Adaptive Confidence Regularization for Multimodal Failure Detection Supplementary Material

Moru Liu[1], Hao Dong[2], Olga Fink[3], Mario Trapp[1,4]
[1]Technical University of Munich [2]ETH Zürich [3]EPFL [4]Fraunhofer IKS

A. Theoretical Analysis on Confidence Degradation

- Let Y be the discrete label variable taking values in $\{1, \ldots, K\}$.
- Let X_S denote the collection of modality inputs whose indices lie in $S \subseteq \{1, \ldots, M\}$. In particular X_M is the full multimodal input and X_T with $T \subset S$ denotes a subset (less information).
- The conditional entropy of Y on X is $H(Y \mid X)$.

Theorem 1. *Adding more modalities cannot increase the conditional entropy:*

$$H(Y \mid X_S) \leq H(Y \mid X_T), \quad \textit{for } T \subset S.$$

Proof. If $T \subset S$, then by the data processing inequality [2], we have $I(Y; X_T) \leq I(Y; X_S)$, meaning that adding modalities cannot decrease the mutual information between the input and the label. Since $I(Y; X) = H(Y) - H(Y \mid X)$ and $H(Y)$ is fixed, it follows directly that $H(Y \mid X_S) \leq H(Y \mid X_T)$. ■

From Theorem 1, we know that when all modalities are predictive of the target, adding additional input information (i.e., more modalities) cannot increase the conditional entropy. Consequently, the model's uncertainty should not increase as more modalities are provided. Confidence degradation, where predictions become less confident after adding modalities, therefore reflects an undesirable and theoretically inconsistent behavior in multimodal learning.

Theorem 2. *Confidence degradation is linked to a higher theoretical bound on prediction error.*

Proof. Let $P_e = Pr(Y \neq \hat{Y})$ denote the probability of prediction error. By Fano's inequality [35], the error rate is bounded by the true conditional entropy:

$$P_e \geq \frac{H(Y|X) - 1}{\log_2(|Y|)}.$$

Confidence degradation refers to the phenomenon where the fused multimodal prediction has *lower* confidence than at least one of its unimodal predictions. Lower predictive confidence implies greater uncertainty about the output distribution, which corresponds to an increase in $H(Y|X)$. According to Fano's inequality, a higher conditional entropy raises the minimum theoretical bound on error. Thus, while not a direct cause, confidence degradation reflects a state where the model's perceived uncertainty is higher, which is strongly associated with a higher potential for misclassification. ■

Theorem 2 formalizes the key phenomenon we observe in practice: failures in multimodal systems frequently occur together with confidence degradation.

B. Broader Impact, Limitations, and Future Work

Broader Impact. The ACR framework presented in this work has the potential to generate a significant positive societal impact. As AI systems, especially multimodal ones, are increasingly integrated into safety-critical applications such as autonomous driving, medical diagnosis, and industrial robotics, the ability to reliably detect potential failures (i.e., misclassifications) is paramount. ACR directly contributes to enhancing the safety and trustworthiness of these systems by providing a dedicated mechanism to identify and flag uncertain predictions before they can lead to adverse outcomes. This can foster greater public trust and accelerate the responsible adoption of AI technologies in areas where reliability is non-negotiable. Furthermore, by improving the understanding of failure modes in multimodal models, such as the identified confidence degradation phenomenon, our work can guide the development of more robust and dependable AI systems across various domains, ultimately leading to safer and more effective human-AI collaboration.

Limitations. While ACR demonstrates strong gains across multiple datasets and modalities, it has several limitations. First, while we demonstrate robustness under certain distribution shifts and OOD scenarios, the behavior of ACR against sophisticated adversarial attacks specifically designed to fool the misclassification detector itself remains an open area. Second, our current study primarily focuses on specific

types of modalities (e.g., video, optical flow, and audio). The generalization of ACR to a very broad range of disparate modalities (e.g., text with medical imaging, or sensor data with acoustic signals) without specific adaptations warrants further investigation.

Future Work. Building upon the contributions of this paper, several avenues for future research present themselves. First, we plan to explore the integration of ACR with online learning and continual learning paradigms. This would allow the misclassification detector to adapt dynamically to evolving data distributions and novel failure modes encountered during real-world deployment, which is crucial for long-term operational reliability. Second, extending and evaluating ACR across a wider spectrum of modalities (e.g., language, audio, tabular data) and a more diverse set of complex, safety-critical tasks (e.g., real-time robotic interaction) is a key direction. Finally, investigating the robustness of ACR against targeted adversarial attacks and developing defense mechanisms will be crucial for deployment in adversarial settings.

C. Related Work

C.1. Failure Detection

Foundational work on failure prediction originates from selective classification, which formalizes the trade-off between prediction accuracy and abstention under uncertainty [4]. In the deep learning era, this concept has re-emerged as *failure detection* (FD), where the goal is to identify test samples on which a model is likely to be incorrect [10, 33]. Thresholding the maximum softmax probability (MSP) [13] remains a strong baseline, yet its vulnerability to overconfidence limits its reliability [32, 33]. Beyond MSP, a large body of work aims to more effectively estimate prediction risk.

One stream of research focuses on training auxiliary modules, such as heads or separate networks, to explicitly predict correctness based on intermediate features or model logits [5, 11, 17]. For instance, ConfidNet [5] introduces a dedicated confidence estimation head that operates on penultimate-layer features. A distinct line of research adopts an integrated approach, jointly optimizing FD functionalities with the primary model's training objective. Representative techniques in this category include improving the separability of correct and incorrect representations [28], enhancing confidence ranking [29], regularizing based on sample typicality [27], and enforcing flatter loss landscapes around misclassified examples [46]. Concurrently, data augmentation strategies have proven effective, primarily by exposing the model to synthesized failure cases during training [3, 12, 23, 47]. However, all previous approaches were designed for unimodal scenarios, without accounting for the interaction and complementary nature of diverse modalities.

C.2. Out-of-Distribution Detection

OOD detection shares a similar objective with FD but addresses fundamentally different challenges [30]. Specifically, OOD detection aims to identify test samples that exhibit semantic shifts from the training distribution, typically without compromising in-distribution (ID) classification accuracy. OOD detection has been extensively investigated in recent years [42, 44], encompassing diverse approaches such as post-hoc scoring [13, 24, 26], feature-based techniques [21, 39], outlier exposure [14, 41, 43], and reconstruction-based methods [7, 45]. Multimodal OOD detection [9, 22] is also an emerging research area. Although OOD detection methods are often employed as baselines for FD, recent studies [15, 46, 47] demonstrate that techniques optimized for OOD detection generally exhibit suboptimal performance on FD tasks. This finding underscores the necessity for developing specialized FD approaches.

C.3. Interpretability-Based Failure Prediction

A complementary line of work investigates predicting failures by analyzing a model's internal reasoning rather than relying solely on output confidence [16, 38]. Recent methods [31, 34] leverage post-hoc interpretability techniques (e.g., Grad-CAM [36], saliency [1], concept activation [19]) and train auxiliary meta-predictors to detect unreliable explanations. These approaches identify patterns such as noisy attention, background-biased focus, and lack of class-discriminative localization, which often correlate with misclassifications. While informative, interpretability-driven failure prediction requires generating explanations for every test sample, incurring significant computational overhead, and remains decoupled from the model's training dynamics. In contrast, ACR embeds failure awareness directly into training and enables failure prediction in a single forward pass, eliminating the cost and complexity associated with post-hoc explanations.

D. Further Details on Datasets

Our framework is primarily evaluated on four action recognition datasets from the MultiOOD benchmark [9]: HMDB51 [20], EPIC-Kitchens [6], HAC [8], and Kinetics-600 [18]. For evaluating multimodal Out-of-Distribution (OOD) detection in ablation studies, we additionally employ the UCF101 dataset [37], also from the MultiOOD benchmark. Fig. 1 illustrates the four main datasets used for multimodal failure detection.

HMDB51 is an action recognition dataset comprising 6,766 video clips distributed across 51 action categories. The videos are sourced from digitized movies and YouTube. This dataset includes both video and optical flow modalities. **EPIC-Kitchens** is an egocentric video dataset capturing daily kitchen activities recorded by 32 participants. For our

Figure 1. An illustration of the four main datasets we used for multimodal failure detection.

experiments, we utilize a subset of 4,871 video clips from participant P22, encompassing eight common actions (*put*, *take*, *open*, *close*, *wash*, *cut*, *mix*, and *pour*). The dataset provides video and optical flow modalities. **Kinetics-600** is a large-scale action recognition dataset containing approximately 480,000 10-second clips distributed across 600 action classes. Following [9], we utilize a subset of 100 classes, resulting in 24,981 video clips for our study. This dataset offers video, audio, and optical flow modalities. **HAC** contains 3,381 video clips featuring seven action categories (e.g., *sleeping*, *watching TV*, *eating*, *running*) performed by humans, animals, and cartoon characters. The dataset includes video, optical flow, and audio modalities. **UCF101** is a diverse video action recognition dataset consisting of 13,320 clips across 101 action classes. The videos, sourced from YouTube, exhibit significant variation in camera motion, object appearance, scale, pose, viewpoint, and background. This dataset provides video and optical flow modalities.

E. More Implementation Details

Pseudo Code for Multimodal Feature Swapping. We provide the pseudo code for multimodal outlier synthesis in Algorithm 1, where we dynamically swap multimodal feature embeddings and assign them corresponding soft labels. MFS naturally generalizes beyond the two-modality setting. In Algorithm 2, we illustrate how to extend MFS to three modalities. We extend MFS to this setting by randomly selecting a start index for each modality and swapping them cyclically: $\mathbf{E}^1 \rightarrow \mathbf{E}^2$, $\mathbf{E}^2 \rightarrow \mathbf{E}^3$, and $\mathbf{E}^3 \rightarrow \mathbf{E}^1$, thus creating outlier features for all three modalities. This generalized formulation enables flexible synthesis of multimodal failure cases, capturing higher-order inconsistencies (e.g., tri-modal conflicts) and scaling effectively to heterogeneous sensor configurations commonly found in real-world systems.

Extension of ACL to More Modalities. Our framework is not limited to two modalities and can be easily extended to M modalities. Given a training sample $\mathbf{x}$ with M modalities, we obtain prediction confidence score *conf* from the combined embeddings of all modalities, and $conf_1$, $conf_2$, ..., $conf_M$ from each modality. The Adaptive Confidence Loss

Algorithm 1 Multimodal Feature Swapping

Input: ID feature $\mathbf{E} = [\mathbf{E}^1, \mathbf{E}^2]$, where $\mathbf{E}^1$ and $\mathbf{E}^2$ are from modality 1 and 2; minimum and maximum number n_{min} and n_{max} for swapping; ground-truth one-hot label $\mathbf{y}_{\text{true}}$ for $\mathbf{E}$ and $\mathbf{y}_{\text{outlier}} = C + 1$.

Pseudo Code:

Sample $n_{\text{swap}} \sim \mathcal{U}(n_{\text{min}}, n_{\text{max}})$, $\lambda = \frac{n_{\text{swap}}}{n_{\text{max}}}$;

Randomly select start indices s_1, s_2 for modality 1 and 2;

Clone features $\widetilde{\mathbf{E}}^1 \leftarrow \mathbf{E}^1$, $\widetilde{\mathbf{E}}^2 \leftarrow \mathbf{E}^2$;

Swap n_{swap} dimensions across modalities:

$$\widetilde{\mathbf{E}}^1[s_1{:}s_1{+}n_{\text{swap}}] \leftarrow \mathbf{E}^2[s_2{:}s_2{+}n_{\text{swap}}]$$

$$\widetilde{\mathbf{E}}^2[s_2{:}s_2{+}n_{\text{swap}}] \leftarrow \mathbf{E}^1[s_1{:}s_1{+}n_{\text{swap}}]$$

Generate label for outlier feature $\mathbf{y}_{\text{swapped}} = (1 - \lambda)\mathbf{y}_{\text{true}} + \lambda \mathbf{y}_{\text{outlier}}$.

Output: Multimodal outlier feature $\mathbf{E}_o = [\widetilde{\mathbf{E}}^1, \widetilde{\mathbf{E}}^2]$ and label $\mathbf{y}_{\text{mixed}}$.

Algorithm 2 Multimodal Feature Swapping with Three Modalities

Input: ID feature $\mathbf{E} = [\mathbf{E}^1, \mathbf{E}^2, \mathbf{E}^3]$, where $\mathbf{E}^1$, $\mathbf{E}^2$, and $\mathbf{E}^3$ are from modality 1, 2, and 3; minimum and maximum number n_{min} and n_{max} for swapping; ground-truth one-hot label $\mathbf{y}_{\text{true}}$ for $\mathbf{E}$ and $\mathbf{y}_{\text{outlier}} = C + 1$.

Pseudo Code:

Sample $n_{\text{swap}} \sim \mathcal{U}(n_{\text{min}}, n_{\text{max}})$, $\lambda = \frac{n_{\text{swap}}}{n_{\text{max}}}$;

Randomly select start indices s_1, s_2, s_3 for modality 1, 2, and 3;

Clone features $\widetilde{\mathbf{E}}^1 \leftarrow \mathbf{E}^1$, $\widetilde{\mathbf{E}}^2 \leftarrow \mathbf{E}^2$, $\widetilde{\mathbf{E}}^3 \leftarrow \mathbf{E}^3$;

Swap n_{swap} dimensions across modalities:

$$\widetilde{\mathbf{E}}^1[s_1{:}s_1{+}n_{\text{swap}}] \leftarrow \mathbf{E}^3[s_3{:}s_3{+}n_{\text{swap}}]$$

$$\widetilde{\mathbf{E}}^2[s_2{:}s_2{+}n_{\text{swap}}] \leftarrow \mathbf{E}^1[s_1{:}s_1{+}n_{\text{swap}}]$$

$$\widetilde{\mathbf{E}}^3[s_3{:}s_3{+}n_{\text{swap}}] \leftarrow \mathbf{E}^2[s_2{:}s_2{+}n_{\text{swap}}]$$

Generate label for outlier feature $\mathbf{y}_{\text{swapped}} = (1 - \lambda)\mathbf{y}_{\text{true}} + \lambda \mathbf{y}_{\text{outlier}}$.

Output: Multimodal outlier feature $\mathbf{E}_o = [\widetilde{\mathbf{E}}^1, \widetilde{\mathbf{E}}^2, \widetilde{\mathbf{E}}^3]$ and label $\mathbf{y}_{\text{mixed}}$.

	AURC↓	AUROC↑	FPR95↓	ACC↑
128	29.08	88.27	49.57	86.66
256	25.11	90.55	46.22	86.43
512	25.34	90.98	43.90	85.97

Table 1. Effect of n_{max} in MFS.

	AURC↓	AUROC↑	FPR95↓	ACC↑
0.2	24.42	90.19	46.15	86.66
0.5	24.18	90.28	46.22	87.00
1.0	23.11	90.93	42.11	86.55
2.0	19.97	92.02	41.96	87.23

Table 2. Effect of weight λ_{acl} for ACL.

	AURC↓	AUROC↑	FPR95↓	ACC↑
MSP	30.24	86.63	62.75	86.03
ACR	**15.38**	**87.79**	**58.06**	**91.51**

Table 3. Results on Office-Home dataset by fusing images from distinct domains.

can then be defined as:

$$\mathcal{L}_{\text{acl}} = \frac{1}{M}\sum_{i=1}^{M} \max(0, \mathit{conf}_i - \mathit{conf}). \quad (1)$$

F. More Ablation Studies

Parameter Sensitivity. We evaluate the sensitivity of our framework to two key hyperparameters using the HMDB51 dataset. First, the maximum swapping dimension n_{max} for Multimodal Feature Swapping (MFS) was varied among 128, 256, and 512, with results presented in Table 1. An n_{max} value of 256 yielded the optimal balance, achieving robust performance across all evaluation metrics. Subsequently, with n_{max} fixed at 256, the weight λ_{acl} for Adaptive Confidence Loss (ACL) was evaluated over the set 0.2, 0.5, 1.0, and 2.0 (detailed in Table 2). A value of $\lambda_{\text{acl}} = 2.0$ consistently delivered the strongest FD performance. Importantly, the framework's performance remained stable across both parameter sweeps, underscoring its robustness to variations in these hyperparameters.

Generalize to Image Classification. We evaluated on Office-Home Dataset [40] by treating images from distinct domains (i.e., Art and RealWorld) as different modalities and fusing them. We show that ACR can be generalized to the image classification task in Tab. 3.

G. Further Discussions on Proposed Modules

G.1. Mechanism of ACL

In Sec. 2.2, we show failures in multimodal systems frequently coincide with confidence degradation. We provided theoretical analysis in Appendix Sec. 1 and show that when distinct modalities are predictive of a target, adding modalities cannot increase conditional entropy. Hence, we introduce ACL to enforce better information integration, ensuring that fused evidence yields more separable confidence between correct and incorrect samples. The mechanism relies on the interplay between ACL and the cross-entropy loss ($\mathcal{L}_{\text{cls}}$). When predictions are ***correct***, $\mathcal{L}_{\text{cls}}$ and ACL work together to drive the fused confidence higher. On ***incorrect*** samples, $\mathcal{L}_{cls}$ pushes fused confidence down. If a unimodal branch remains overconfident, ACL remains high. To minimize total loss, gradients propagate to suppress the overconfident unimodal encoder, as the fused head cannot rise without violating $\mathcal{L}_{cls}$. This ensures that ***ACL boosts confidence only for correct predictions while avoiding overconfidence in incorrect predictions (Tab. 4)***, hence (together with MFS) widening the confidence separation that directly benefits failure detection.

Method	Correct Predictions (↑)				Incorrect Predictions (↓)			
	HMDB51	EPIC	HAC	Kinetics	HMDB51	EPIC	HAC	Kinetics
w/o ACL	0.95	0.91	0.94	0.92	0.80	0.73	0.74	0.64
w/ ACL	**0.96**	**0.93**	**0.96**	**0.93**	**0.63**	**0.69**	**0.62**	**0.57**

Table 4. Average fused confidence on **correct** and **incorrect** predictions.

How Adaptive Confidence Loss Addresses Unimodal Overconfidence. Imagine a situation where one modality (e.g., audio) is ambiguous or corrupted, causing its corresponding unimodal network to make a confidently wrong prediction (high conf_1). The other modality (e.g., video) provides clear evidence for the correct class, leading to a correct, high-confidence prediction (conf_2). A standard fusion model might struggle to integrate these conflicting signals. If the fusion process results in a moderate fused confidence conf that is lower than the erroneously high conf_1, it will incur a large ACL penalty from the $\max(0, \mathit{conf}_1 - \mathit{conf})$ term. To minimize this loss, the model can't easily force the fused confidence up without a good reason, as that would be penalized by the cross-entropy loss term $\mathcal{L}_{\text{cls}}$ if the prediction is wrong. Instead, a more effective way to reduce the ACL loss is to lower the confidence of the overconfident unimodal prediction (conf_1). Through repeated exposure to such conflicting examples during training, the unimodal feature extractor learns a valuable lesson. It learns that producing a high-confidence prediction that is likely to be contradicted by another modality is "expensive" in terms of the overall loss. Therefore, it adjusts its weights to become more cautious and better calibrated. The unimodal network learns to associate maximum confidence not just with strong internal features, but with features that are also robust and likely to lead to cross-modal agreement. As shown in Fig. 2 and Fig. 3, training with ACL effectively alleviates unimodal overconfidence on incorrect predictions across all datasets.

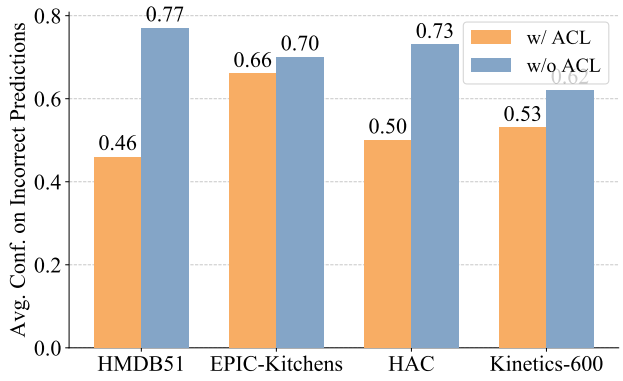

Figure 2. The average confidence on incorrect predictions for video modality w/ and w/o ACL.

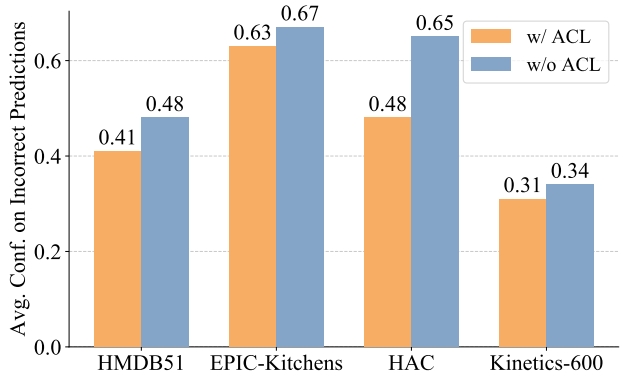

Figure 3. The average confidence on incorrect predictions for optical flow modality w/ and w/o ACL.

G.2. Mechanism of MFS

Multimodal Feature Swapping. Our MFS module isn't designed to replicate the entire, complex distribution of all possible real-world failures, as this would be intractable. Instead, MFS serves as a principled and targeted regularizer that exposes the model to a critical and common failure mode in multimodal systems: cross-modal inconsistency. Our core hypothesis is that many real-world errors arise from conflicting or ambiguous signals between modalities. MFS directly simulates this failure mode by swapping feature segments. Unlike generic outlier exposure (OE) or manifold mixup, MFS creates challenging, near-ID hard negatives by preserving intra-modal semantics while breaking cross-modal consistency. Ultimately, the model learns *which structurally ambiguous cases should stay low-confidence*.

The feature swapping operation preserves local modality structure within unswapped segments, maintaining realistic intra-modal patterns while disrupting cross-modal alignment through swapped segments, simulating realistic sensor conflicts or temporal misalignments. It can also generate outliers of varying difficulty by adjusting the n_{swap} parameter, ranging from subtle inconsistencies when n_{swap} is small to severe cross-modal conflicts when n_{swap} is large.

Training on these synthesized samples forces the model to develop a more robust fusion mechanism that must critically assess the semantic agreement between modalities. This closer examination of cross-modal consistency improves its ability to detect real-world misclassifications, which often exhibit similar signal conflicts. The effectiveness of this approach is validated by our extensive empirical results, which demonstrate that learning to reject these synthetic inconsistencies directly translates to better identification of real-world errors. Fig. 4 gives a detailed illustration of how MFS enables the detection and rejection of uncertain predictions, thereby improving model reliability. The soft labeling parameter λ provides calibrated training signals proportional to the degree of synthetic corruption, allowing the model to learn fine-grained uncertainty estimation.

Why Contiguous Swapping and Effective? Real multimodal failures are typically localized/structured rather than i.i.d. white noise (e.g., a few seconds of missing audio, a region of blurred video, or temporal misalignment). Swapping contiguous blocks provides a simple inductive bias that introduces localized disruption instead of globally destroying the representation. Importantly, this choice is *empirically validated*: contiguous swapping outperforms random perturbations (noise/drop/random mixing) in our ablations in Tab. 7.

Semantic Preservation. MFS swaps only a subset of feature dimensions, leaving most dimensions unchanged and thereby *preserving intra-modal semantics*. As evidence, predictions from original and MFS features agree in 99.2% of cases *(indicating preserved semantics)*, with the MFS features yielding 0.1 lower confidence on average.

Why MFS Succeeds where OE Fails? OE mainly trains rejection under semantic shift (OOD). In contrast, *multimodal failures often occur within ID semantics* but with conflicting/ambiguous evidence across modalities. MFS explicitly synthesizes such cross-modal inconsistency, producing samples that remain semantically ID yet structurally unreliable.

Difference between MFS and Feature Mixing. While both MFS and Feature Mixing [25] generate outliers by swapping feature dimensions between modalities, their primary goal and supervision strategies are fundamentally different. MFS is designed for misclassification detection. Its goal is to teach the model to identify when it's making a mistake on in-distribution data. The model learns to classify the outlier using a specific soft label that mixes the true class with a new "outlier" class. This helps it recognize failure-aware samples that have conflicting internal signals. Feature Mixing is designed for out-of-distribution detection. Its goal is to teach the model to identify data that comes from unknown classes not seen during training. The model learns to be confused by the outlier. The entropy loss forces the model to be uncertain, outputting a flat, uniform probability distribution over all known classes. This teaches it to assign low confidence to

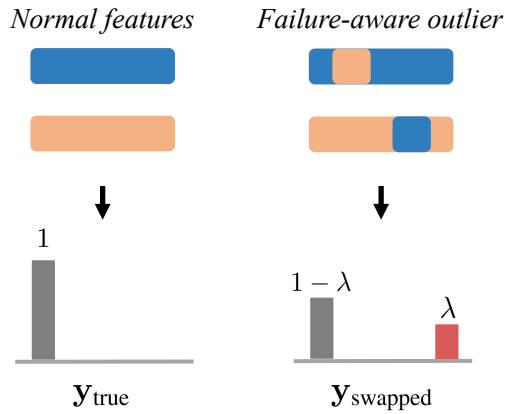

Figure 4. An illustration of how MFS enables the detection and rejection of uncertain predictions, thereby improving model reliability. The failure-aware outliers generated by MFS are treated as uncertain because they introduce cross-modal inconsistency. Accordingly, their soft label $\mathbf{y}_{\text{swapped}}$ reduces the ground-truth class probability from 1 to $1-\lambda$, compared to $\mathbf{y}_{\text{true}}$. Training with cross-entropy loss on $\mathbf{y}_{\text{swapped}}$ ensures that the prediction confidence on failure-aware outliers remains lower than on normal features, which is exactly the desired outcome.

anything that doesn't look like in-distribution data.

As for the implementation, MFS selects a single contiguous block of feature dimensions per modality (sample start indices s_1, s_2, and swap n_{swap} consecutive dims). The size n_{swap} is sampled from $\mathcal{U}(n_{\text{min}}, n_{\text{max}})$, so it can control how close and far the synthesized outliers lie from ID. Feature Mixing randomly samples an arbitrary subset of N feature dimensions from each modality (non-contiguous, independent indices) and swaps those positions between modalities (simple index-based swap). Our ablation results show that MFS achieves superior performance compared to Feature Mixing in the context of multimodal failure detection.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 2

[2] Normand J Beaudry and Renato Renner. An intuitive proof of the data processing inequality. *arXiv preprint arXiv:1107.0740*, 2011. 1

[3] Zhen Cheng, Fei Zhu, Xu-Yao Zhang, and Cheng-Lin Liu. Breaking the limits of reliable prediction via generated data. *IJCV*, 2024. 2

[4] C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. 2

[5] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *NeurIPS*, 2019. 2

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 2

[7] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *CVPR*, 2021. 2

[8] Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. SimMMDG: A simple and effective framework for multi-modal domain generalization. In *NeurIPS*, 2023. 2

[9] Hao Dong, Yue Zhao, Eleni Chatzi, and Olga Fink. Multiood: Scaling out-of-distribution detection for multiple modalities. In *NeurIPS*, 2024. 2, 3

[10] Hao Dong, Moru Liu, Jian Liang, Eleni Chatzi, and Olga Fink. To trust or not to trust your vision-language model's prediction. *arXiv preprint arXiv:2505.23745*, 2025. 2

[11] Federica Granese, Marco Romanelli, Daniele Gorla, Catuscia Palamidessi, and Pablo Piantanida. Doctor: A simple method for detecting misclassification errors. In *NeurIPS*, 2021. 2

[12] Shuangpeng Han and Mengmi Zhang. Unveiling ai's blind spots: An oracle for in-domain, out-of-domain, and adversarial errors. *arXiv preprint arXiv:2410.02384*, 2024. 2

[13] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 2

[14] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 2

[15] Paul F Jaeger, Carsten T Lüth, Lukas Klein, and Till J Bungert. A call to reflect on evaluation practices for failure detection in image classification. *arXiv preprint arXiv:2211.15259*, 2022. 2

[16] Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. *arXiv preprint arXiv:2206.14754*, 2022. 2

[17] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *NeurIPS*, 2018. 2

[18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[19] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, pages 2668–2677. PMLR, 2018. 2

[20] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 2

[21] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018. 2

[22] Shawn Li, Huixian Gong, Hao Dong, Tiankai Yang, Zhengzhong Tu, and Yue Zhao. Dpu: Dynamic prototype updating for multimodal out-of-distribution detection. *arXiv preprint arXiv:2411.08227*, 2024. 2

[23] Yuting Li, Yingyi Chen, Xuanlong Yu, Dexiong Chen, and Xi Shen. Sure: Survey recipes for building reliable and robust deep networks. In *CVPR*, 2024. 2

[24] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 2

[25] Moru Liu, Hao Dong, Jessica Kelly, Olga Fink, and Mario Trapp. Extremely simple multimodal outlier synthesis for out-of-distribution detection and segmentation. *arXiv preprint arXiv:2505.16985*, 2025. 5

[26] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020. 2

[27] Yijun Liu, Jiequan Cui, Zhuotao Tian, Senqiao Yang, Qingdong He, Xiaoling Wang, and Jingyong Su. Typicalness-aware learning for failure detection. *Advances in Neural Information Processing Systems*, 37:11456–11478, 2024. 2

[28] Yan Luo, Yongkang Wong, Mohan S Kankanhalli, and Qi Zhao. Learning to predict trustworthiness with steep slope loss. In *NeurIPS*, 2021. 2

[29] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *ICML*, 2020. 2

[30] Vivek Narayanaswamy, Yamen Mubarka, Rushil Anirudh, Deepta Rajan, Andreas Spanias, and Jayaraman J Thiagarajan. Know your space: Inlier and outlier construction for calibrating medical ood detectors. *arXiv preprint arXiv:2207.05286*, 2022. 2

[31] Kien X Nguyen, Tang Li, and Xi Peng. Interpretable failure detection with human-level concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 26326–26334, 2025. 2

[32] Francesco Pinto, Harry Yang, Ser Nam Lim, Philip Torr, and Puneet Dokania. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. In *NeurIPS*, 2022. 2

[33] Xin Qiu and Risto Miikkulainen. Detecting misclassification errors in neural networks with a gaussian process model. In *AAAI*, 2022. 2

[34] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 ieee conference on secure and trustworthy machine learning (satml)*, pages 464–483. IEEE, 2023. 2

[35] Jonathan Scarlett and Volkan Cevher. An introductory guide to fano's inequality with applications in statistical estimation. *arXiv preprint arXiv:1901.00555*, 2019. 1

[36] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[37] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014. 2

[38] Rakshith Subramanyam, Kowshik Thopalli, Vivek Narayanaswamy, and Jayaraman J Thiagarajan. Decider: Leveraging foundation model priors for improved model failure detection and explanation. In *ECCV*, 2024. 2

[39] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *ICML*, 2022. 2

[40] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017. 4

[41] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *ICCV*, 2021. 2

[42] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024. 2

[43] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *ICCV*, 2019. 2

[44] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023. 2

[45] Yibo Zhou. Rethinking reconstruction autoencoder-based out-of-distribution detection. In *CVPR*, 2022. 2

[46] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Rethinking confidence calibration for failure prediction. In *ECCV*, 2022. 2

[47] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Openmix: Exploring outlier samples for misclassification detection. In *CVPR*, 2023. 2