RESEARCH ARTICLE OPEN ACCESS

MultiModal Counterfactuals (MM-CF): A Framework for Explainable Vision-Language Models in Medical Diagnostics

[1.]P. Boomangai [2.] S. Nisha [3.] T. Sivaranjani [4.]R.Sowmiya

[1.]Assistant Professor Dept of CSE
[2,3,4.] UG Students

Computer Science and Engineering, A.V.C College of Engineering, Mayiladuthurai,TamilNadu

ABSTRACT

Vision-language models (VLMs) are increasingly used in medical diagnostics, but their black-box nature limits clinical adoption. Existing explainability methods rely on unimodal saliency maps that fail to capture cross-modal interactions. This paper introduces **MultiModal Counterfactuals (MM-CF)** , a framework generating counterfactual explanations spanning visual and textual modalities using joint latent space optimization to identify minimal, clinically plausible perturbations that flip model predictions. Evaluated on chest X-ray classification (MIMIC-CXR) and skin lesion diagnosis (HAM10000), MM-CF achieves 94% validity with 2.3 average edits. A clinician user study (n=8 specialists) shows MM-CF improves diagnostic confidence by 31% and reduces decision time by 40% versus baselines.

Keywords: Explainable AI, Counterfactual Explanations, Multimodal Learning, Vision Language Models, Medical Diagnostics

1. Introduction

Vision-language models (VLMs) like CheXzero [1] have achieved remarkable performance in medical imaging, yet their black-box nature limits clinician trust and adoption. Clinicians require transparent, actionable explanations for high-stakes decisions.

Existing explainable AI (XAI) methods suffer critical limitations. Saliency maps like Grad CAM [2] highlight image regions but cannot explain *why* decisions are made nor capture cross-modal interactions between images and text. Counterfactual explanations offer a superior alternative by answering: *"What minimal changes would flip the model's prediction?"* [3], aligning with clinical reasoning where physicians consider alternative diagnoses based on hypothetical changes.

However, existing counterfactual methods are unimodal—operating on images [4] or

text [5] alone. Medical diagnostics inherently involves multimodal reasoning; a radiologist interprets chest X-rays alongside clinical context. Explanations must reflect this synergy.

We introduce **MultiModal Counterfactuals (MM-CF)** , the first framework for cross-modal counterfactual explanations for medical VLMs. **Contributions:** (1) Joint latent space optimization for coordinated visual-textual counterfactuals with clinical plausibility constraints, (2) Quantitative evaluation on two medical datasets, (3) Clinician study showing 31% confidence improvement and 40% faster decisions.

2. Related Work

XAI in Medical Imaging: Gradient methods like Grad-CAM [2] generate heatmaps of influential regions. Perturbation methods like LIME [6] approximate local behavior. Both identify correlation rather than causation and cannot answer "what if" questions [7].

Counterfactual Explanations: Wachter et al. [3] formalized counterfactuals as minimal input changes altering predictions. DiVE [4] generates image counterfactuals using GANs; Polyjuice [5] modifies text. None address cross-modal coordination.

Medical VLMs: Models like CheXzero [1] and BioViL [8] leverage contrastive learning on paired image-text data. Interpretability work remains nascent, primarily visualizing attention [9] without counterfactual capabilities.

Table 1 positions MM-CF against existing methods. No prior work enables cross-modal counterfactual generation with clinical plausibility constraints.

Table 1: Comparison with Existing XAI Methods

Grad-CAM [2]	Image	Saliency	No	Limited
LIME [6]	Image/Text	Perturbation	No	Limited
DiVE [4]	Image	Counterfactual	No	No
Polyjuice [5]	Text	Counterfactual	No	Yes
MM-CF (Ours)	Both	Counterfactual	Yes	Yes

3. Methodology

3.1 Problem Formulation

Let $x_{img} \in X_{img}$ be a medical image and $x_{txt} \in X_{txt}$ clinical text. A multimodal model $f: X_{img} \times X_{txt} \rightarrow Y$ predicts $\hat{y} = f(x_{img}, x_{txt})$. A counterfactual (x_{img}', x_{txt}') must satisfy: **Validity** $(f(x_{img}', x_{txt}') \neq \hat{y})$, **Proximity** (minimal changes), and **Plausibility** (clinical realism).

3.2 Joint Latent Space

MM-CF uses a pre-trained multimodal encoder $\mathcal{E}$ (CheXzero [1]) mapping images and text to shared embeddings: $\mathbf{z}_{\mathbf{x}} = \mathcal{E}_{\mathbf{x}}(\mathbf{x}) \in \mathbb{R}^{d}$, $\mathbf{z}_{\mathbf{t}} = \mathcal{E}_{\mathbf{t}}(\mathbf{t}) \in \mathbb{R}^{d}$. The joint representation $\mathbf{z} = [\mathbf{z}_{\mathbf{x}}; \mathbf{z}_{\mathbf{t}}]$ enables coordinated optimization.

3.3 Counterfactual Optimization

For target class $y_{\text{target}} \neq y$, we minimize:

$$\min_{\mathbf{x}', \mathbf{t}'} \mathcal{L}_{\text{valid}} + \lambda_1 \mathcal{L}_{\text{prox}}(\mathbf{x}, \mathbf{x}') + \lambda_2 \mathcal{L}_{\text{prox}}(\mathbf{t}, \mathbf{t}') + \lambda_3 \mathcal{L}_{\text{plaus}}$$

where $\mathcal{L}_{\text{valid}}$ encourages prediction change, $\mathcal{L}_{\text{prox}}$ penalizes latent deviation, and $\mathcal{L}_{\text{plaus}}$ enforces clinical realism via anatomical constraints and medical lexicon validation.

Algorithm 1 iteratively updates latent codes via gradient descent until validity is achieved, returning top-k diverse candidates.

3.4 Modality-Specific Perturbations

Visual: Localized modifications guided by anatomical segmentation (lung fields, lesion boundaries). Operations include region removal, appearance modification.

Text: Controlled generation with medical lexicon constraints using BioBERT [10]. Operations include negation insertion, synonym substitution.

Cross-Modal Coordination: Consistency loss $\mathcal{L}_{\text{cross}} = 1 - \cos(\mathbf{x}(\mathbf{x}'), \mathbf{t}(\mathbf{t}'))$ ensures semantic alignment.

3.5 Plausibility & Diversity

Plausibility enforced via automated validation (CLIP score, discriminator) and clinical constraints. Diversity achieved via determinantal point processes (DPP) [11].

4. Results and Discussion

4.1 Experimental Setup

Datasets: MIMIC-CXR [12] (377,110 chest X-rays, pneumonia classification), HAM10000 [13] (10,015 dermoscopic images, melanoma vs. benign).

Baselines: Grad-CAM, LIME, DiVE, Polyjuice, attention visualization.

Metrics: Validity (flip rate), Proximity (latent distance), Plausibility (clinician rating 1-5), Sparsity (% modified).

4.2 Quantitative Results

Table 2: Performance Comparison (MIMIC-CXR)

DiVE	0.82 ± 0.04	0.34 ± 0.05	3.2 ± 0.3	15.3 %
Polyjuice	0.71 ± 0.06	0.41 ± 0.05	3.4 ± 0.4	22.1 %
MM-CF	**0.94 ± 0.03**	**0.22 ± 0.04**	**4.3 ± 0.3**	**8.7 %**

Table 3: Performance Comparison (HAM10000)

Di VE	0.79 ± 0.05	0.38 ± 0.06	3.1 ± 0.3	17.2 %
Poly juice	0.68 ± 0.07	0.44 ± 0.05	3.3 ± 0.4	24.8 %
M M-CF	**0.91 ± 0.04**	**0.25 ± 0.05**	**4.1 ± 0.4**	**9.5 %**

MM-CF achieves 94% validity on chest X-rays and 91% on dermatology, significantly outperforming unimodal baselines. Lower proximity and sparsity indicate minimal, targeted edits.

Table 4: Ablation Study (MIMIC-CXR)

MM-CF (full)	0.94	4.3	2.3
- cross-modal coordination	0.78	3.1	3.1
- plausibility validator	0.89	2.9	2.1
- anatomical constraints	0.85	3.4	2.8

Cross-modal coordination is critical: removing it reduces validity by 16% and plausibility by 1.2 points.

4.3 Qualitative Examples

Figure 1: Counterfactual Explanation for Pneumonia Classification

[Image: Original chest X-ray with right upper lobe opacity + report "Opacity present" → Pneumonia. Counterfactual: Same X-ray with opacity inpainted + "Clear lung fields" → Normal.]

Case 1: Original opacity and text predict pneumonia. Counterfactual removes opacity and changes text, flipping prediction.

Case 2: Original irregular lesion predicts melanoma. Counterfactual regularizes border, flips to benign.

Case 3: Subtle interstitial markings. Minimal reduction flips prediction, revealing model sensitivity.

4.4 User Study Results

Participants: 5 radiologists, 3 dermatologists. Reviewed 50 challenging cases per specialty. **Table 5: User Study Results**

Grad	No			MM	
Diagnostic Confidence (%)	68	71	76	74	**89**
Time-to-Decision	47	44	41	43	**28**

(s)					
Utility Rating (1-5)	-	2.4	3.1	3.2	**4.3**
"Clinically Useful" (%)	-	34%	62%	58%	**78%**

Figure 2: User Study Results Visualization

[Image: Bar chart showing confidence improvement (68%→89%) and time reduction (47s→28s) across methods.]

MM-CF improves diagnostic confidence by 31% and reduces decision time by 40%. Radiologists rated MM-CF "clinically useful" in 78% of cases vs. 34% for Grad-CAM.

Clinician Feedback:

- *"This tells me what the model would need to see to change its mind—I can double check my own reasoning."*
- *"Multiple counterfactuals help me understand model uncertainty."*

4.5 Error Analysis

Common failure modes (<8% of cases):

- Anatomically impossible edits (mitigated by segmentation)
- Text-image misalignment (cross-modal loss reduced from 12% to 4%)
- Rare pathologies with limited training data

5. Conclusion and Future Work

This paper introduced MM-CF, the first framework for cross-modal counterfactual explanations in medical vision-language models.

Key Findings:

- **Methodology:** Joint latent space optimization enables coordinated visual-textual counterfactuals with clinical plausibility
- **Performance:** 94% validity on MIMIC-CXR, 91% on HAM10000, outperforming unimodal baselines
- **Clinical Impact:** 31% confidence improvement, 40% faster decisions (n=8 specialists)

MM-CF bridges the trust gap limiting black-box AI deployment in healthcare by providing testable "what-if" scenarios aligned with clinical reasoning.

Limitations:

- Computational cost: 3.2 seconds per explanation
- Requires paired image-text data
- Potential for misleading explanations if constraints fail

Future Work:

- Reduce latency to sub-second via model distillation
- Extend to multi-class differential diagnosis
- Multi-center clinical trials with patient outcome measures
- Integration with electronic medical record systems

References

[1] E. Tiu et al., "Expert-level detection of pathologies from chest X-rays," *Nature Biomedical Engineering*, 2022.

[2] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks," *ICCV*, 2017.

[3] S. Wachter et al., "Counterfactual explanations without opening the black box," *Harvard Journal of Law & Technology*, 2018.

[4] L. Goyal et al., "DiVE: Diverse visual counterfactual explanations," *ICML Workshop*, 2019.

[5] A. Wu et al., "Polyjuice: Generating counterfactuals for explaining models," *ACL*, 2021.

[6] R. Guidotti et al., "A survey of methods for explaining black box models," *ACM Computing Surveys*, 2018.

[7] C. Rudin, "Stop explaining black box machine learning models," *Nature Machine Intelligence*, 2019.

[8] S. Zhang et al., "Biomedical vision-language model for multimodal inference," *arXiv:2304.05390*, 2023.

[9] S. Chen et al., "Cross-modal attention for vision-language models," *CVPR Workshop*, 2023.

[10] K. Lee et al., "BioBERT: A pre-trained biomedical language representation model," *Bioinformatics*, 2020.

[11] A. Kulesza, B. Taskar, "Determinantal point processes for machine learning," *Foundations and Trends in ML*, 2012.

[12] A. E. W.

Johnson et al.,

"MIMIC-CXR-JPG,

" *Scientific Data*,

2019. [13] P.

Tschandl et al., "The