i) **Modular Architecture:** We introduce a plug-in concept bottleneck model (CBM) enabling concept-level interpretability for off-the-shelf single-encoder medical models via linear alignment to BioMedCLIP; (ii) **Expert Validation:** We validate the clinical utility of our concept-based explanations with domain experts, demonstrating tight alignment with clinical reasoning; (iii) **Causal Discovery:** We leverage intervenable CBMs for counterfactual concept interventions to systematically isolate causal concepts from spurious features; and (iv) **Intervention-Guided Refinement:** We utilize our intervention strategy to apply counterfactual concept perturbations to held-out data to fine-tune the classification head and boost downstream predictive performance.

We evaluate this framework on the Mayo Clinic ultrasound and CheXpert 5×200 chest X-ray datasets. With this modular plug-in tool, we are able to identify spurious and causal correlations in a highly data-efficient manner, using clinical concepts to train models to dial into the correct diagnostic features within the image and more closely align human medical learning with traditional machine learning. While human clinicians excel at learning a great deal from a small amount of data, we argue that with a Concept Bottleneck Model (CBM), machines can do so as well. Crucially, our framework achieves this paradigm shift while maintaining baseline performance and systematically boosting downstream capabilities (2–3%) through our targeted intervention strategy.

2 Related Work

Feature Attribution vs. Concept Interpretability While widely adopted, literature demonstrates that these pixel-level heatmaps inherently suffer from notable vulnerabilities, frequently offering purely qualitative visualizations that lack structured semantic grounding while exhibiting high sensitivity to network architectures and minor input perturbations [1,12]. Concept-based paradigms fundamentally deviate from this approach; instead of attempting to interpret ungrounded pixel-level features post-hoc, they bypass these established visual-

ization constraints by routing downstream predictions directly through an intermediate layer of explicit, human-understandable semantic representations [5].

Vision-Language Alignment and Post-Hoc CBMs Standard CBMs typically require dense, manual concept annotations and end-to-end retraining, limiting their utility for pre-trained clinical models [8]. To bypass this bottleneck, recent post-hoc or label-free CBMs leverage the joint semantic space of vision–language models like CLIP [10] and BioMedCLIP [14]. These approaches project single-modality image representations onto textual concept spaces without complete model retraining [6]. We build on this paradigm, demonstrating that a lightweight affine mapping is sufficient to anchor complex, single-modality medical encoders within a clinically grounded concept space.

Concept Intervention and Model Refinement Test-time concept interventions allow researchers to simulate counterfactual states, probe model robustness, and analyze causal failure modes [13], [11], [5]. However, existing frameworks typically assume access to full concept labels during training or couple concept prediction tightly with primary task learning. Utilizing expert-guided concept interventions specifically to debug and fine-tune frozen, single-modality medical models remains largely underexplored. Our work bridges this gap by transforming test-time interventions into an active mechanism for targeted latent data augmentation and classification head refinement.

3 Methodology

Our methodology enables concept-based interpretability and controlled refinement of single-modality medical imaging models by aligning their latent representations to a shared vision–language semantic space. The framework consists of three core phases: (i) **Post-Hoc Mapping** extracts features from a task-specific clinical image encoder and linearly aligns its representation space to the BioMedCLIP embedding space; (ii) **Concept Bottleneck Construction** generates semantically grounded concept predictions using textual abstractions derived from clinical knowledge; and (iii) **Counterfactual Intervention** leverages expert-guided concept perturbations to isolate causal factors, flag spurious concepts, and drive targeted counterfactual data augmentation for downstream classification head fine-tuning. Overview of the pipeline is illustrated in Figure 1

3.1 Post-Hoc Alignment and CBM Construction

We begin with a pre-trained, off-the-shelf single-modality medical imaging model that operates purely on images and lacks intrinsic textual supervision. To obtain a stable and semantically rich representation, we extract image features from an intermediate layer prior to the final classification layer. Specifically, features are isolated after global average pooling and before the final nonlinearity, as this

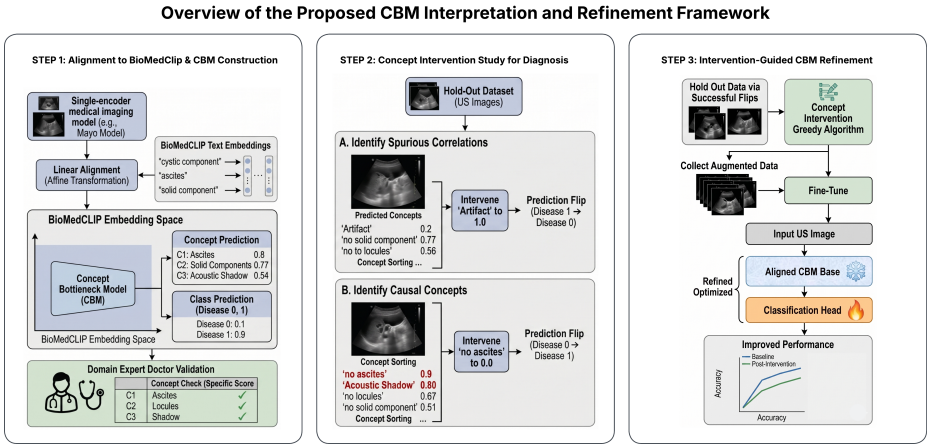

Fig. 1: Overview of the proposed framework. A single-modality medical imaging model is linearly aligned to the BioMedCLIP embedding space to enable concept-based interpretability via a concept bottleneck model (CBM). Concept interventions on held-out data identify causal versus spurious concepts through prediction flips. These interventions are then used to augment data and fine-tune the CBM classification head, improving predictive performance.

representation preserves linear structure while retaining highly discriminative information. These extracted features serve as fixed image embeddings.

To enable semantic grounding, we align the feature space of this single-modality model to the embedding space of BioMedCLIP [14], a large-scale vision–language model pre-trained on approximately 15 million medical image–report pairs. We explicitly note that foundation models like BioMedCLIP are fundamentally pre-trained on public scientific literature rather than native, institutional Electronic Health Record (EHR) systems. Furthermore, clinical radiology notes typically describe global, multi-view findings across an entire study or scan volume rather than localizing concepts to specific 2D images or individual slices. While a production-level deployment would ideally be trained natively on paired local EHR data, this work serves as an intentional proof-of-concept to establish the baseline feasibility of post-hoc cross-model alignment under these broader domain constraints. Following prior work on cross-model alignment [6], we learn an affine transformation that maps the single-modality image features directly into this shared BioMedCLIP embedding space.

3.2 Counterfactual Concept Intervention and Diagnosis

We formulate model validation as a problem of counterfactual concept-level intervention to cleanly isolate spurious correlations from informative clinical features. Formally, an intervention operator $\mathcal{I}(\hat{\mathbf{c}}, \mathcal{M})$ overwrites a targeted subset of predicted concept activations $\hat{\mathbf{c}}$ with alternative binary states specified by an intervention map $\mathcal{M}$, generating a counterfactual concept vector $\tilde{\mathbf{c}}$.

To efficiently identify the minimal concept set necessary to alter the model's decision, we implement the gradient-guided greedy strategy detailed in Algorithm 1 (inspired by [11]). This protocol prioritizes candidate interventions by ranking concepts according to their absolute gradient magnitude with respect to the target class logit, $\nabla_{\hat{\mathbf{c}}} g_{y^*}(\hat{\mathbf{c}})$.

Algorithm 1 Greedy Concept-Level Intervention

Require: Concept activations $\hat{\mathbf{c}} \in \mathbb{R}^m$, classification head $g(\cdot)$, target label y^*, intervention operator $\mathcal{I}(\cdot)$, maximum concept search depth K
Ensure: Minimal set of concept interventions $\mathcal{M}$ that flips the prediction to y^*
Compute original baseline prediction $\hat{y} = \arg\max g(\hat{\mathbf{c}})$
if $\hat{y} = y^*$ **then**
  return $\emptyset$ {Target prediction already satisfied}
end if
Compute gradient vector $\mathbf{G} = \nabla_{\hat{\mathbf{c}}} g_{y^*}(\hat{\mathbf{c}})$ of the target class logit with respect to concept activations
Rank concepts $\{c_1, \ldots, c_m\}$ by descending absolute gradient magnitude of $\mathbf{G}$
for $k = 1$ to K **do**
  Isolate top-k ranked concepts $\mathcal{C}_k$
  for each binary counterfactual assignment $\mathbf{v} \in \{0, 1\}^k$ **do**
    Construct intervention map $\mathcal{M} = \{(c_i \leftarrow v_i) \mid c_i \in \mathcal{C}_k\}$
    Apply intervention to synthesize counterfactual: $\tilde{\mathbf{c}} = \mathcal{I}(\hat{\mathbf{c}}, \mathcal{M})$
    Compute counterfactual prediction $\tilde{y} = \arg\max g(\tilde{\mathbf{c}})$
    if $\tilde{y} = y^*$ **then**
      return $\mathcal{M}$ {Minimal flipping set discovered}
    end if
  end for
end for

return $\emptyset$ {No successful intervention set found within depth K}

Interventions are applied sequentially to the top-k concepts by pushing their activations toward binary extremes. The smallest intervention set that successfully induces a prediction flip provides a concise counterfactual explanation of the model's decision boundary.

3.3 Intervention-Guided Refinement

We leverage successful prediction flips from the diagnostic phase to drive targeted model refinement. When Algorithm 1 identifies an intervention map $\mathcal{M}$ that corrects an erroneous baseline prediction, the resulting counterfactual concept vector $\tilde{\mathbf{c}}$ is aggregated into a concept-guided counterfactual augmentation dataset $\mathcal{D}_{aug} = \{(\tilde{\mathbf{c}}_n, y_n^*)\}$. Operating entirely within the latent bottleneck layer avoids computationally expensive or ungrounded pixel-level image synthesis while generating targeted training signals that expose decision-boundary vulnerabilities.To prevent representation collapse and preserve feature robustness, the image encoder $f_s(\cdot)$ and linear alignment matrix W remain frozen. Fine-tuning is strictly restricted to the linear task classification head $g(\cdot)$. Optimizing $g(\cdot)$ over

Table 1: Performance evaluation on held-out Mayo Clinic test splits comparing the baseline ensemble, pre-intervention, and post-intervention fine-tuned CBM.

Model Configuration	Test Set	Accuracy	F1	AUROC
Mayo Ensemble (Baseline)	Test Set 1	0.890	0.880	0.920
	Test Set 2	0.730	0.740	0.840
CBM (Pre-Intervention)	Test Set 1	0.850	0.850	0.916
	Test Set 2	0.740	0.740	0.839
CBM (Post-Intervention)	Test Set 1	0.850	0.850	**0.917**
	Test Set 2	**0.753**	**0.749**	0.839

Table 2: CheXpert per-label performance comparison between TorchXrayVision baseline and our CBM framework before and after intervention-based fine-tuning.

Pathology Class	TorchXrayVision Baseline				CBM (Pre-Intervention)				CBM (Post-Intervention)			
	Prec	Rec	F1	AUC	Prec	Rec	F1	AUC	Prec	Rec	F1	AUC
Cardiomegaly	0.63	**0.68**	0.65	**0.86**	**0.73**	0.58	0.65	0.85	**0.73**	0.58	0.65	0.85
Consolidation	**0.50**	0.05	0.09	0.71	0.39	**0.71**	0.51	0.69	0.43	**0.71**	**0.54**	**0.75**
Edema	0.43	0.21	0.28	0.70	**0.48**	0.45	0.47	**0.83**	**0.48**	**0.50**	**0.49**	0.80
Effusion	0.35	**0.66**	0.46	0.77	**0.73**	0.61	**0.67**	0.89	**0.73**	0.61	**0.67**	**0.90**
Atelectasis	0.43	**0.63**	0.51	0.76	0.91	0.50	0.65	0.84	**0.92**	0.55	**0.69**	**0.90**
Macro Avg	0.47	0.45	0.40	0.76	0.65	0.57	0.59	0.82	**0.66**	**0.59**	**0.61**	**0.84**
Weighted Avg	0.47	0.45	0.40	—	0.64	0.57	0.58	—	**0.65**	**0.59**	**0.60**	—

$\mathcal{D}_{aug}$ explicitly penalizes reliance on spurious artifacts and reinforces sensitivity to causal clinical markers, improving diagnostic accuracy without sacrificing baseline generalization.

4 Results

We evaluate our framework on a clinical ovarian ultrasound dataset from Mayo Clinic (525 samples; 361 train, 164 test) alongside an additional out-of-distribution validation set of 644 samples. To demonstrate domain generalization, we further evaluate on CheXpert 5×200, a balanced multi-label benchmark consisting of 1,000 chest X-rays across five clinical findings, utilizing a pre-trained DenseNet backbone from TorchXrayVision [4] as our base model.

CLIP Alignment Sanity Check To address our first claim, we verify that mapping features to the BioMedCLIP space preserves representation quality. On the CheXpert benchmark in Table 3, the aligned zero-shot model achieves 33.3% accuracy compared to the CLIP baseline of 33.9%, while demonstrating a minor 1.8% improvement in AUROC (64.1% vs. 62.3%). This demonstrates that our learned projection layer retains downstream representation quality within a 1% absolute performance margin, serving as a vital sanity check that we preserve clinical feature discrimination without intensive end-to-end backbone retraining.

Table 3: Overall performance comparison across models on CheXpert 5×200.

Model	Accuracy	Macro F1	AUROC	Interpretability
CLIP Zero-shot	33.9	27.4	62.3	No
Aligned Model (Zero-shot)	33.3	24.5	64.1	Limited
X-ray Vision (Linear Probe)	53.7	52.0	82.1	No
CBM (Pre-intervention)	57.0	59.0	82.0	Yes
CBM (Post-intervention)	**59.0**	**61.0**	**84.0**	Yes

Concept Bottleneck and Intervention Performance Table 1 tracks the progression of performance metrics across the baseline ensemble and our CBM variants on two held-out ultrasound test splits. On the in-distribution Test Set 1, routing predictions through an explicit concept layer preserves the high task accuracy of the uninterpretable baseline, while post-intervention fine-tuning yields a marginal edge in AUROC. On the more complex, out-of-distribution Test Set 2, the pre-intervention CBM performs consistently with baseline trends, whereas targeted counterfactual refinement delivers noticeable improvements in both accuracy and F1-score. These results verify that our framework optimizes downstream decision boundaries without risking performance degradation or representation collapse.

Additional Dataset Study Table 2 presents fine-grained per-label performance on the CheXpert 5x200, contrasting the baseline TorchXrayVision model against our framework. Compared to the uninterpretable baseline, the pre intervention CBM establishes a significantly more balanced trade-off between precision and recall, correcting severe baseline sensitivity flaws on pathologies such as Consolidation. Post-intervention counterfactual fine-tuning delivers further across-the-board gains, improving macro-averaged recall, F1-score, and AUROC.

Finally, Table 3 compiles the global performance metrics across all evaluated frameworks on the CheXpert benchmark. While standard zero-shot and linear probing settings provide competitive baselines, our intervention-guided CBM achieves the highest overall accuracy and Macro F1 while uniquely providing intrinsic model interpretability.

Concept-Based Explanation and Intervention We qualitatively evaluate our framework's debugging utility via case studies from Mayo Clinic ultrasound and CheXpert 5×200 datasets (Figure 2). Unlike traditional image-only models that can overfit to non-clinical artifacts like scanner types or text annotations, our architecture isolates these variables to allow test-time counterfactual interventions. Specifically, adjusting spurious concept weights rectifies false predictions on ovarian masses (left column) and resolves multi-label ambiguities in chest X-rays (right column). This parameter-free auditing mechanism demonstrates how concept-level interventions compensate for imperfect datasets, enabling high-quality clinical modeling with superior sample efficiency and robust insulation from spurious features.

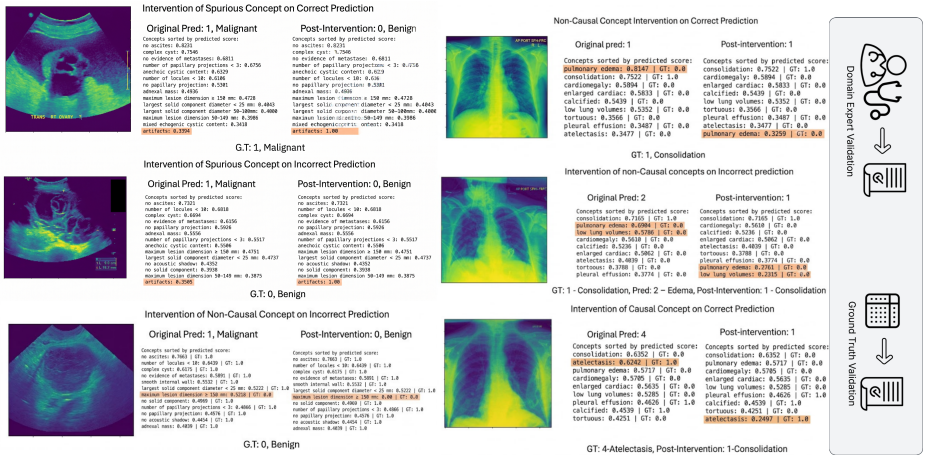

Fig. 2: Interpretation and intervention case studies across ultrasound (Mayo Clinic) and chest X-ray (CheXpert) modalities

5 Conclusion

We present a plug-in framework for concept-based interpretability and refinement of off-the-shelf medical deep learning models [5]. By linearly mapping a pre-trained encoder to the BioMedCLIP embedding space, our approach constructs an intervenable Concept Bottleneck Model (CBM) without architectural modifications. Beyond passive explanation, gradient-guided concept interventions enable domain experts to perform counterfactual debugging and targeted classification head fine-tuning, directly correcting systematic failure modes to enhance trust in clinical deployments.

A key limitation is the spatial and structural disconnect between global clinical narratives and specific image frames. While rich semantic concepts are natively documented within radiology notes at the time of diagnosis, these clinical summaries describe entire multi-view studies globally and lack granular, instance-level association with individual images or slices. Consequently, rather than relying on uncalibrated, off-the-shelf foundation spaces, future research will focus on developing weakly-supervised alignment techniques to dynamically map text-mined concepts from native clinical reports directly to corresponding image volumes [2], mitigating data alignment bottlenecks across diverse clinical environments.

Prospects of Application: This framework can be deployed as an auditing plug-in for legacy clinical workflows. In interactive environments, it enables real-time, human-in-the-loop decision debugging, allowing clinicians to dynamically verify model reasoning, suppress misleading artifacts, and safely refine diagnostic heads at the point of care without end-to-end retraining.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)
2. Bhalla, U., Oesterling, A., Srinivas, S., Calmon, F., Lakkaraju, H.: Interpreting clip with sparse linear concept embeddings (splice). Advances in Neural Information Processing Systems **37**, 84298–84328 (2024)
3. Bhati, D., Neha, F., Amiruzzaman, M.: A survey on explainable artificial intelligence (xai) techniques for visualizing deep learning models in medical imaging. Journal of Imaging **10**(10), 239 (2024)
4. Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., et al.: Torchxrayvision: A library of chest x-ray datasets and models. In: International Conference on Medical Imaging with Deep Learning. pp. 231–249. PMLR (2022)
5. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
6. Moayeri, M., Rezaei, K., Sanjabi, M., Feizi, S.: Text-to-concept (and back) via cross-model alignment. In: International Conference on Machine Learning. pp. 25037–25060. PMLR (2023)
7. Oikarinen, T., Weng, T.W.: Clip-dissect: Automatic description of neuron representations in deep vision networks. arXiv preprint arXiv:2204.10965 (2022)
8. Patrício, C., Neves, J.C., Teixeira, L.F.: Coherent concept-based explanations in medical image and its application to skin lesion diagnosis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3799–3808 (2023)
9. Patrício, C., Neves, J.C., Teixeira, L.F.: Explainable deep learning methods in medical image classification: A survey. ACM Computing Surveys **56**(4), 1–41 (2023)
10. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
11. Shin, S., Jo, Y., Ahn, S., Lee, N.: A closer look at the intervention procedure of concept bottleneck models. In: International Conference on Machine Learning. pp. 31504–31520. PMLR (2023)
12. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. In: ICML Workshop on Visualization for Deep Learning (2017)
13. Yeh, C.K., Kim, B., Arik, S., Li, C.L., Pfister, T., Ravikumar, P.: On completeness-aware concept-based explanations in deep neural networks. Advances in neural information processing systems **33**, 20554–20565 (2020)
14. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)