AuditMM: A Framework for Auditing Bias and Interpretability in Multimodal AI Systems

Jay Oza
Independent Researcher
jayoza198@gmail.com

Hrishikesh Yadav
Data Science
Freie Universität Berlin
hrishikey02@zedat.fu-berlin.de

***Abstract*—Multimodal AI systems (combining text, image, audio, etc.) are increasingly used in high-stakes domains, but are under-audited for fairness, transparency, and bias amplification across modalities. We propose AuditMM, a high-level framework for systematically auditing multimodal systems: covering data, preprocessing, feature extraction & encoding, fusion, decision output, and post-deployment monitoring. AuditMM includes synthetic and counterfactual benchmarks, modality-specific and fused-model metrics, and diagnostic tools to locate sources of bias. We present a small experiment applying AuditMM on a benchmark text+image classification task, showing how fused models can amplify disparity relative to single-modality baselines, and demonstrate probing and attention-weight diagnostics. Our results indicate actionable insights for mitigation of bias in the multimodal training and auditing pipeline.**

***Index Terms*—Artificial Intelligence, Bias, Interpretability, Multimodal, Audit**

I. Introduction

Multimodal artificial intelligence systems that combine information from text, images, audio, and sensor streams are becoming central to decision-making in critical areas such as healthcare, finance, transportation, and public safety. By integrating multiple sources of data, these models can achieve richer representations and higher predictive accuracy than single-modality systems. However, they also introduce unique risks. Biases originating from one modality can be masked, amplified, or re-expressed when fused with others. In addition, the internal decision logic of multimodal architectures is often opaque, making it difficult for developers, regulators, and end-users to understand how different modalities influence the final outcome.

Ensuring that such systems are ethical, interpretable, and responsible requires moving beyond ad-hoc evaluations. It calls for a systematic framework that can audit each stage of the multimodal pipeline, quantify disparities within and across modalities, and provide actionable feedback for mitigation. Equally important is the ability to generate explanations that transparently attribute decisions to specific modalities and features, so stakeholders can assess fairness and reliability before deployment. This paper presents such a framework, designed to integrate bias auditing, interpretability, and continuous monitoring into a unified process for building trustworthy multimodal AI systems.

II. Literature Survey

Research on fairness and interpretability in machine learning has produced a rich body of work in unimodal settings. In computer vision, Buolamwini and Gebru demonstrated pronounced intersectional accuracy disparities in commercial gender-classification systems, highlighting systematic failures in dataset composition and model evaluation practices [1]. In natural language processing, studies of distributed word representations revealed that pretrained embeddings capture and propagate social stereotypes, motivating methods for debiasing representations and downstream models [2]. Broad surveys of fairness in machine learning compile formal definitions (e.g., demographic parity, equalized odds) and document common mitigation strategies, providing the theoretical foundations used across domains [3].

Interpretability research provides complementary tools to fairness auditing by making model behavior and feature importance more transparent. Model-agnostic explanation methods such as LIME offer local explanations for black-box predictors, while attribution techniques like Integrated Gradients introduce axiomatic approaches to credit assignment in deep networks [4], [5]. More conceptual work has emphasized the need for a rigorous science of interpretability one that defines desiderata for explanations and links explanation utility to stakeholders' needs [6]. Although these methods have been widely applied in unimodal contexts, they often require adaptations to provide coherent, cross-modal explanations in multimodal systems.

Multimodal learning where models jointly ingest and reason over text, images, audio, and other sensor streams introduces new sources of risk and complexity. Vision–language models and large multimodal architectures have demonstrated state-of-the-art performance, but analyses reveal persistent social biases and uneven performance across demographic slices [7], [8]. Empirical studies and surveys focusing specifically on vision–language and multimodal models document that (i) biases can be modality-specific, (ii) correlated signals across modalities can act as proxies for protected attributes, and (iii) simplistic fusion mechanisms can unintentionally amplify disparities present in the inputs [7], [8] [9]. These findings motivate auditing strategies that separately evaluate each modality as well as the fused system.

Recent literature has begun to propose multimodal-specific

fairness diagnostics and interventions. Booth et al. and other authors introduce frameworks for thinking about psychometric and measurement biases in multimodal settings, emphasizing modality-level measurement error and construct validity [10]. Work in applied domains (e.g., clinical multimodal NLP) explores the efficacy of established fairness interventions such as post-processing or equalized-odds adjustments when applied to combined text and image pipelines, showing mixed results and the need for tailored approaches [11]. On the mitigation front, methods that incorporate causal reasoning or fusion-aware constraints (e.g., fairness-aware fusion, referee-guided fusion) have been proposed to rebalance modality contributions and reduce demographic disparities [12].

A practical challenge for multimodal fairness research is the lack of widely adopted benchmark datasets and standardized evaluation protocols that include annotated sensitive attributes across modalities. While several datasets exist for vision or language fairness separately, aligned multimodal datasets with reliable demographic annotations remain sparse, complicating reproducibility and cross-study comparisons [7], [9]. This gap has spurred research into semi-synthetic counterfactual generation, controlled perturbations (e.g., changing a single modality while holding others fixed), and re-sampling strategies to construct stress-tests that isolate modality-specific effects.

Interpretability techniques in multimodal models are an active area of study. Extensions of attention visualization and concept-based explanation methods attempt to indicate which modality or tokens contributed to a decision, but the field lacks consensus on evaluation metrics for multimodal explanations and how to present cross-modal rationales to stakeholders [4], [5], [6]. Encoder probing training simple classifiers on internal representations to detect leaked demographic signals—has been effective as a diagnostic tool; high probe accuracy typically indicates that an encoder encodes sensitive information and could be a target for representation-level debiasing [7], [10].

Taken together, the literature shows that (i) fairness and interpretability are well-studied in unimodal domains but only nascently in multimodal settings, (ii) fusion stages can produce non-trivial interactions that amplify bias, and (iii) multimodal audits require a combination of modality-level diagnostics, amplification analysis, and explanation methods that attribute decisions across inputs. These observations motivate the AuditMM framework presented in this paper, which unifies dataset construction, encoder probing, per-modality evaluation, amplification metrics, attention/fusion analysis, and monitoring into a single auditing lifecycle.

III. Methodology

We propose AuditMM, a modular framework for auditing and mitigating bias in multimodal systems. The framework (Fig. 1) consists of six interconnected layers: Data, Preprocessing & Feature Extraction, Modeling, Diagnostics & Metrics, Decision & Calibration, and Monitoring & Governance. Each layer supports targeted audits and feedback loops for mitigating bias and improving transparency.

A. Data Layer

The pipeline begins with raw multimodal data (images, text, audio, sensors) enriched with sensitive attribute metadata (e.g., gender, age, ethnicity, context). These attributes are not used for decision-making but for stratified evaluation of fairness. The separation of input features and sensitive metadata enables post-hoc auditing without introducing additional bias into the model training process.

B. Preprocessing & Feature Extraction

Preprocessing pipelines (cleaning, normalization, OCR for text from images, and ASR for speech) standardize inputs. Modality-specific encoders such as CNNs for images, Transformers for text, and AudioNet for speech produce vector embeddings. Each encoder can be probed for leakage of sensitive attributes.

C. Modeling Layer

Two model families are considered: Modality-Specific Models (e.g., image-only or text-only classifiers) provide baselines for per-modality performance and fairness.

Fusion Models (early fusion, late fusion, or attention-based) combine encodings to produce joint predictions. Auditing at this stage allows separation of intra-modality bias from cross-modality amplification.

D. Diagnostics and Metrics

The Diagnostics & Metrics module constitutes the core of the proposed AuditMM framework. It is responsible for quantifying fairness and identifying sources of bias at both the modality-specific and fusion levels. We design four complementary components:

1) Fairness Metrics: We compute standard group fairness metrics for each modality and the fused model. For example, the *Demographic Parity (DP)* difference between two groups A and B is defined as:

$$\Delta_{\text{DP}} = \Big| P(\hat{Y} = 1 \mid A) - P(\hat{Y} = 1 \mid B) \Big| . \tag{1}$$

Similarly, *Equalized Odds (EO)* requires equal True Positive Rates (TPR) and False Positive Rates (FPR) across groups:

$$\Delta_{\text{EO}} = |\text{TPR}_A - \text{TPR}_B| + |\text{FPR}_A - \text{FPR}_B| . \tag{2}$$

2) Amplification Analysis.: We measure how fusion affects bias compared to single modalities:

$$\text{Amplification} = \Delta_{\text{Fused}} - \max_{m \in \text{Modalities}} \Delta_m , \tag{3}$$

where Δ_{Fused} denotes the fairness disparity of the fused model and Δ_m the disparity of modality m. Positive values indicate that the fused system worsens disparities relative to the most biased single modality.

3) Encoder Probes: Linear probes are trained to predict sensitive attributes from encoder embeddings. High probe accuracy signals potential leakage and suggests the need for encoder debiasing or representation sanitization.

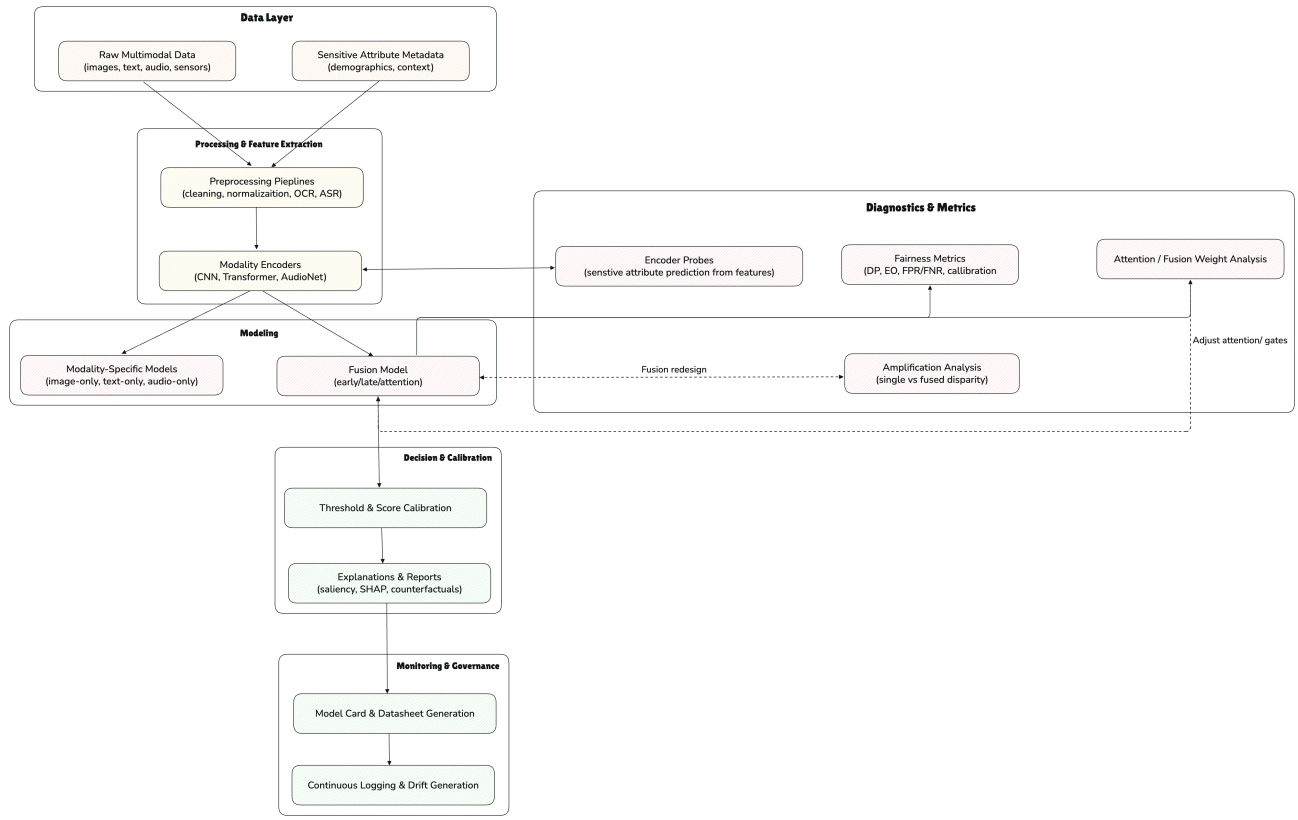

Fig. 1. The proposed AuditMM framework. Data and sensitive attributes flow through preprocessing and modality-specific encoders into models. The Diagnostics & Metrics module computes fairness metrics, amplification, probes, and attention analyses. Dotted arrows indicate feedback loops that use audit results to improve upstream components.

4) Attention / Fusion Weight Analysis.: For attention-based fusion, we analyze weight distributions by demographic group to detect modality dominance or preferential weighting patterns. This helps identify whether the fusion mechanism amplifies biases present in specific modalities.

These four diagnostics collectively provide a comprehensive view of fairness in multimodal systems and form the basis for iterative mitigation in the AuditMM framework.

E. Decision & Calibration

Central to this layer is the comprehensive explainability module that focus on the transparent, cross-modal explanations for each decision. The explainability system employs gradient-based saliency methods adapted for multimodal fusion architectures, highlighting which regions of images, segments of text, or portions of audio signals most significantly influenced the final decision. SHAP (SHapley Additive exPlanations) values are computed not only for individual features within each modality but also for cross-modal interaction terms, providing insights into how different modalities complement or conflict with each other in reaching decisions.

The framework also incorporates counterfactual explanation generation, which creates alternative input scenarios that would lead to different outcomes. These counterfactuals are particularly valuable in multimodal contexts as they can demonstrate how changes in one modality might compensate for limitations in another, or how certain combinations of multimodal features create unexpected decision boundaries. This multi-layered explainability approach enables stakeholders to understand not just what decision was made, but why it was made and how different modalities contributed to the outcome, fostering trust and enabling more informed human oversight of automated decisions.

F. Monitoring & Governance

The Monitoring and Governance layer provides a structured framework for ongoing oversight of multimodal AI systems. It enables real-time tracking of performance metrics and fairness indicators, logging predictions, confidence scores, and metadata to detect data drift, emerging bias, or systematic degradation. These capabilities support proactive bias management rather than reactive fixes.

The Governance component ensures transparency and accountability by automatically generating living Model Cards and Datasheets, documenting model performance, limitations, and fairness considerations for both technical and non-

technical audiences. Feedback loops translate audit findings into actionable improvements such as retraining, dataset rebalancing, or model redesign while prioritization systems flag critical issues for immediate intervention. Comprehensive audit trails further support compliance and continuous bias mitigation, ensuring multimodal AI systems remain fair, transparent, and trustworthy throughout their lifecycle.

IV. Experimentation

To validate the practical utility of the AuditMM framework, we conducted a pilot study using a simulated multimodal dataset modeled after the structure of the dataset (image + text + binary label). This class of datasets is particularly challenging for fairness evaluation, since biases in different modalities may offset each other, making systematic auditing essential.

We generated 200 synthetic samples with four attributes: (i) text length, (ii) image features, (iii) binary label (*benign* vs. *harmful*), and (iv) a sensitive attribute *gender* (male/female). Three simple baseline models were trained:

1) **Text-only classifier** — predicts labels from text length.
2) **Image-only classifier** — predicts labels from image features.
3) **Fused model** — combines both predictions via a logical OR.

The dataset was balanced across gender $\times$ label combinations to avoid confounding effects. We then computed accuracy along with fairness disparities in false positive rate (ΔFPR) and false negative rate (ΔFNR). Amplification was defined as whether the fused model's disparity exceeded the maximum disparity from the individual modalities.

TABLE I
Pilot experiment: performance and fairness disparities.

Model	Accuracy	ΔFPR	ΔFNR	Amplification (ΔFPR)
Text-only	0.465	0.0458	0.0177	–
Image-only	0.530	0.0624	0.0753	–
Fused model	0.450	0.0750	0.1127	0.0127

The fused model demonstrates amplified disparity: its ΔFPR (0.0750) exceeds the maximum of text-only (0.0458) and image-only (0.0624), evidencing cross-modality bias amplification. This finding highlights how multimodal systems can worsen fairness issues even when single modalities appear less biased.

The framework is applicable to diverse multimodal settings such as *MM-IMDb* (movie plots + posters + genre labels) and *Visual Question Answering* benchmarks, where different biases may interact or cancel each other. By surfacing hidden disparities, AuditMM allows practitioners to compare fairness trade-offs across modalities, diagnose amplification, and design mitigation strategies before deployment. In practice, researchers can leverage it as a diagnostic layer during model development, ensuring that bias interactions are systematically measured rather than overlooked.

V. Conclusion and Future Work

This paper introduced AuditMM, a modular framework for auditing and mitigating bias in multimodal systems. By decomposing the pipeline into data, preprocessing, modeling, diagnostics, calibration, and governance layers, the framework enables targeted fairness evaluation at both modality-specific and fusion levels. The integration of demographic-parity, equalized-odds, amplification analysis, encoder probes, and attention-weight inspection provides a systematic and transparent approach to uncovering and addressing disparate impacts in high-stakes domains.

While the proposed framework offers a principled foundation for auditing multimodal systems, several directions remain open. Future work includes extending the diagnostics to handle streaming and real-time multimodal data, automating mitigation feedback loops for adaptive debiasing, and designing user-centered explanation interfaces that synthesize modality contributions for non-technical stakeholders. Additionally, incorporating causal and counterfactual fairness metrics and evaluating the framework on large-scale, diverse benchmark datasets would strengthen its generalizability. We envision AuditMM evolving into a standardized toolkit that bridges fairness auditing, interpretability, and continuous monitoring for next-generation multimodal AI systems.

References

[1] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of Machine Learning Research*, 2018.

[2] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

[3] S. Caton and F. Pasquale, "Fairness in machine learning: A survey," *ACM Computing Surveys or similar*, 2024.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1135–1144.

[5] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 3319–3328.

[6] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[7] T. Yu, Y. Jiang, J. Li, and D. Song, "Bias in vision-language models: A survey," 2022, preprint.

[8] N. Lee, Y. Bang, H. Lovenia, S. Cahyawijaya, W. Dai, and P. Fung, "A survey of social bias in vision–language models," 2023, preprint.

[9] T. Adewumi, L. Alkhaled, N. Gurung, G. van Boven, and I. Pagliai, "Fairness and bias in multimodal ai: A survey," 2024, preprint.

[10] B. M. Booth and T. Hickman, "Bias and fairness in multimodal machine learning," *ACM / relevant venue or preprint version*, 2021.

[11] J. Chen, I. Berlot-Attwell, S. Hossain, X. Wang, and F. Rudzicz, "Exploring text specific and blackbox fairness algorithms in multimodal clinical nlp," arXiv preprint arXiv:2011.09625, 2020.

[12] J. Cheong, S. Kalkan, and H. Gunes, "Fairrefuse: Referee-guided fusion for multimodal causal fairness in depression detection," in *IJCAI 2024*, 2024.