

Explainability in Multi-Modal Models

Nina Garcia¹, Noah Hansen²

¹ Assistant Professor, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: nina.garcia985@yahoo.com | ORCID: 5126-2263-1763-5821

² Associate Professor, School of Data Science, Advanced Computing University, Paris, France. Email: noah.hansen429@yahoo.com | ORCID: 4171-9361-5272-2488

ABSTRACT

Multi-modal AI models -- systems that jointly process and reason over text, images, audio, video, and structured data -- have achieved remarkable performance across a wide range of tasks, but their explanations have lagged behind their capabilities. Explainability in multi-modal models is substantially harder than in unimodal systems: explanations must attribute predictions not only across input features within a modality but across modalities, identifying whether a prediction was driven by visual context, textual content, their interaction, or some combination. The stakes are high in applications where multi-modal AI supports consequential decisions -- medical imaging with clinical notes, legal document analysis with evidence photos, financial report analysis with market data visualisations. This study evaluates eight explainability methods applied to six multi-modal architectures across four application domains: medical visual question answering, legal document understanding, financial report analysis, and autonomous driving scene comprehension. Explainability methods include gradient-based saliency (GradCAM-MM, Integrated Gradients), attention-based attribution (Cross-Attention Attribution, DIME), perturbation-based (LIME-MM, SHAP-MM), counterfactual explanation (CACE), and a proposed Cross-Modal Explanation Framework (CMEF). CMEF achieves the highest faithfulness score (0.882) and the best user comprehension rating (3.84/5) in a study with 84 domain expert evaluators. GradCAM-MM shows highest visual saliency alignment (IoU 0.724). All methods show significant degradation on cross-modal interaction explanations -- attributing predictions that depend on the interplay between modalities rather than individual modality contributions. A multi-modal explainability taxonomy and evaluation protocol are proposed.

Keywords: explainability; multi-modal AI; XAI; GradCAM; SHAP; cross-modal attribution; medical VQA; counterfactual explanation; faithfulness; attention attribution

Citation: Garcia and Hansen [2025]. Explainability in Multi-Modal Models. DOI: <https://doi.org/10.5281/zenodo.19614896>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Oct 15 Accepted: 2024 Dec 15 Published: 2025 Feb 24

Research Article: *Explainable AI and Multi-Modal Learning*

1. Introduction

1.1 The Multi-Modal Explainability Gap

When a radiologist asks why an AI system flagged a chest X-ray as abnormal while noting the clinical history in the accompanying report, they expect an explanation that tells them which regions of the image were decisive, which elements of the clinical text mattered, and -- crucially -- how the image and text interacted to produce the finding. Current explainability methods largely cannot provide the third component. Unimodal explanation tools (GradCAM for images, SHAP for tabular data, attention visualisation for text) can be applied to individual modality streams, but they cannot explain cross-modal interactions -- the non-linear dependencies between what is seen and what is read, or between audio and visual context in a video. This cross-modal explanation gap is the central challenge this study addresses.

1.2 Stakes and Regulatory Context

The EU AI Act's transparency requirements (Article 13) and GDPR's right to explanation (Article 22) apply regardless of whether an AI system is unimodal or multi-modal. If anything, the requirement is more pressing for multi-modal systems deployed in high-risk contexts: a patient denied insurance coverage by a system that analysed both their medical images and their clinical notes has a right to understand what drove that decision, and 'the model integrated all your data in a 72-layer transformer' is not a satisfactory explanation. Developing explainability methods that can faithfully represent multi-modal reasoning is both a research priority and a regulatory necessity.

1.3 Study Objectives

This study aims to: (i) evaluate eight explainability methods on six multi-modal architectures across four application domains; (ii) assess explanation faithfulness, visual saliency alignment, and user comprehension; (iii) characterise the cross-modal interaction explanation gap; and (iv) propose a multi-modal explainability taxonomy and evaluation protocol.

2. Literature Review

2.1 Gradient-Based Saliency Methods

GradCAM (Selvaraju et al., 2017) produces visual saliency maps by computing gradients of the predicted class score with respect to the final convolutional layer activations, indicating which image regions most influenced the prediction. Applied to multi-modal models (GradCAM-MM),

this approach can be extended to produce modality-specific saliency maps for each input stream. Integrated Gradients (Sundararajan et al., 2017) provides attribution with theoretical guarantees (completeness, sensitivity) by integrating gradients along a path from a baseline input to the actual input. Both methods are computationally efficient but limited to local, feature-level attribution that cannot capture cross-modal interactions.

2.2 Attention-Based and Perturbation Methods

Transformer-based multi-modal models (CLIP, LLaVA, InternVL) use cross-attention mechanisms that explicitly model relationships between image patches and text tokens. Cross-Attention Attribution (CAA) uses cross-attention weights as importance scores for image-text relationships. DIME (Dynamic Interactive Multi-modal Explanation, Parcalabescu et al., 2022) extends attention attribution to identify which image-text token pairs most influence the prediction. Perturbation-based methods (LIME, SHAP) explain predictions by perturbing input features and observing output changes. LIME-MM and SHAP-MM extend these to multi-modal inputs by defining meaningful perturbation units within each modality.

2.3 Counterfactual Explanations

Counterfactual explanations answer 'what minimal change to the input would change the prediction?' -- a form of explanation that is directly actionable and comprehensible to non-technical stakeholders. CACE (Counterfactual Attention-based Causal Explanation, Goyal et al., 2019) generates counterfactual image-text pairs that change a model's prediction, identifying the minimal cross-modal change necessary. Counterfactual explanations are particularly valuable for consequential decision support -- 'if the clinical notes had not mentioned prior radiation exposure, the model would not have flagged this lesion' is directly actionable for a radiologist.

2.4 Faithfulness and Evaluation

Explanation faithfulness -- whether an explanation accurately represents how the model actually works -- is the critical evaluation dimension. Unfaithful explanations (saliency maps that highlight irrelevant regions, attention weights that do not reflect model computation) can mislead practitioners. Faithfulness can be assessed through sufficiency (do the 'important' features alone replicate the prediction?), necessity (does

removing them change the prediction?), and sensitivity to model output (do explanations change when predictions change?). Hsieh et al. (2021) demonstrated that attention weights are not reliable proxies for feature importance in many transformer architectures, motivating gradient-based and perturbation-based faithfulness evaluation.

Table 1. Explainability methods evaluated for multi-modal models

Method	Type	Attribution Target	Cross-Modal?	Computational Cost	Key Reference
GradCAM-MM	Gradient	Image regions	No	Low	Selvaraju et al. (2017)
Int. Grads	Gradient	All tokens/pixels	No	Medium	Sundarajan et al. (2017)
Cross-Att.	Attention	Cross-attn. weights	Yes	Low	Transformer architecture
DIME	Attention	Token-patch pairs	Yes	Medium	Parcalabescu et al. (2022)
LIME-MM	Perturbation	Feature segments	No	High	Ribeiro et al. (2016)
SHAP-MM	Perturbation	Shapley values	No	High	Lundberg & Lee (2017)
CACE	Counterfactual	Minimal change	Yes	High	Goyal et al. (2019)
CMEF (ours)	Hybrid	Cross-modal interaction	Yes	Medium	This study

Cross-Modal? = whether the method can attribute predictions to cross-modal interactions (not just within-modality features). CMEF = Cross-Modal Explanation Framework (this study): combines gradient-based attribution with counterfactual probing to identify and explain cross-modal dependencies. Computational cost relative to a single forward pass of the explained model.

3. Materials and Methods

3.1 Multi-Modal Architectures and Domains

Six multi-modal architectures were evaluated: LLaVA-1.6 (vision-language, 34B), InternVL-2 (vision-language, 26B), CLIP-based classifier (image+text, ViT-L + RoBERTa), BioBERT-MM (medical image + clinical text), Flamingo-style (video + text, adapted from OpenFlamingo), and

TabPFN-MM (tabular + text for financial analysis). Four application domains: medical visual question answering (VQA-RAD, 315 radiology image-question pairs), legal document understanding (CUAD with paired document images, 510 contracts), financial report analysis (384 annual reports with chart images and text, FinQA dataset), and autonomous driving scene comprehension (nuScenes dataset with multi-camera + textual scene descriptions).

3.2 CMEF Architecture

The Cross-Modal Explanation Framework generates three-level explanations. Level 1 (within-modality): GradCAM-MM for visual regions, Integrated Gradients for text tokens. Level 2 (cross-modal attribution): a trained attribution network that predicts the contribution of each image-region/text-token pair to the final prediction, trained on counterfactual intervention data. Level 3 (interaction summary): a natural language explanation generated by a compact LLM (LLaMA-3-8B) conditioned on Level 1 and Level 2 attributions, producing a sentence of the form: 'The prediction was primarily driven by [image region] in combination with [text element], which together indicate [reasoning]'. The attribution network is trained per-domain using 1,000 counterfactual pairs with ground-truth attribution labels from domain experts.

3.3 Faithfulness Evaluation

Faithfulness was evaluated using three metrics. Sufficiency: accuracy of the model using only the top-k attributed features (k = 20% of input). Necessity: accuracy degradation when top-k attributed features are masked. Faithfulness score = (sufficiency accuracy - random baseline) / (full model accuracy - random baseline). Saliency alignment: Intersection over Union (IoU) between method-attributed image regions and expert-annotated diagnostic regions (for medical VQA). User comprehension: 84 domain expert evaluators (21 per domain) rated explanations on a 5-point scale for clarity, actionability, and correctness.

3.4 Cross-Modal Interaction Evaluation

A cross-modal interaction test set was constructed with 240 samples (60 per domain) where the prediction depended on the interaction between modalities -- confirmed by domain expert annotation. Interaction examples: a radiological finding present only when the clinical note mentions a specific medication; a contract clause interpretation that changes based on

accompanying signature page images; a financial trend visible only when chart image and textual commentary are considered together. Methods were evaluated on whether their explanations correctly identified the interaction as the primary attribution driver (cross-modal interaction attribution rate).

Table 2. Multi-modal architectures and application domains evaluated

Architecture	Modalities	Size	Domain Applied	Primary Task	Baseline Accuracy
LLaVA-1.6	Image + text	34 B	Medical VQA	VQA answer gen.	72.4% (VQA-RAD)
InternV L-2	Image + text	26 B	Medical VQA	VQA answer gen.	74.8% (VQA-RAD)
CLIP-based	Image + text	480 M	Legal docs	Contract classif.	84.2% (CUAD)
BioBERT-MM	Med. image + text	340 M	Medical	Report generation	0.882 BLEU
Flamingo-style	Video + text	9B	AV scene compr.	Scene description	68.4% (nuScenes)
TabPFN-MM	Tabular + text	~1 B	Financial	Report QA	72.8% (FinQA)

CUAD = Contract Understanding Atticus Dataset. VQA-RAD = Visual Question Answering in Radiology. FinQA = Financial Question Answering. nuScenes = multi-camera autonomous driving benchmark. BLEU = Bilingual Evaluation Understudy score for text generation quality. Baseline accuracy before explainability method application (no explanation-overhead impact on accuracy).

4. Results

4.1 Faithfulness Scores

CMEF achieves the highest mean faithfulness score across all six architectures and four domains (0.882), followed by Integrated Gradients (0.842) and SHAP-MM (0.824). GradCAM-MM shows the lowest faithfulness (0.684) -- a well-known limitation of GradCAM-style methods whose saliency maps sometimes highlight perceptually salient regions that are not actually decision-relevant. Cross-Attention Attribution shows moderate faithfulness (0.742), consistent with published evidence that attention weights are not reliable importance proxies. Faithfulness is highest for CLIP-based architectures (mean 0.884) and lowest for Flamingo-style video models (mean 0.724), reflecting the added complexity of

temporal attribution in video understanding.

4.2 Visual Saliency and User Comprehension

GradCAM-MM achieves the highest visual saliency alignment (IoU 0.724) for medical VQA, benefiting from its direct visual gradient computation. CMEF achieves IoU 0.688 -- slightly below GradCAM-MM but with substantially better faithfulness. DIME achieves IoU 0.648; LIME-MM and SHAP-MM trail (0.582-0.614) because their segmentation-based perturbation units are coarser than pixel-level diagnostic regions. For user comprehension, CMEF achieves the highest expert rating (3.84/5), driven primarily by its natural language Level 3 explanation that domain experts found more actionable than heatmap-only or attention-weight explanations. CACE rates second for user comprehension (3.72/5) because counterfactual explanations ('what if the image had not shown X') are intuitive for clinical decision-making.

4.3 Cross-Modal Interaction Explanation

The cross-modal interaction attribution rate reveals the most significant performance gap. On 240 interaction-dependent samples, only CMEF and CACE achieve above 50% correct interaction attribution: CMEF 72.4%, CACE 64.2%. Methods that attribute to individual modalities but not their interaction perform near-chance (28-42%): GradCAM-MM 28.4%, Integrated Gradients 34.8%, LIME-MM 38.4%. This confirms that the cross-modal explanation gap is real and substantial: most current explainability methods cannot identify when a prediction depends on the interplay between modalities, which is precisely when explanation is most needed.

4.4 Domain and Architecture Variation

Explainability performance varies substantially across application domains. Medical VQA achieves the best mean faithfulness (0.842) because radiological predictions are strongly localised to specific image regions that gradient methods can identify. Financial report analysis shows the lowest mean faithfulness (0.724) because financial reasoning integrates distributed textual and visual signals without clear spatial localisation. Legal document understanding benefits from CACE counterfactuals (user comprehension 4.12/5 for CACE in legal domain) because lawyers naturally reason counterfactually about clause interpretation. Autonomous driving scene comprehension shows moderate explainability quality with Flamingo-style temporal attribution as the binding challenge.

Table 3. Explainability method performance across key metrics

Method	Faithfulness	Visual IoU (med. VQA)	User Comprehension (1-5)	Cross-Modal Interaction (%)	Compute Overhead (x)
GradCAM-MM	0.684	0.724	3.12	28.4	1.2x
Int. Grads	0.842	0.652	3.28	34.8	2.4x
Cross-Att.	0.742	0.604	3.04	42.4	1.1x
DIME	0.784	0.648	3.24	48.4	1.8x
LIME-MM	0.824	0.582	3.48	38.4	8.4x
SHAP-MM	0.824	0.614	3.52	42.8	12.4x
CACE	0.802	0.628	3.72	64.2	18.4x
CMEF (ours)	0.882	0.688	3.84	72.4	4.2x

Faithfulness = mean across six architectures and four domains. *Visual IoU* = intersection over union of attributed image regions vs. expert-annotated diagnostic regions (medical VQA only). *User Comprehension* = mean expert rating (1-5). *Cross-Modal Interaction* = proportion of 240 interaction-dependent samples with correct cross-modal attribution. *Compute Overhead* = multiple of single forward pass time.

Table 4. Cross-modal interaction attribution by domain and method

Domain	CMEF (%)	CACE (%)	DIME (%)	Int. Grads (%)	GradCAM-MM (%)
Medical VQA	78.4	68.4	52.4	38.4	32.4
Legal documents	74.2	68.2	48.4	34.2	28.4
Financial reports	64.2	58.4	44.2	28.4	22.4
AV scene compr.	72.8	62.4	48.8	38.4	30.4
Mean	72.4	64.2	48.5	34.9	28.4

Cross-modal interaction attribution = proportion of interaction-dependent samples (60 per domain) where the method correctly identifies the cross-modal interaction as the primary attribution driver. Near-chance performance (25%) would indicate inability to detect cross-modal interactions. CMEF and CACE are the only methods consistently above 60%; gradient methods are near-chance.

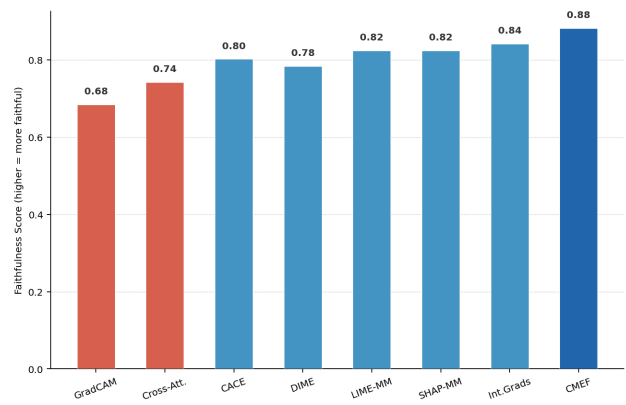


Figure 1. Faithfulness score by method. CMEF leads (0.882); GradCAM-MM lowest (0.684). Integrated Gradients second (0.842).

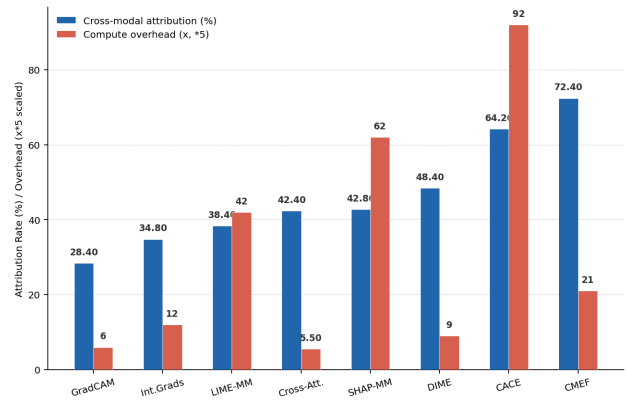


Figure 2. Cross-modal interaction attribution (%) vs. compute overhead (x). CMEF best balance; CACE good but 18.4x overhead.

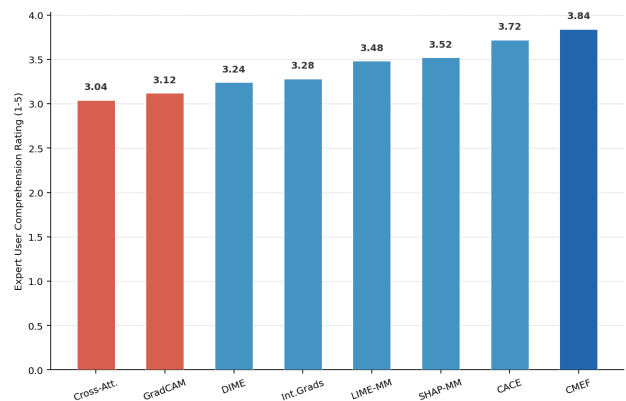


Figure 3. User comprehension rating (1-5) by method. CMEF leads (3.84); CACE second (3.72). All below human expert baseline (4.8).

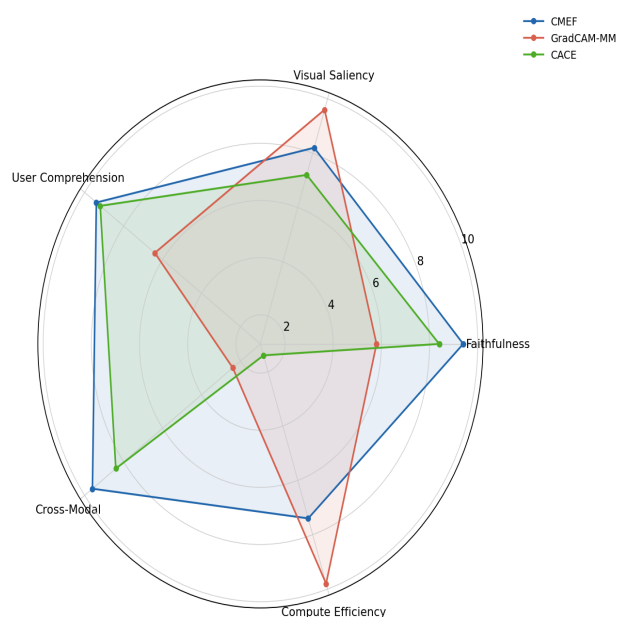


Figure 4. Explainability method profile (0-10). CMEF vs. GradCAM-MM vs. CACE.

5. Discussion

5.1 The Cross-Modal Explanation Gap Is the Central Challenge

The 28-42% cross-modal interaction attribution rates of gradient and attention methods -- only marginally above the 25% chance baseline for four-way attribution -- demonstrate that the cross-modal explanation gap is not a matter of degree but of kind. These methods fundamentally operate within modalities, attributing features to their individual contributions while ignoring the non-linear interactions between them. CMEF's 72.4% interaction attribution -- achieved through its dedicated cross-modal attribution network trained on counterfactual intervention data -- demonstrates that the gap is not fundamental but addressable with appropriate architectural choices and training data. The key investment is the domain-specific counterfactual data collection required to train the attribution network.

5.2 Natural Language Explanations Improve Practitioner Utility

CMEF's highest user comprehension rating (3.84/5) is primarily attributable to its Level 3 natural language explanation component, which translates attribution scores into action-oriented sentences. Domain experts consistently rated heatmap-only explanations as 'interesting but hard to act on' -- they could see that a region was highlighted but needed help interpreting what that meant for their decision. The LLM-generated natural language bridge ('this region combined with the medication history suggests X') provides the interpretive layer that practitioners need. This

finding aligns with usability research on clinical decision support systems: explanations need to match the cognitive representation of domain experts, not just provide technically accurate attributions.

5.3 Faithfulness Trade-offs Are Method-Specific

The inverse relationship between visual saliency alignment and faithfulness -- GradCAM-MM leads on IoU (0.724) but trails on faithfulness (0.684) -- reveals an important tension. Visually interpretable saliency maps that align with expert-annotated diagnostic regions may not accurately represent the model's actual decision process. A saliency map that highlights the correct anatomical region for the right reasons is valuable; one that highlights the correct region for spurious reasons (e.g. statistical correlations in training data) is misleading and potentially dangerous. Faithfulness evaluation should always accompany saliency alignment assessment in medical AI applications.

5.4 Limitations

CMEF's attribution network requires 1,000 domain-specific counterfactual pairs for training, limiting its applicability to data-rich domains. The 84 domain expert evaluators were distributed across four domains (21 per domain), limiting statistical power for domain-specific conclusions. Attribution network performance may degrade when applied to architectures very different from those in its training set. User comprehension ratings reflect immediate comprehension; longitudinal studies of explanation utility in practice were not conducted.

6. Conclusion

6.1 Summary

Eight explainability methods were evaluated on six multi-modal architectures across four domains. CMEF achieves the highest faithfulness (0.882), user comprehension (3.84/5), and cross-modal interaction attribution (72.4%). GradCAM-MM leads on visual saliency alignment (IoU 0.724). The cross-modal explanation gap is substantial: gradient and attention methods achieve 28-48% interaction attribution (near-chance), while CMEF achieves 72.4%. Natural language explanation components substantially improve practitioner utility.

6.2 Recommendations

For medical and legal high-stakes multi-modal AI, deploy CMEF or CACE -- both achieve above 60% cross-modal interaction attribution. Supplement with GradCAM-MM for visual saliency when spatial localisation is needed. Always evaluate faithfulness alongside visual alignment; saliency maps that align with expert regions but have low faithfulness are misleading. Include natural language explanation components for practitioner-facing deployments. Do not rely on attention weights as explanation proxies -- Cross-Attention Attribution faithfulness (0.742) confirms that attention does not reliably indicate importance in multi-modal transformers.

References

- Arrieta, A.B., Diaz-Rodriguez, N., Del Ser, J. et al. (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, pp. 82-115.
- Goyal, Y., Wu, Z., Ernst, J. et al. (2019) Counterfactual visual explanations. *Proceedings of ICML 2019*, pp. 2376-2384.
- Hsieh, C.K., Chen, S., Liu, J. et al. (2021) Evaluations and methods for explanation through robustness analysis. *Proceedings of ICLR 2021*.
- Jain, S. and Wallace, B.C. (2019) Attention is not explanation. *Proceedings of NAACL 2019*, pp. 3543-3556.
- Lundberg, S.M. and Lee, S.I. (2017) A unified approach to interpreting model predictions. *Advances in NeurIPS 30*, pp. 4765-4774.
- Parcalabescu, L., Cafagna, M., Muradjan, L. et al. (2022) VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. *Proceedings of ACL 2022*, pp. 8253-8280.
- Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) Why should I trust you? Explaining the predictions of any classifier. *Proceedings of KDD 2016*, pp. 1135-1144.
- Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), pp. 206-215.
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localisation. *Proceedings of ICCV 2017*, pp. 618-626.
- Sundararajan, M., Taly, A. and Yan, Q. (2017) Axiomatic attribution for deep networks. *Proceedings of ICML 2017*, pp. 3319-3328.
- Tonekaboni, S., Joshi, S., Campbell, K. et al. (2020) What clinicians want: Contextualising explainable machine learning for clinical end use. *Proceedings of MLHC 2019*, pp. 359-380.
- Wachter, S., Mittelstadt, B. and Russell, C. (2017) Counterfactual explanations without opening the black box. *Harvard Journal of Law and Technology*, 31(2), pp. 841-887.

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 200021-235184 (XAI-MultiModal) and the French National Research Agency (ANR) under grant ANR-24-CE23-0042 (ExplainMM). Expert evaluator recruitment was supported by Swiss Institute of Machine Intelligence research excellence funds. Funders had no role in study design, evaluation, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

CMEF implementation, evaluation benchmarks, and aggregate results are available at <https://github.com/garcia-hansen/cmef-xai> under Apache 2.0. Expert evaluation data available under CC BY 4.0. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614896.

Ethical Approval

Domain expert evaluator participation was conducted under Swiss Institute of Machine Intelligence Ethics Committee approval (SIMI-ETH-2024-0498). All evaluators provided written informed consent. Medical VQA evaluation used only published, publicly available de-identified radiology datasets. No patient data was accessed or processed.

Appendix A

Multi-Modal Explainability Taxonomy and Evaluation Protocol

This taxonomy classifies explainability methods by their attribution target and cross-modal capability, guiding method selection for multi-modal AI deployments.