

Quality and Reliability Auditing of Multimodal AI Systems: Explanation Faithfulness, Failure Prediction, and Calibration in Deepfake Detection

Md Anas Biswas

School of Computing, University of Portsmouth
Portsmouth, PO1 2UP, United Kingdom
up2082724@myport.ac.uk

Abstract

As multimodal AI enters high-stakes settings like forensics and moderation, reliability, not benchmark accuracy alone, determines whether outputs can be trusted. Three properties are decisive: whether a model's confidence is calibrated, whether its failures can be anticipated, and whether its explanations reflect the evidence it uses. We audit a representative system, audio-visual deepfake detection, on LAV-DF, which provides ground-truth temporal annotations of every manipulation. Using two in-domain detectors, visual and audio, we examine explanation faithfulness, failure prediction, and calibration. The central finding: classification accuracy and reliability are decoupled. The audio detector is the stronger classifier (AUROC 0.999) yet the worst calibrated and least faithfully explained: most overconfident (Expected Calibration Error 0.247), with integrated-gradients explanations barely localizing the manipulation (1.14 times chance) despite its decisions localizing strongly (7.38 times), showing the attribution method, not the model, fails. The weaker visual detector (AUROC 0.667) is more faithfully explained and readily recalibrated. For failure prediction, confidence is the reliable signal (AUROC 0.85): abstaining on the least-confident half cuts the error rate from 22% to under 4%, whereas cross-modal explanation disagreement performs no better than chance. Benchmark accuracy is therefore a poor proxy for deployable reliability, and reliability must be audited directly.

Keywords

Reliability auditing; model calibration; failure prediction; explanation faithfulness; deepfake detection.

1. Introduction

Generative models can now synthesize audio-visual media in which a subject's lip movements, facial appearance, or voice have been altered so convincingly that the manipulation is imperceptible to human observers (Khalid et al. 2021; Cai et al. 2022). Because such media enables fraud, disinformation, and non-consensual content, automatic deepfake detection has become an active research area, and reported accuracy on standard benchmarks has improved rapidly (Cai et al. 2023). For research benchmarks this is encouraging; for deployment it is insufficient. A detector that is occasionally and unpredictably wrong, that reports high confidence on the cases it gets wrong, or whose explanations indicate the wrong evidence, cannot be relied upon by a human analyst or an automated moderation pipeline operating under legal, journalistic, or platform-governance constraints.

What matters for deployment is the quality and reliability of a model's outputs: the calibration of its confidence, the faithfulness of its explanations, and the predictability of its failures. These reliability properties are studied for unimodal image classifiers (Guo et al. 2017; Hendrycks and Gimpel 2017) but remain comparatively unexamined for multimodal systems such as audio-visual deepfake detectors, whose two-stream structure raises questions absent in the unimodal case. When a clip's audio and video streams can each be independently authentic or forged, does a fused detector's confidence remain calibrated? Do gradient-based explanations recover where in time a manipulation occurs when manipulations are brief and localized? And when the detector errs, can a cross-modal signal, such as a

disagreement between the two streams about whether or where the evidence lies, anticipate the error better than the detector's own confidence?

This paper answers these questions through a systematic quality and reliability audit of audio-visual deepfake detection, a representative multimodal AI system, on the LAV-DF dataset (Cai et al. 2022). We deliberately use in-domain detectors rather than off-the-shelf pretrained models because, as we show, off-the-shelf detectors fail to transfer to LAV-DF under simple preprocessing, a manifestation of the well-documented cross-dataset generalization gap (Rössler et al. 2019; Li et al. 2020). LAV-DF is especially suited to this audit because it annotates the ground-truth temporal location of every manipulation, enabling a quantitative measurement of explanation faithfulness against a known answer rather than qualitative saliency inspection.

1.1 Objectives

The objectives of this study are: (i) to construct reliable in-domain audio and visual deepfake detectors on a temporally annotated dataset and to characterize a localized-manipulation scoring pitfall and a cross-dataset transfer failure that condition such an audit; (ii) to measure explanation faithfulness against ground-truth manipulation timing, distinguishing failures of the model from failures of the attribution method; (iii) to benchmark reliability signals, including a cross-modal explanation-disagreement signal, for predicting detector failures, and to quantify the operational value of confidence-based selective prediction; and (iv) to assess per-stream and post-fusion calibration. The unifying contribution articulated in the conclusion is empirical evidence that classification accuracy and reliability are decoupled in this multimodal setting, so that reliability properties must be audited directly rather than inferred from benchmark accuracy.

2. Literature Review

Research on deepfake detection has progressed along several largely separate lines: dataset construction, cross-dataset generalization, calibration and uncertainty, explanation faithfulness, and failure prediction. We review each and position the present work, which concerns the reliability of detectors that are already accurate rather than the pursuit of accuracy itself.

Datasets and detection. Early work targeted purely visual face manipulations, supported by FaceForensics++ (Rössler et al. 2019), Celeb-DF (Li et al. 2020), and the DeepFake Detection Challenge dataset (Dolhansky et al. 2020). The audio-visual setting adds a second manipulable modality and the possibility of cross-modal inconsistency. FakeAVCeleb (Khalid et al. 2021) pairs face-swapped or reenacted video with synthesized speech. LAV-DF (Cai et al. 2022) and its journal extension (Cai et al. 2023) instead alter short, strategically chosen segments using lip-sync reenactment and text-to-speech, making LAV-DF the first dataset built for temporal forgery localization; AV-Deepfake1M (Cai et al. 2024) scales this paradigm, and localization methods such as UMMAFormer (Zhang et al. 2023) have been benchmarked on it. We use LAV-DF not to advance localization but as a testbed whose temporal annotations enable an audit of detector reliability.

Audio-visual fusion and the multimodal threat. Multimodal detectors combine evidence from both streams, typically by score-level or feature-level fusion, on the assumption that a forgery in either modality, or an inconsistency between them, betrays manipulation. This assumption underlies the audio-visual datasets above (Khalid et al. 2021; Cai et al. 2022) and the temporal localization methods built on them (Zhang et al. 2023; Cai et al. 2024). Fusion, however, introduces reliability questions that do not arise for a single stream: a confident and accurate modality can dominate an uncertain one and drive the fused decision, the calibration of the fused score need not follow from that of its components, and an explanation must now be reconciled across two time axes. These properties of the fused system, rather than the fusion mechanism itself, are what the present audit examines.

Cross-dataset generalization. A recurring limitation is poor generalization across datasets and manipulation families: a model trained on one forgery type often performs near chance on another (Rössler et al. 2019; Li et al. 2020). The same fragility is reported for audio anti-spoofing across corpora (Müller et al. 2022). This effect is central to our methodology and motivates in-domain training.

Calibration and uncertainty. Modern neural networks are systematically overconfident, with softmax outputs exceeding empirical accuracy (Guo et al. 2017). Calibration is commonly quantified by the Expected Calibration Error (Naeini et al. 2015) and improved post hoc by temperature scaling (Guo et al. 2017); uncertainty can be estimated via deep ensembles (Lakshminarayanan et al. 2017) or Monte Carlo dropout (Gal and Ghahramani 2016). Within deepfake detection, point predictions alone are operationally insufficient when calibration degrades under cross-dataset shift,

so proper scoring rules and calibration error capture reliability that accuracy does not. Our audit contributes a multimodal, per-stream and post-fusion calibration analysis.

Explanation faithfulness. Attribution methods such as Grad-CAM (Selvaraju et al. 2017), LIME (Ribeiro et al. 2016), SHAP (Lundberg and Lee 2017), and integrated gradients (Sundararajan et al. 2017) indicate which inputs drive a decision, but a persistent concern is faithfulness: an explanation should reflect the model's actual evidence rather than a plausible but disconnected narrative (Jacovi and Goldberg 2020). Faithfulness is hard to evaluate without ground truth; common proxies overlap saliency with a manipulation mask, perturb inputs and measure the effect (Hooker et al. 2019; Petsiuk et al. 2018), or apply sanity checks (Adebayo et al. 2018). We supply a temporal analogue of mask-overlap faithfulness on LAV-DF.

Failure prediction. Predicting when a classifier is wrong underlies selective classification, in which a model abstains on likely errors (Geifman and El-Yaniv 2017), and misclassification detection, whose canonical baseline is the maximum softmax probability (Hendrycks and Gimpel 2017). Because confidence is miscalibrated, recent work re-examines the link between calibration and failure prediction (Zhu et al. 2022). We ask whether cross-modal disagreement, a signal unavailable in the unimodal setting, can outperform confidence at predicting the failures of a fused detector. In summary, prior work optimizes or evaluates accuracy and localization; the reliability of already-accurate audio-visual detectors, and the relationship among calibration, faithfulness, and failure prediction, remains largely unexamined, which is the gap this study addresses.

3. Methods

Detectors. Both detectors are EfficientNet-B0 backbones (Tan and Le 2019) initialized from ImageNet weights (Deng et al. 2009) and fine-tuned for binary real/fake classification with the Adam optimizer (learning rate $3e-4$) for five epochs. The visual detector operates on 224×224 face crops, with forged crops sampled from inside the fake periods of video-fake clips so that forged-labeled frames genuinely contain manipulated pixels. The audio detector operates on 128-band log-mel spectrograms of two-second windows. Because manipulations are localized in time, clip-level scores are obtained by taking the maximum forgery score over per-segment windows rather than averaging over the whole clip. Each stream is trained on a balanced set of 1,500 authentic and 1,500 forged examples drawn from the identity-disjoint training split, with faces located by a Haar cascade detector, and reaches high validation accuracy (0.887 for the visual stream and 0.982 for the audio stream). A fused detector is formed by averaging the two streams' forgery probabilities and thresholding at 0.5.

Explanation faithfulness. Each detector is explained with integrated gradients (Sundararajan et al. 2017), reduced to a temporal saliency curve so that audio and visual explanations share a common time axis. LAV-DF's fake periods provide a ground-truth temporal mask, enabling a quantitative faithfulness ratio (FR): the fraction of saliency mass falling inside the manipulated period, normalized by the fraction of the clip that the period occupies,

$$FR = (\text{saliency mass in fake period} / \text{total saliency mass}) / (\text{fake-period length} / \text{clip length}).$$

$FR = 1$ indicates saliency distributed at chance with respect to the manipulation; $FR > 1$ indicates localization beyond chance. We compute FR for both the integrated-gradients attribution (attribution localization) and the detector's own per-segment score curve (decision localization); comparing the two separates failures of the attribution method from failures of the model.

Failure prediction. For each clip we compute the fused forgery probability and candidate reliability signals: the maximum softmax probability (MSP), predictive entropy, and boundary margin (the confidence family); cross-modal prediction disagreement and label disagreement; and cross-modal explanation disagreement (CMED), defined from the Spearman rank correlation of the two time-aligned saliency curves so that CMED is large when the modalities localize evidence at different times:

$$CMED = 1 - \rho_s(s_{\text{visual}}, s_{\text{audio}}),$$

where ρ_s is the Spearman correlation and $s_{\text{visual}}, s_{\text{audio}}$ are the per-segment saliency curves. Each signal is evaluated by the area under the ROC curve (AUROC) for predicting the binary error outcome, with 95% confidence intervals from 1,000 bootstrap resamples. We additionally summarize selective prediction by ranking clips by a signal and abstaining on the least confident, reporting the Area Under the Risk-Coverage curve (AURC), the average selective risk over all coverage levels, where lower is better (Geifman and El-Yaniv 2017).

Calibration. We quantify calibration with the Expected Calibration Error (Naeini et al. 2015), which partitions the N predictions into $B = 10$ equal-width probability bins and averages, weighted by bin occupancy, the gap between each bin's mean predicted probability and its empirical forgery frequency:

$$ECE = \sum_b (n_b / N) |acc(b) - conf(b)|, \quad b = 1, \dots, B,$$

where bin b holds n_b of the N predictions, $acc(b)$ is the empirical forgery rate in the bin, and $conf(b)$ is the mean predicted probability. We complement ECE with the Brier score (the mean squared error of the predicted probabilities) and the maximum calibration error (MCE, the largest single-bin gap), and report all three for the visual stream, the audio stream, and the fused score, before and after post-hoc temperature scaling (Guo et al. 2017), with the optimal temperature T ($T > 1$ indicates overconfidence, $T < 1$ underconfidence). Each stream is evaluated against the label it predicts: the visual stream against the video-manipulation flag, the audio stream against the audio-manipulation flag, and the fused score against the clip-level any-fake label (video or audio manipulated).

4. Data Collection

Dataset. LAV-DF (Cai et al. 2022) comprises 136,304 talking-head clips of 153 subjects sourced from VoxCeleb2, of which 36,431 are authentic and 99,873 contain localized manipulations, with identity-disjoint training, development, and test splits of 78,703, 31,501, and 26,100 clips. Each clip is annotated with the precise start and end times of every manipulation and with flags indicating whether the video and audio streams were modified. Because the streams are manipulated independently, every clip falls into one of four cells denoted by a two-letter video/audio code: RVRA (real video, real audio; the only authentic cell), FVRA (fake video, real audio), RVFA (real video, fake audio), and FVFA (both fake). Manipulations are brief: across the 356 forged clips in our evaluation subset they occupy a median of 8.2% of clip duration (interquartile range 5.4-11.8%). We draw a class-balanced evaluation subset of 500 clips from the test split; 473 yield usable face crops and constitute the evaluation set, with all detectors trained only on the disjoint training split.

Off-the-shelf detectors do not transfer. An audit is meaningful only if the detector discriminates on the data. Four publicly available pretrained detectors evaluated on the subset under our preprocessing, a FaceForensics++-trained Vision Transformer, a CLIP-based cross-dataset detector, a Celeb-DF-trained Vision Transformer, and the AASIST audio anti-spoofing model (Jung et al. 2022), all failed to discriminate: the visual models showed no separation between authentic and forged clips, the Celeb-DF model's ordering was inverted, and AASIST achieved a real-versus-fake-audio AUROC of 0.484, indistinguishable from chance (Table 1). These outcomes are consistent with the cross-dataset generalization gap and motivate in-domain training.

Table 1. Off-the-shelf pretrained detectors evaluated on the LAV-DF subset under our preprocessing. None discriminates between authentic and forged clips (chance corresponds to AUROC 0.50).

Pretrained detector	Modality	Outcome on LAV-DF subset
Vision Transformer (FaceForensics++)	Visual	No separation (AUROC $\approx$ 0.50)
CLIP-based cross-dataset detector	Visual	No separation (AUROC $\approx$ 0.50)
Vision Transformer (Celeb-DF)	Visual	Inverted ordering (AUROC $<$ 0.50)
AASIST anti-spoofing (Jung et al. 2022)	Audio	At chance (AUROC 0.484)

Localized-manipulation scoring pitfall. Because manipulations are localized, the scoring protocol materially affects measured performance. Scoring the in-domain audio detector on the full-clip mel-spectrogram yields an AUROC of only 0.612, whereas scoring two-second windows at a one-second hop and taking the maximum raises it to 0.999 on the same clips and detector. The 0.39 gap is attributable entirely to the protocol. We therefore adopt maximum-over-window scoring throughout and flag this as a methodological hazard: evaluations on localized-manipulation datasets that use uniform or full-clip scoring may substantially underestimate detector performance and misrank detectors.

5. Results and Discussion

5.1 Numerical Results

Table 2 reports detector performance. The audio detector is near-perfect (AUROC 0.999), the visual detector is, by design, far weaker (0.667), and the mean-fused detector reaches 0.933 with an overall error rate of 0.224 (106 of 473

clips) at the 0.5 threshold. The fused detector attains high precision (0.992) but lower recall (0.708), reflecting that its errors are concentrated in clips the weak visual stream cannot resolve.

Table 2. In-domain detector performance on the LAV-DF evaluation subset (AUROC with 95% bootstrap confidence interval; accuracy, precision, recall, and F1 at threshold 0.5).

Detector	AUROC (95% CI)	Acc	P	R	F1
Visual (face crops)	0.667 [.618, .717]	0.603	0.575	0.851	0.687
Audio (mel-spectrogram)	0.999 [.998, 1.00]	0.981	0.967	0.996	0.981
Fused (mean probability)	0.933 [.910, .953]	0.776	0.992	0.708	0.826

Table 3 reports explanation faithfulness. Both detectors' decisions localize the manipulation strongly (visual 3.86 times chance, audio 7.38 times). The visual attribution is also faithful (3.39 times), but the audio attribution barely separates from chance (1.14 times on the full clip, 0.95 times under windowed scoring), far below its own decision (7.38 times). The 95% intervals confirm the dissociation: the audio attribution ratio does not overlap either the audio decision or the visual attribution. Thus the audio detector clearly determines where the forgery is, but integrated gradients fails to surface it; the unfaithfulness lies with the attribution method, not the model.

Table 3. Explanation faithfulness ratio (saliency mass inside the ground-truth fake period, normalized by chance) with 95% bootstrap confidence intervals. Values above 1 indicate localization beyond chance.

Signal	FR	95% CI	> chance
Visual decision	3.86	[3.43, 4.37]	77.0%
Visual attribution (IG)	3.39	[2.91, 3.95]	60.4%
Audio decision	7.38	[6.95, 7.83]	100%
Audio attribution (IG)	1.14	[1.11, 1.18]	69.4%

Table 4 reports failure-prediction AUROC. The confidence family (MSP, entropy, and margin, which coincide for a binary classifier) dominates at 0.853. Cross-modal prediction disagreement matches it (0.848) but adds nothing beyond confidence, because stream disagreement is precisely what drives the fused score toward the decision boundary. Cross-modal explanation disagreement (CMED) is at chance (0.519); its interval excludes that of confidence by a wide margin. Errors are concentrated almost entirely in the FVRA cell (error rate 0.843), where the confident audio stream out-votes the uncertain visual stream and the average lands just below threshold; these are near-boundary, low-confidence predictions that confidence captures directly and that CMED, being flat across cells, cannot.

Table 4. Failure-prediction AUROC of each reliability signal with 95% bootstrap confidence intervals (chance = 0.50; fused error rate = 0.224).

Reliability signal	AUROC	95% CI
Confidence (MSP = entropy = margin)	0.853	[.818, .891]
Cross-modal prediction disagreement	0.848	[.806, .888]
Cross-modal explanation disagreement (CMED)	0.519	[.455, .585]

Table 5 reports calibration. All three scores are miscalibrated, and the pattern again contradicts the accuracy ordering. The most accurate (audio) detector is the worst calibrated, with the highest ECE (0.247), Brier score (0.242), and maximum calibration error (0.876), and an optimal temperature at the search ceiling (5.0); temperature scaling barely helps it. The visual detector is overconfident but readily recalibrated (ECE 0.135 to 0.087). The fused score is, in contrast, underconfident ($T = 0.80$) despite both component streams being overconfident, because averaging two near-binary scores pulls the fused probability toward the middle; fusion even yields the lowest Brier (0.125). Calibration therefore does not compose under fusion and must be assessed on the deployed system.

Table 5. Calibration of the two streams and the fused score: ECE (with 95% bootstrap CI), ECE after temperature scaling (TS), Brier score, maximum calibration error (MCE), and optimal temperature T.

Detector	ECE (95% CI)	ECE-TS	Brier	MCE	Opt. T
Visual	0.135 [.113, .179]	0.087	0.169	0.546	2.60
Audio	0.247 [.208, .287]	0.236	0.242	0.876	5.00
Fused	0.135 [.112, .168]	0.134	0.125	0.447	0.80

5.2 Graphical Results

Figure 1 visualizes the faithfulness asymmetry. Both decisions localize the manipulation well above chance, and the visual attribution does too, but the audio attribution sits near the chance line, confirming that an accurate model can be paired with an unfaithful gradient explainer. Figure 2 shows a representative clip: the visual saliency concentrates near the manipulated region, whereas the audio saliency is diffuse across the clip despite a correct classification.

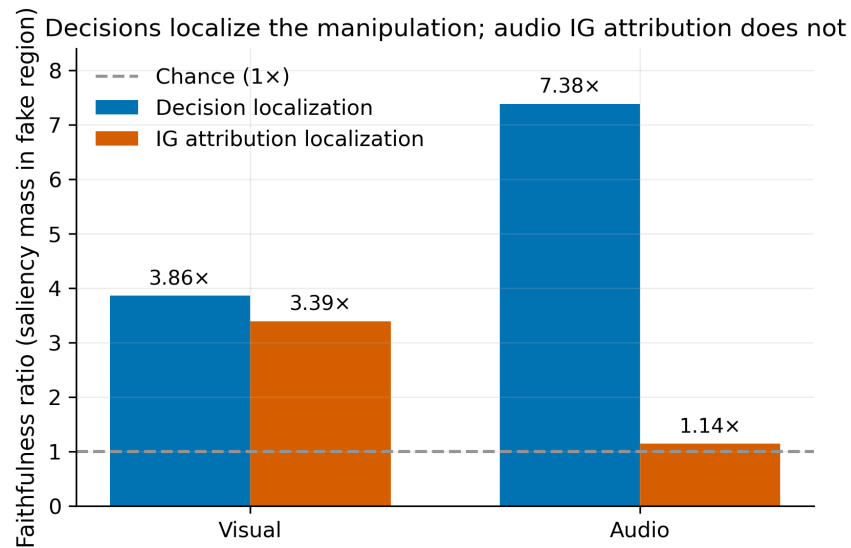


Figure 1. Explanation faithfulness, decision versus attribution, per modality. Bars give the faithfulness ratio; the dashed line marks chance. The audio attribution is unfaithful despite a strongly faithful audio decision.

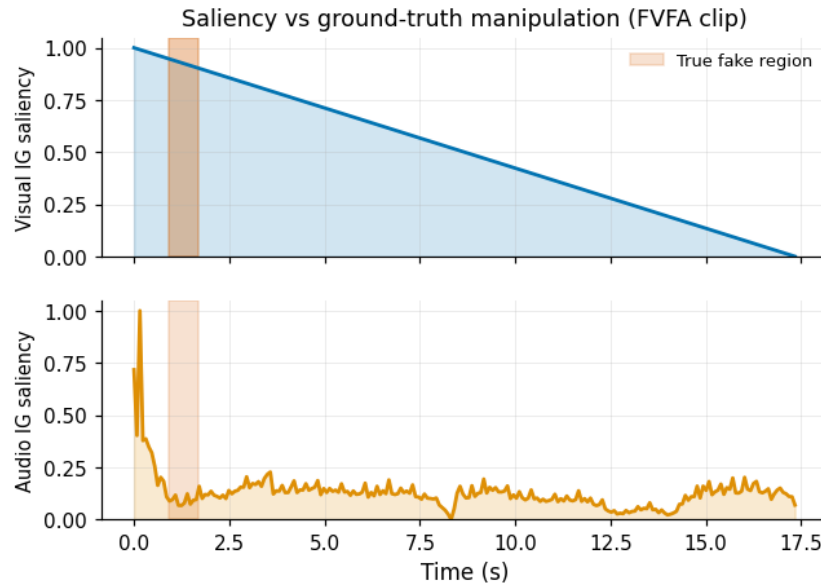


Figure 2. Qualitative example for one clip. Visual saliency (top) concentrates near the true manipulated region (shaded); audio saliency (bottom) is diffuse across the clip despite the audio detector classifying it correctly.

Figure 3 makes the failure structure visible. The combined-detector errors concentrate almost entirely in the FVRA cell, where the confident audio stream out-votes the uncertain visual stream, while cross-modal explanation disagreement is flat across all four cells. This is precisely why prediction confidence predicts failures and explanation disagreement does not.

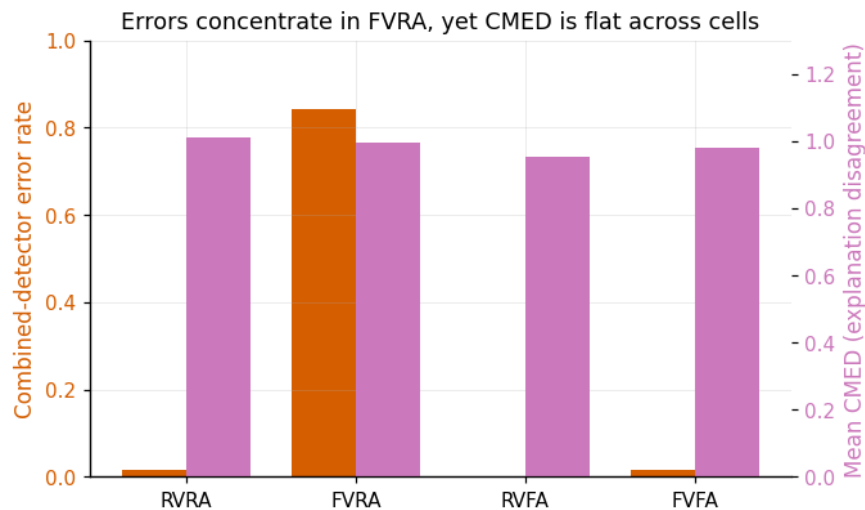


Figure 3. Per-cell combined-detector error rate (left axis) versus mean cross-modal explanation disagreement (right axis). Errors concentrate in the FVRA cell while disagreement is flat across cells.

Figure 4 shows the operational consequence for failure prediction as a risk-coverage curve. Ranking clips by confidence (MSP) and abstaining on the least-confident sharply reduces the selective risk, whereas ranking by cross-modal explanation disagreement hugs the no-skill base error rate and never lowers it.

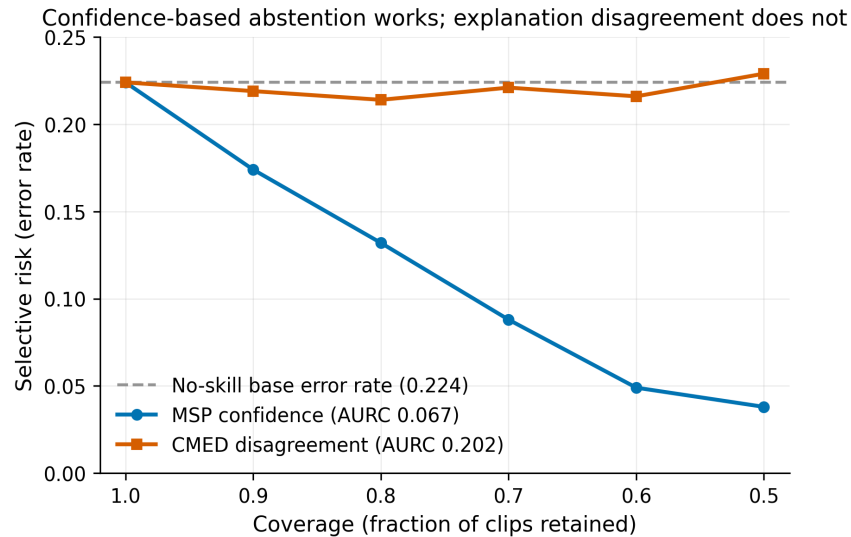


Figure 4. Risk-coverage curves for selective prediction. Abstaining by confidence reduces the error rate steeply; abstaining by explanation disagreement tracks the no-skill base rate. Lower curves and lower AURC are better.

Figures 5 and 6 summarize calibration. The optimal temperatures (Figure 5) show that the visual and audio detectors are overconfident ($T > 1$) while the fused score is underconfident ($T < 1$), confirming that calibration does not compose under fusion. Figure 6 shows that temperature scaling substantially improves the visual detector but barely the audio detector, whose overconfidence is too severe to fully correct.

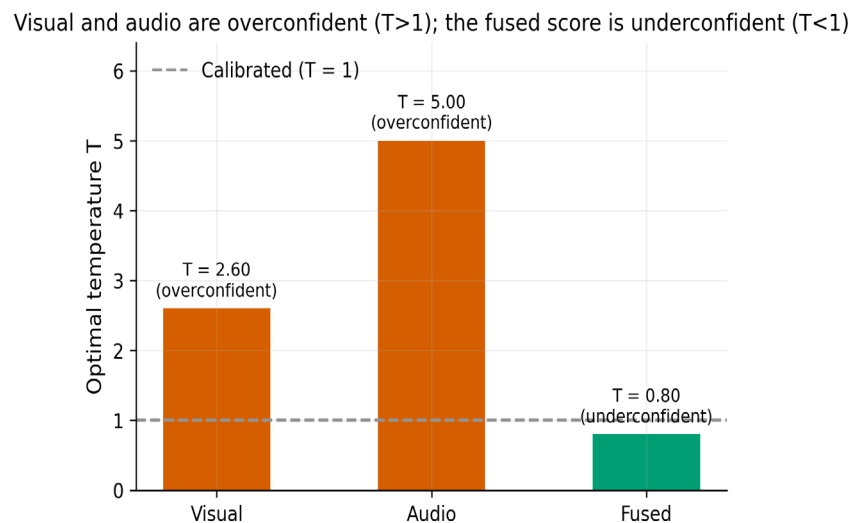


Figure 5. Optimal temperature T per detector against the calibration reference $T = 1$. Bars above 1 indicate overconfidence (visual and audio); the fused bar below 1 indicates underconfidence, so calibration does not compose under fusion.

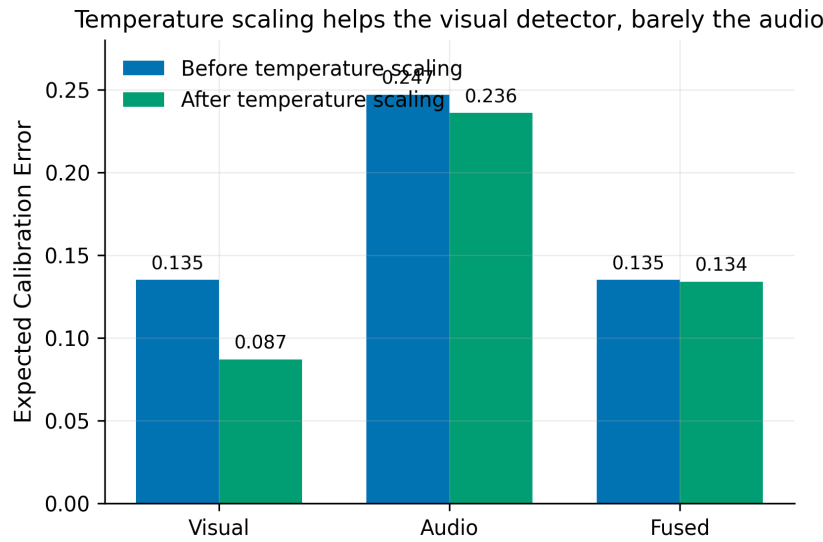


Figure 6. Expected Calibration Error before and after temperature scaling. Scaling helps the visual detector but barely improves the audio detector.

5.3 Proposed Improvements

The audit yields concrete, deployable guidance rather than a new model. First, because confidence is the reliable failure signal, confidence-based selective prediction should be adopted operationally: on this system, abstaining on the least-confident half of clips reduces the error rate roughly six-fold, from 22.4% to 3.8%, with an Area Under the Risk-Coverage curve of 0.067 versus a no-skill base rate of 0.224. This converts the failure-prediction result into an actionable lever for human-in-the-loop review, in which low-confidence clips are routed to an analyst. Second, gradient explanations of accurate audio detectors should not be trusted without validation, because they can direct an analyst to the wrong temporal evidence; perturbation- or attention-based explanations are candidate replacements for the unfaithful gradient attribution and a direction for future work. Third, calibration should be measured on the deployed, fused system and recalibrated there, since it does not inherit from the component streams and can even invert their over-confidence into under-confidence. These recommendations are inexpensive, require no retraining, and follow directly from the reliability measurements.

5.4 Validation

The findings are supported by statistical hypothesis tests and robustness checks. All headline quantities carry 95% confidence intervals from 1,000 bootstrap resamples (Tables 2 to 5). The faithfulness dissociation is significant: within audio-fake clips the saliency mass in the manipulated region is far lower for the attribution than for the decision (Wilcoxon signed-rank test, $p < 10^{-39}$), and the audio-attribution interval does not overlap the audio-decision or visual-attribution intervals. The calibration ranking is significant: the audio ECE interval [.208, .287] does not overlap the visual [.113, .179] or fused [.112, .168] intervals, so the audio detector is significantly worse calibrated than both. The negative result for cross-modal explanation disagreement is not an artifact of a single metric: re-evaluating CMED under seven formulations, replacing the Spearman correlation with Pearson correlation, cosine distance, Jensen-Shannon divergence, Wasserstein distance, peak-location distance, and top-k temporal overlap, leaves all seven at chance (AUROC between 0.48 and 0.52) and below confidence. Finally, the methodological checks are themselves validated against ground truth: the scoring pitfall is demonstrated on the same detector and clips (full-clip AUROC 0.612 versus windowed 0.999), and the cross-dataset transfer failure is confirmed across four off-the-shelf detectors.

5.5 Limitations

Several limitations bound the scope of these conclusions and motivate the planned extension. First, the audit uses a single dataset, LAV-DF, whose manipulations are short, co-located lip-sync edits; whether the failure of cross-modal explanation disagreement is specific to co-located manipulations or also holds for independent, whole-clip manipulations is precisely the question the FakeAVCeleb extension is designed to answer. Second, we study two detectors built on a single backbone (EfficientNet-B0) rather than a broad architecture sweep, and the visual stream is

intentionally left weak so that fusion behaviour can be observed; the qualitative reliability findings are expected to be architecture-independent, but this is not established here. Third, explanation faithfulness is measured with one attribution method, integrated gradients, so the diagnosis that the audio attribution is unfaithful is specific to that method; perturbation- and attention-based explanations may behave differently and are deferred to future work. Fourth, faithfulness is quantified against temporal masks, which requires temporally localized ground truth and does not transfer directly to datasets whose manipulations span the whole clip. These limitations concern the external validity of the findings rather than the internal validity of the reported measurements, each of which is supported by bootstrap confidence intervals and hypothesis tests.

6. Conclusion

We conducted a quality and reliability audit of audio-visual deepfake detection, a representative multimodal AI system, across explanation faithfulness, failure prediction, and calibration, using in-domain detectors and the ground-truth temporal annotations of LAV-DF. All four objectives are met. We built reliable in-domain detectors and characterized a localized-manipulation scoring pitfall and a cross-dataset transfer failure; we measured faithfulness against ground truth and isolated the audio explanation failure to the attribution method rather than the model; we benchmarked reliability signals and showed that confidence supports effective selective prediction while cross-modal explanation disagreement does not; and we analyzed per-stream and post-fusion calibration. The unifying contribution is empirical evidence that accuracy and reliability are decoupled: the most accurate (audio) detector is the most overconfident and the least faithfully explained, even though its decisions localize manipulations precisely, while the weaker visual detector is more faithful and more readily calibrated, and the fused system's errors are confidence-boundary errors best anticipated by simple confidence. Benchmark accuracy is therefore a poor proxy for deployable reliability in multimodal AI, and reliability properties must be audited directly. Future work will extend this audit to a second dataset with independent, whole-clip manipulations (FakeAVCeleb) to test whether the cross-modal disagreement result is specific to co-located manipulations, and to perturbation-based faithfulness where temporal masks are unavailable.

References

- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. and Kim, B., Sanity checks for saliency maps, *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- Cai, Z., Stefanov, K., Dhall, A. and Hayat, M., Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization, *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1-10, 2022.
- Cai, Z., Ghosh, S., Dhall, A., Gedeon, T., Stefanov, K. and Hayat, M., Glitch in the matrix: a large scale benchmark for content driven audio-visual forgery detection and localization, *Computer Vision and Image Understanding*, vol. 236, 103818, 2023.
- Cai, Z., Ghosh, S., Adatia, A. P., Hayat, M., Dhall, A., Gedeon, T. and Stefanov, K., AV-Deepfake1M: a large-scale LLM-driven audio-visual deepfake dataset, *Proceedings of the ACM International Conference on Multimedia*, pp. 7414-7423, 2024.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L., ImageNet: a large-scale hierarchical image database, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255, 2009.
- Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M. and Ferrer, C. C., The DeepFake Detection Challenge (DFDC) dataset, *arXiv preprint arXiv:2006.07397*, 2020.
- Gal, Y. and Ghahramani, Z., Dropout as a Bayesian approximation: representing model uncertainty in deep learning, *Proceedings of the International Conference on Machine Learning*, pp. 1050-1059, 2016.
- Geifman, Y. and El-Yaniv, R., Selective classification for deep neural networks, *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q., On calibration of modern neural networks, *Proceedings of the International Conference on Machine Learning*, pp. 1321-1330, 2017.
- Hendrycks, D. and Gimpel, K., A baseline for detecting misclassified and out-of-distribution examples in neural networks, *Proceedings of the International Conference on Learning Representations*, 2017.
- Hooker, S., Erhan, D., Kindermans, P.-J. and Kim, B., A benchmark for interpretability methods in deep neural networks, *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- Jacovi, A. and Goldberg, Y., Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness?, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 4198-4205, 2020.

- Jung, J., Heo, H.-S., Tak, H., Shim, H., Chung, J. S., Lee, B.-J., Yu, H.-J. and Evans, N., AASIST: audio anti-spoofing using integrated spectro-temporal graph attention networks, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6367-6371, 2022.
- Khalid, H., Tariq, S., Kim, M. and Woo, S. S., FakeAVCeleb: a novel audio-video multimodal deepfake dataset, *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.
- Lakshminarayanan, B., Pritzel, A. and Blundell, C., Simple and scalable predictive uncertainty estimation using deep ensembles, *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- Li, Y., Yang, X., Sun, P., Qi, H. and Lyu, S., Celeb-DF: a large-scale challenging dataset for DeepFake forensics, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3207-3216, 2020.
- Lundberg, S. M. and Lee, S.-I., A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- Müller, N., Czempin, P., Diekmann, F., Froghyar, A. and Böttinger, K., Does audio deepfake detection generalize?, *Proceedings of Interspeech*, pp. 2783-2787, 2022.
- Naeini, M. P., Cooper, G. F. and Hauskrecht, M., Obtaining well calibrated probabilities using Bayesian binning, *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2901-2907, 2015.
- Petsiuk, V., Das, A. and Saenko, K., RISE: randomized input sampling for explanation of black-box models, *Proceedings of the British Machine Vision Conference*, 2018.
- Ribeiro, M. T., Singh, S. and Guestrin, C., Why should I trust you? Explaining the predictions of any classifier, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144, 2016.
- Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Nießner, M., FaceForensics++: learning to detect manipulated facial images, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1-11, 2019.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., Grad-CAM: visual explanations from deep networks via gradient-based localization, *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618-626, 2017.
- Sundararajan, M., Taly, A. and Yan, Q., Axiomatic attribution for deep networks, *Proceedings of the International Conference on Machine Learning*, pp. 3319-3328, 2017.
- Tan, M. and Le, Q., EfficientNet: rethinking model scaling for convolutional neural networks, *Proceedings of the International Conference on Machine Learning*, pp. 6105-6114, 2019.
- Zhang, R., Wang, H., Du, M., Liu, H., Zhou, Y. and Zeng, Q., UMMAFormer: a universal multimodal-adaptive transformer framework for temporal forgery localization, *Proceedings of the ACM International Conference on Multimedia*, pp. 8749-8759, 2023.
- Zhu, F., Cheng, Z., Zhang, X.-Y. and Liu, C.-L., Rethinking confidence calibration for failure prediction, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 518-536, 2022.

Biography

Md Anas Biswas is a master's student on the MSc in Artificial Intelligence and Machine Learning in the School of Computing at the University of Portsmouth. His research focuses on the reliability and trustworthiness of machine-learning systems, including calibration, explanation faithfulness, fairness auditing, and recourse under deployment stress, with applications in media forensics and intrusion detection.