UNCERTAINTY AS HETEROGENEOUS EVIDENCE: MUSE FOR VISION-LANGUAGE MODEL FAILURE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–Language Models (VLMs) are increasingly used in high-stakes settings such as disaster response, where identifying unreliable predictions is essential for effective decision-making. Most current VLM failure-prediction methods rely on a single uncertainty perspective, which may overlook important reliability signals. We propose that VLM failures produce diverse reliability signatures that cannot be fully captured by one perspective alone. To address this, we present MUSE, a Multi-perspective Uncertainty Synthesis Ensemble that integrates five explicitly defined reliability perspectives: cross-modal alignment, semantic competition, intra-class geometry, class-conditional density, and neighborhood consistency. MUSE operates on frozen VLM representations without retraining or modifying the underlying model and does not require access to the training objective, enabling retrospective reliability estimation across pretrained backbones. Across 20 classification dataset–task settings, MUSE achieves mean AUROCs of 94.26%, 95.02%, and 94.36% on CLIP, SigLIP, and OpenCLIP, respectively, outperforming the strongest baseline by +7.99, +5.33, and +6.51 AUROC percentage points. Correlation analyses demonstrate distinct complementarity among these reliability perspectives, while cross-modal retrieval experiments further demonstrate applicability to real-world multimodal disaster-response scenarios. These findings support multi-perspective uncertainty synthesis as a robust approach to VLM failure prediction.

Keywords: VLMs; Failure Prediction; Uncertainty Quantification; Multimodal Learning

1 INTRODUCTION

VLMs are proposed to be used for zero-shot visual reasoning components in time-sensitive decision-support pipelines, including disaster management systems that sift large volumes of social-media imagery for situational awareness. In these high-stakes settings, overconfident errors can trigger inappropriate triage decisions, which makes per-instance reliability estimation essential in addition to accuracy.

A practical reliability primitive is failure prediction: given a frozen VLM and a frozen downstream zero-shot decision rule, predict whether each individual output is likely correct. However, prior work on VLM failure or uncertainty prediction is often organized around a single uncertainty estimator or a single modeling perspective (e.g., confidence-based scores such as Maximum Concept Matching (MCM) (27), learned visual-only heads such as Learning Visual Uncertainties (LVU) (8; 52; 20), or multimodal variants such as the Vision–Language Uncertainty quantification framework (ViLU) (22)). This one-estimator viewpoint can under-address failure modes that are not well captured by the chosen signal, such as failures stemming from weak cross-modal alignment, ambiguous semantic competition, off-manifold samples, density mismatch, or neighborhood consistency. This motivates a methodological gap: reliability estimation for deployed VLMs should be (i) post hoc (i.e., applicable after the VLM is trained and does not require its internal weights), (ii) non-invasive (no retraining or modification of the VLM for the failure prediction task), and (iii) multi-perspective so that complementary estimators can cover distinct failure modes that a single estimator may miss.

We hypothesize that uncertainty for VLM outputs is best treated as heterogeneous evidence: an estimator that succeeds on one type of failure may fail on another, while a complementary estimator may succeed exactly where the first fails. To test this hypothesis, we propose **MUSE** (**M**ulti-perspective **U**ncertainty **S**ynthesis **E**nsemble), an post hoc failure-prediction framework that integrates five complementary reliability signals computed from pretrained VLM representations: (i) **ATLAS** (Alignment-based Text-visual Learned Assessment System), a discriminative alignment detector; (ii) **SOARER** (Second-Order Affinity-based Reliability Estimation Readout), which models class-similarity distribution shape and second-order competition; (iii) **ICS** (Intra-class Statistics), which captures intra-class structure and off-manifold behavior; (iv) **AugLikelihoodQDA** (Augmented Likelihood-based Quadratic Discriminant Analysis), a class-conditional density/likelihood signal; and (v) **SWNC** (Similarity-Weighted Neighborhood Consistency), which measures dual-view neighborhood consistency. MUSE operates purely on frozen embeddings and downstream

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

outputs, requiring neither access to the VLM's internal weights nor its original training loss, thereby providing a loss-agnostic retroactive reliability layer.

We validate MUSE on classification and cross-modal retrieval failure prediction across diverse benchmarks and three frozen VLM backbones: **CLIP** (Contrastive Language–Image Pretraining) (34), **SigLIP** (Sigmoid Loss for Language–Image Pretraining) (54), and **OpenCLIP** (an open-source implementation of CLIP) (6). We evaluate detectors using AUROC (or AUC), FPR@95 (False Positive Rate at 95% Recall), and Cohen's d, averaging over multiple random seeds with hyperparameters selected strictly on validation splits, and an uncontaminated test set. Our contributions are: We (i) formulate VLM failure prediction as a multi-perspective uncertainty synthesis problem and treat uncertainty as heterogeneous evidence rather than a single confidence measure; (ii) introduce **MUSE**, a non-invasive, post hoc framework comprising ATLAS, SOARER, ICS, AugLikelihoodQDA, and SWNC to enable reliability estimation without retraining the underlying VLM; (iii) provide extensive empirical validation across datasets, backbones and seeds, showing improvements over individual estimators and sota (state of the art) baselines.

Prior work has been criticized for degrading under strong distribution shifts(24) (10). Compared to standard corpora specifically processed for training industrial scale VLMs, MUSE is evaluated under a substantially more severe form of distribution shift, that is, multimodal disaster tweet pair datasets (CrisisMMD(2), MHII (Multimodal Humanitarian Information Identification)(2), UC Irvine Multimodal(28)), which contain actual social media data and exhibit pragmatic grounding mismatch, colloquial text, and event-specific phrasing where sota VLM zero-shot retrieval accuracy drops to as low as 12%. MUSE maintains 82–91% AUC in these regimes, validating that heterogeneous evidence fusion is robust precisely where single-signal methods collapse.

We propose that VLM failures form a partitioned evidence space with five mathematically distinct axes: (i) first-order cross-modal alignment (ATLAS), (ii) second-order semantic competition geometry (SOARER), (iii) intra-class manifold topology (ICS), (iv) generative class-conditional density (AugLikelihoodQDA), and (v) non-parametric dual-view neighborhood consistency (SWNC), where prior work captures only one or two axes. Our empirical Spearman correlation analysis (ρ = 0.63–0.83 between axes) provides direct evidence that these axes carry statistically distinct information. MUSE is the only framework that simultaneously (i) integrates multiple heterogeneous perspectives, (ii) requires no auxiliary vision encoder, (iii) evaluates cross-modal retrieval, and (iv) reports statistical significance across seeds. MUSE requires a labeled training pool with observed correctness outcomes. In safety-critical deployments where reliability estimation is essential, a small labeled failure-history pool is precisely the resource that becomes available after initial deployment. We frame MUSE as a retrospective reliability layer designed for exactly this operational regime.

2 Related Work

A widely used baseline for misclassification and out-of-distribution (OOD) detection is to score samples using softmax-derived confidence such as Maximum Class Probability (MCP) or entropy (16), but these scores can be miscalibrated and overconfident under shift. Prior work improves failure prediction by learning alternative confidence targets (e.g., True Class Probability, TCP) (8), predicting losses via auxiliary modules (52), designing black-box misclassification detectors (13), and learning relative uncertainty measures from probability patterns (9). Other directions explicitly model uncertainty using distributional outputs (e.g., evidential/Dirichlet formulations) (39), transfer uncertainty via pretrained uncertainty modules (20), and improve robustness by increasing estimator diversity, e.g., heterogeneous ensembles (26). However, such ensemble approaches require the training and storage of multiple full models, and have a large operational footprint than post-hoc scoring on cached embeddings (MUSE).

Vision–Language Models (VLMs) enable strong zero-shot classification and retrieval, but their embeddings are usually deterministic and do not directly provide calibrated reliability. MCM extends MCP-style scoring to VLM image–text similarities (27). Post-hoc probabilistic treatments aim to quantify uncertainty over similarity scores without retraining the backbone (3), while ViLU uses cross-attention to learn a multimodal uncertainty predictor operating on frozen vision–language embeddings (22). Certain VLM OOD methods exploit structure in label space, e.g., semantic consistency across related in-distribution labels (56), or strengthen global/local alignment via additional training objectives (50); these approaches are effective but typically emphasize a single primary signal family (e.g., confidence, probabilistic similarity, or alignment).

Safety-critical perception systems may incorporate explicit failure detection modules, such as LT-SAT which detects tracking failures online via Bayesian inference and supports recovery (33). Relatedly, foundation-model-based anomaly detection leverages generative trajectories (e.g., diffusion paths) as OOD signals without per-dataset retraining (1). A recurring limitation is reliance on a single estimator or inductive bias, despite heterogeneous real-world

failure modes. Our work addresses this by treating VLM reliability as heterogeneous evidence and synthesizing complementary post-hoc signals within MUSE, without retraining or modifying the underlying VLM.

3 Problem Formulation, Notation, and System Architecture

Problem Formulation Given a frozen VLM and a frozen zero-shot target classifier g, the classification failure detection task is to learn a scoring function

$$u : (\mathbf{v}, \mathbf{T}) \mapsto u(\mathbf{v}, \mathbf{T}) \in \mathbb{R},$$

such that higher values of u correspond to samples more likely to be misclassified by g. We frame this as a per-sample binary discrimination problem: the detector u is trained to approximate $P(e_i = 1 \mid \mathbf{v}_i, \mathbf{T}, \mathbf{s}_i)$, evaluated by the separability it induces between correct-prediction and incorrect-prediction populations. Neither the VLM nor the target classifier is fine-tuned.

Notation Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a dataset of image-label pairs, where x_i is an image and $y_i \in \{1, \ldots, K\}$ is its ground-truth class. A frozen vision–language model $\mathcal{F}$ maps images and class-name prompts into a shared d-dimensional embedding space, producing ℓ_2-normalized visual features $\mathbf{v}_i = \mathcal{F}_v(x_i) \in \mathbb{R}^d$ and text features $\{\mathbf{t}_c\}_{c=1}^{K} \subset \mathbb{R}^d$. The stacked text matrix is $\mathbf{T} \in \mathbb{R}^{K \times d}$. A zero-shot target classifier g produces class scores $\mathbf{s}_i = g(\mathbf{v}_i, \mathbf{T}) \in \mathbb{R}^K$ and a predicted label $\hat{y}_i = \arg\max_c s_{i,c}$. The probability vector is $\mathbf{p}_i = \sigma(\mathbf{s}_i)$, where σ denotes either softmax or element-wise sigmoid. The binary failure indicator is $e_i = \mathbb{I}[\hat{y}_i \neq y_i] \in \{0, 1\}$. Training, validation, and test splits are denoted $\mathcal{D}_{\text{tr}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$.

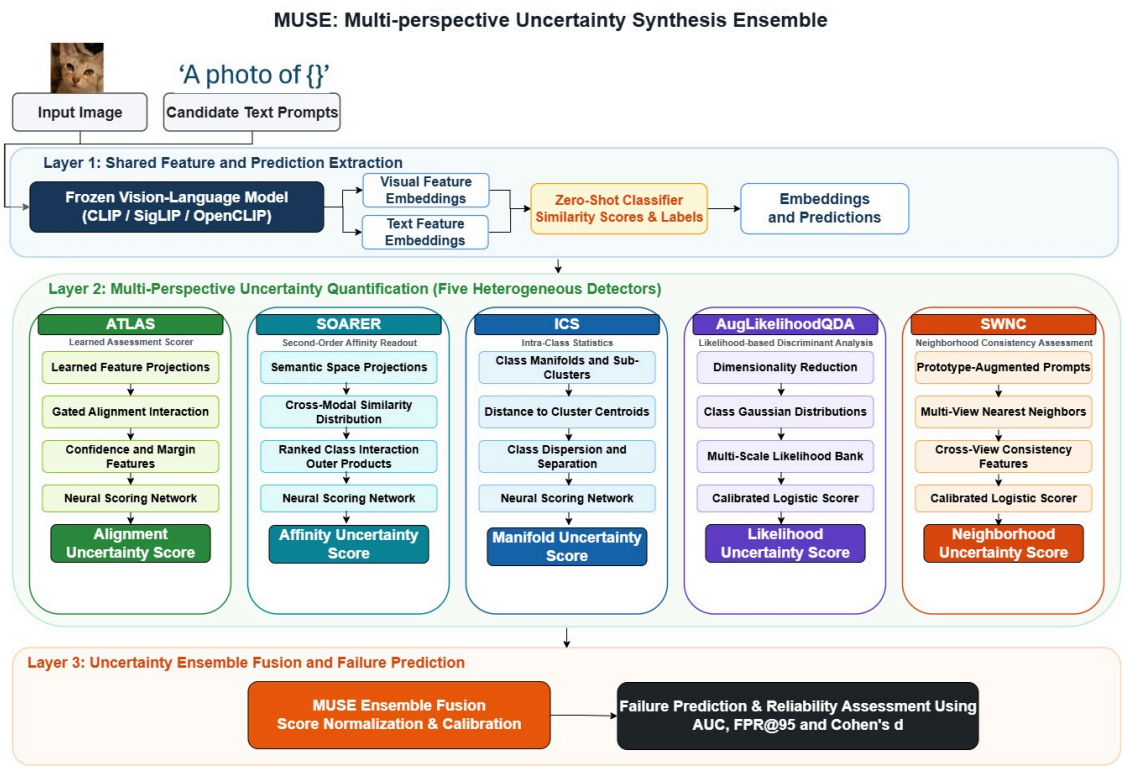


Figure 1: Overview of Failure Prediction Architecture.

System Architecture The framework instantiates five heterogeneous failure-detection models, each embodying a distinct inductive bias about why a zero-shot classification may fail. The system is organized into three layers:

1. **Shared embedding layer.** A frozen VLM provides visual embeddings and text embeddings for class prompts. The target classifier provides predicted classes, probability vectors, and correctness flags on the training pool.

2. **Detection model layer.** Five models consume embeddings and target outputs, each producing a scalar failure score: cross-modal attention alignment (ATLAS), semantic competition geometry (SOARER), intra-class manifold structure (ICS), class-conditional density (AugLikelihoodQDA), and dual-view neighborhood consistency (SWNC).
3. **Evaluation layer.** Each model is validated independently; the best configuration is selected on $\mathcal{D}_{\text{val}}$ and evaluated once on $\mathcal{D}_{\text{test}}$.

4 Proposed Detector Architectures

ATLAS: Alignment-based Text-visual Learned Assessment System. ATLAS is a discriminative neural detector that models cross-modal alignment via learned projections and gating. Its core failure score is computed as

$$f_\theta(\mathbf{v}, \mathbf{T}) = \text{MLP}([\sigma(\langle \mathbf{W}_v\mathbf{v}, \mathbf{W}_t\mathbf{t}_{\hat{y}}\rangle) \cdot \mathbf{W}_v\mathbf{v},\ \mathbf{W}_t\mathbf{t}_{\hat{y}},\ \mathbf{m}]),$$

where $\mathbf{W}_v, \mathbf{W}_t \in \mathbb{R}^{d\times d}$ are projection matrices initialized to identity and frozen for the first n_{freeze} iterations, $\mathbf{t}_{\hat{y}}$ is the predicted-class text embedding, σ is the sigmoid function, $\langle\cdot,\cdot\rangle$ is the inner product, $[\cdot]$ denotes concatenation, and $\mathbf{m} = [1 - p_{\hat{y}},\ p_{\hat{y}} - p_{(2)},\ H(\mathbf{p}),\ p_{\hat{y}},\ p_{(2)},\ \langle \mathbf{v}, \mathbf{t}_{\hat{y}}\rangle]^\top \in \mathbb{R}^6$ is a meta-descriptor summarizing target confidence, with $p_{(2)}$ the second-highest probability and $H(\mathbf{p}) = -\sum_c p_c \log p_c$ the predictive entropy. An auxiliary caption-alignment loss with weight $\lambda_{\text{cap}} = 0.10$ regularizes the projected visual code.

SOARER: Second-Order Affinity-based Reliability Estimation Readout. SOARER is a discriminative neural model that analyzes the full class-similarity distribution. Its failure score is derived from a learned semantic distribution

$$f_\theta(\mathbf{v}, \mathbf{T}) = \text{MLP}([\mathbf{v},\ \mathbf{t}_{\hat{y}},\ \mathbf{d}_s,\ \mathbf{m}]), \quad \mathbf{p}_s = \text{softmax}\left(\frac{(\mathbf{W}_v\mathbf{v})(\mathbf{W}_t\mathbf{T})^\top}{\tau_s}\right),$$

where $\mathbf{W}_v, \mathbf{W}_t \in \mathbb{R}^{d\times d_f}$ are learned bias-free projections to a d_f-dimensional semantic space, $\tau_s > 0$ is a learnable temperature, and $\mathbf{d}_s \in \mathbb{R}^{m^2+m+2}$ is a semantic descriptor comprising the top-m probabilities from $\mathbf{p}_s$, their entropy $H(\mathbf{p}_s)$, margin $p_{s,(1)} - p_{s,(2)}$, and the vectorized outer product (Kronecker product) $\text{vec}(\mathbf{p}_{\text{top}}\mathbf{p}_{\text{top}}^\top)$, equivalently $\mathbf{p}_{\text{top}} \otimes \mathbf{p}_{\text{top}}$, capturing second-order competition structure. The vector $\mathbf{t}_{\hat{y}}$ is the text embedding of the CLIP-predicted class, and $\mathbf{m}$ denotes six frozen CLIP meta features. The inductive bias is that failures are signaled by near-ties or diffuse class competition.

ICS: Intra-Class Statistics. As other detectors work primarily rely on inter-class differences to flag failures, we decided to introduce a detector dependent solely on intra-class statistics. ICS is a manifold-based neural model that evaluates intra-class geometric plausibility. For each class c, a manifold summary is precomputed comprising centroid $\boldsymbol{\mu}_c$, diagonal variance, and J sub-clusters $\{\boldsymbol{\mu}_{c,j}, \sigma_{c,j}, \pi_{c,j}\}_{j=1}^J$ via mini-batch clustering. At inference, the failure score is

$$f_\theta(\mathbf{v}, \mathbf{T}) = \text{MLP}([f_v(\mathbf{v}),\ f_t(\mathbf{t}_{\hat{y}}),\ \mathbf{d}_{\text{inst}},\ \mathbf{d}_{\text{cls}}]),$$

where f_v, f_t are trainable projection networks, $\mathbf{d}_{\text{inst}} \in \mathbb{R}^{J+10}$ is an instance descriptor containing distances to $\boldsymbol{\mu}_{\hat{y}}$, mean distance to k nearest in-class neighbors, and soft sub-cluster assignment $\boldsymbol{\alpha} = \text{softmax}(-[\|\mathbf{v} - \boldsymbol{\mu}_{\hat{y},1}\|/\sigma_{\hat{y},1}, \ldots, \|\mathbf{v} - \boldsymbol{\mu}_{\hat{y},J}\|/\sigma_{\hat{y},J}])$ with its entropy, and $\mathbf{d}_{\text{cls}} \in \mathbb{R}^{2J+3}$ is a fixed class-level summary including sub-cluster radii and inter-sub-cluster separation, with k the neighborhood size and J the number of sub-clusters per class. An auxiliary caption-alignment loss regularizes the projected embeddings. The inductive bias is that misclassifications lie in low-density or sub-cluster-inconsistent regions.

AugLikelihoodQDA: Augmented Likelihood-based Quadratic Discriminant Analysis model. AugLikelihoodQDA is a density-based generative model. It instantiates a Quadratic Discriminant Analysis (QDA) (42) view of the frozen embedding space, treating each class as a Gaussian with its own covariance. Visual embeddings are PCA-reduced to r dimensions, augmented with a bias coordinate $\tilde{\mathbf{x}}_i = [\mathbf{U}_r^\top \mathbf{v}_i; 1] \in \mathbb{R}^{r+1}$, and per-class Gaussians are fit with shrinkage. The failure score is

$$u(\mathbf{v}) = \sigma\big(\mathbf{w}^\top \mathbf{f}_{\text{bank}}(\mathbf{v}) + b\big), \quad \mathbf{f}_{\text{bank}} = [\mathbf{f}_1, \ldots, \mathbf{f}_B], \quad \mathbf{f}_j = \begin{bmatrix} \max_{c\neq\hat{y}} \ell_c - \ell_{\hat{y}} \\ 1 - p_{\hat{y}}^{\text{gen}} \\ \max_c \ell_c - \ell_{\hat{y}} \end{bmatrix},$$

where $\mathbf{U}_r$ is the matrix of top-r principal components, $\ell_c(\tilde{\mathbf{x}}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_c| - \frac{1}{2}(\tilde{\mathbf{x}} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\tilde{\mathbf{x}} - \boldsymbol{\mu}_c)$ is the class-conditional log-likelihood under class c, $\boldsymbol{\Sigma}_c = (1-\alpha)\boldsymbol{\Sigma}_c^{\text{emp}} + \alpha\boldsymbol{\Sigma}_{\text{global}} + \epsilon\mathbf{I}$ is the shrunk covariance with shrinkage coefficient α, $p_{\hat{y}}^{\text{gen}} = \exp(\ell_{\hat{y}})/\sum_{c'}\exp(\ell_{c'})$ is the generative softmax probability, B is the number of QDA configurations (varying r and α), and $\mathbf{w}, b$ are logistic-calibrator parameters fit on $\mathcal{D}_{\text{tr}}$. The inductive bias is that samples unlikely under their predicted-class density are failures.

Algorithm 1 Model-Level Classification Failure Detection

Input: Frozen VLM $\mathcal{F}$, frozen zero-shot target classifier g, class prompts $\mathbf{T}$, trained detector $\mathcal{M}$, input image x, and validation-selected threshold $\theta^\star$
Output: Binary failure flag $\hat{e} \in \{0, 1\}$

$\mathbf{v} \leftarrow \mathcal{F}_v(x)/\|\mathcal{F}_v(x)\|_2$, $\mathbf{s} \leftarrow g(\mathbf{v}, \mathbf{T})$, $\hat{y} \leftarrow \arg\max_c s_c$
if $\mathcal{M} = \text{ATLAS}$ **then**
  $u \leftarrow \text{MLP}([\sigma(\langle \mathbf{W}_v\mathbf{v}, \mathbf{W}_t\mathbf{t}_{\hat{y}} \rangle)\mathbf{W}_v\mathbf{v}, \mathbf{W}_t\mathbf{t}_{\hat{y}}, \mathbf{m}])$
else if $\mathcal{M} = \text{SOARER}$ **then**
  $\mathbf{p}_s \leftarrow \text{softmax}\Big((\mathbf{W}_v\mathbf{v})(\mathbf{W}_t\mathbf{T})^\top/\tau_s\Big)$; construct $\mathbf{d}_s$ from top-m probabilities, entropy, margin, and $\text{vec}(\mathbf{p}_{\text{top}}\mathbf{p}_{\text{top}}^\top)$
  $u \leftarrow \text{MLP}([\mathbf{v}, \mathbf{t}_{\hat{y}}, \mathbf{d}_s, \mathbf{m}])$
else if $\mathcal{M} = \text{ICS}$ **then**
  Compute $\mathbf{d}_{\text{inst}}$ from the predicted-class centroid, k nearest in-class neighbors, and sub-cluster assignments; compute $\mathbf{d}_{\text{cls}}$ from class-level sub-cluster statistics
  $u \leftarrow \text{MLP}([f_v(\mathbf{v}), f_t(\mathbf{t}_{\hat{y}}), \mathbf{d}_{\text{inst}}, \mathbf{d}_{\text{cls}}])$
else if $\mathcal{M} = \text{AugLikelihoodQDA}$ **then**
  Compute class-conditional QDA log-likelihoods $\{\ell_c\}$ and $\mathbf{f}_{\text{bank}}$ from likelihood margins and generative confidence
  $u \leftarrow \sigma(\mathbf{w}^\top\mathbf{f}_{\text{bank}} + b)$
else if $\mathcal{M} = \text{SWNC}$ **then**
  Construct $\tilde{\mathbf{t}}_c$ and retrieve neighborhoods using visual $\mathbf{v}$ and semantic $\mathbf{q}_s$ views; compute neighborhood consistency features
  $u \leftarrow \sigma(\mathbf{w}^\top\mathbf{f}_{\text{SWNC}} + b)$
end if
return $\hat{e} \leftarrow \mathbb{I}[u \geq \theta^\star]$

SWNC: Similarity-Weighted Neighborhood Consistency model. SWNC is a querying-based non-parametric model operating in dual views: visual and semantic. For each class c, a prototype-augmented text embedding $\tilde{\mathbf{t}}_c = \beta\mathbf{t}_c + (1-\beta)\boldsymbol{\pi}_c$ is constructed and normalized, where $\boldsymbol{\pi}_c$ is the mean training embedding in class c and $\beta \in [0, 1]$ is the text-blending weight. The semantic query is

$$u(\mathbf{v}) = \sigma\big(\mathbf{w}^\top\mathbf{f}_{\text{SWNC}}(\mathbf{v}) + b\big), \quad \mathbf{q}_s = \sum_{c=1}^{K} \frac{\exp(\tau_a\langle \mathbf{v}, \tilde{\mathbf{t}}_c \rangle)}{\sum_{c'} \exp(\tau_a\langle \mathbf{v}, \tilde{\mathbf{t}}_{c'} \rangle)} \tilde{\mathbf{t}}_c,$$

where τ_a is the assignment temperature, $\mathbf{q}_s$ is used to retrieve neighbors in the semantic view (visual view retrieves directly with $\mathbf{v}$), and $\mathbf{f}_{\text{SWNC}} \in \mathbb{R}^{2n_{\text{feat}}^{\text{view}}+7}$ concatenates per-view features (predicted-class mass deficit, entropy, margin, miss rate over $k \in \{k_1, \ldots, k_M\}$ neighborhoods with neighbor temperature τ_n) and cross-view features (Jaccard overlap, label agreement), with $n_{\text{feat}}^{\text{view}} = 6M + 8$ per view, where $n_{\text{feat}}^{\text{view}}$ is a scalar feature-count, and it determines the dimensionality of the concatenated detector feature vector, M is the number of k values, and $\mathbf{w}, b$ are the logistic-calibrator parameters. Leave-self-out is enforced during training and validation. The inductive bias is that failures exhibit inconsistent neighborhoods across views.

Algorithm 1 summarizes how the detectors work.

4.1 Cross-Modal Pair Retrieval Failure Detection

In the cross-modal retrieval setting, given a query image $\mathbf{v}_i$ and a gallery of caption embeddings $\mathbf{T}_{\text{gallery}} = \{\mathbf{t}_j\}_{j=1}^{M}$, where M is the number of captions in the split, the target retrieval system returns a ranked list of captions. The top-1 retrieved caption index is $\hat{j}_i = \arg\max_j \langle \mathbf{v}_i, \mathbf{t}_j \rangle$. A retrieval is deemed a failure if the ground-truth paired caption does not appear in the top-k results. We use k = 1. The failure-detection task is to predict which query images will experience retrieval failures. Three of the five models proposed for classification (AugLikelihoodQDA, SWNC, and ICS) are inherently class-based: they rely on per-class manifolds, per-class densities, or per-class neighborhood structures that do not exist in the retrieval setting, which is a class-less setting where the caption space is unstructured and captions are not grouped into discrete categories. Only ATLAS and SOARER are adapted to the retrieval task and are, consequently, the only constituents of the ensemble here.

ATLAS adaptation for retrieval. The ATLAS architecture requires minimal modification. The text bank $\mathbf{T}$ is replaced by the per-split caption gallery $\mathbf{T}_{\text{gallery}} \in \mathbb{R}^{M\times d}$, and the predicted label $\hat{y}$ is replaced by the retrieved caption index $\hat{j}$. The gated interaction becomes

$$\tilde{\mathbf{v}} = \sigma\Big(\langle \mathbf{W}_v\mathbf{v}, \mathbf{W}_t\mathbf{t}_{\hat{j}} \rangle\Big) \cdot \mathbf{W}_v\mathbf{v},$$

where $\mathbf{t}_{\hat{j}}$ is the top-1 retrieved caption embedding. The target-score meta-descriptor is computed over the retrieval probability distribution $\mathbf{p} = \text{softmax}(\tau\mathbf{v}^\top\mathbf{T}_{\text{gallery}}^\top) \in \mathbb{R}^M$, yielding the same six features: $[1 - p_{\hat{j}}, p_{\hat{j}} - p_{(2)}, H(\mathbf{p}), p_{\hat{j}}, p_{(2)}, \langle \mathbf{v}, \mathbf{t}_{\hat{j}} \rangle]$.

SOARER adaptation for retrieval. SOARER similarly replaces the class-text bank with the caption gallery. The semantic similarity distribution is computed as $\mathbf{p}_s = \text{softmax}((\mathbf{W}_v\mathbf{v})(\mathbf{W}_t\mathbf{T}_{\text{gallery}})^\top/\tau_s)$, and the top-$m$ descriptor, entropy, margin, and Kronecker features are extracted identically. The scorer input becomes $[\mathbf{v}, \mathbf{t}_{\hat{j}}, \mathbf{d}_s, \mathbf{m}]$, where all components are defined as in the classification setting but computed over the retrieval distribution.

The retrieval pipeline follows the same three-stage protocol: shared embedding extraction, per-model training and validation-based selection, and single test evaluation. Training labels are derived from ground-truth caption pairings: $e_i = \mathbb{I}[\hat{j}_i \neq j_i^{\text{true}}]$, where j_i^{true} is the index of the true paired caption in the gallery.

5 Results

We compare five proposed failure detectors (ATLAS, SOARER, ICS, AugLikelihoodQDA, SWNC) against baselines and report an ensemble that combines the proposed detectors. Experiments use three frozen zero-shot vision–language backbones (CLIP ViT-B/32, SigLIP ViT-B/16, OpenCLIP ViT-B/32). The metrics are averaged over 5 random seeds per method. We report AUROC and FPR at 95% recall (both as percentages) and Cohen's d (unitless).

5.1 Classification failure detection

Table 1: Failure detection on Easy tasks (Ensemble AUROC > 0.96). Target model: CLIP (ViT-B/32). Best **bold**; second-best underlined. "Acc" = zero-shot top-1 accuracy. AUC = AUROC, FPR = FPR@95, d = Cohen's d. ASONAM17 Name implies the ASONAM17 disaster name task (30).

Method	Flowers-102 (31) *Acc: 63.65%*			EuroSAT (15) *Acc: 42.74%*			ASONAM17 Name *Acc: 70.62%*			UC Irvine (28) *Acc: 43.20%*			Caltech-101 (23) *Acc: 85.99%*			CIFAR-10 (21) *Acc: 88.13%*			UEC Food 100 (43) *Acc: 40.89%*		
	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑
Baselines																					
MCM (27)	85.56	45.23	1.60	74.70	66.77	0.97	75.18	76.06	0.90	65.58	74.54	0.56	91.27	26.81	1.87	89.76	37.26	1.82	81.72	60.77	1.32
LVU (8; 52; 20)	77.77	78.75	1.10	90.19	44.84	1.94	96.26	15.58	3.08	83.23	67.80	1.39	91.98	33.40	2.10	93.60	21.91	2.38	80.23	64.17	1.20
ViLU (22)	95.22	22.12	3.01	98.98	3.09	5.23	98.69	5.86	4.10	96.37	16.01	3.21	95.33	25.43	2.54	97.56	11.41	3.07	87.98	48.26	1.76
Proposed Detectors																					
ATLAS	95.44	22.62	2.75	98.99	4.04	4.91	98.71	6.74	4.24	98.63	9.27	3.96	97.22	13.95	3.33	97.27	13.79	**3.38**	89.40	42.62	1.88
SOARER	92.10	31.96	2.17	98.39	8.20	4.10	98.68	6.10	4.25	98.45	9.89	3.84	95.97	12.98	2.97	96.21	17.82	2.99	87.61	49.94	1.71
QDA	99.68	1.50	5.64	98.85	4.88	4.66	92.73	40.10	2.30	97.46	16.45	3.21	97.70	10.49	3.28	96.53	15.43	3.20	95.99	21.38	2.73
ICS	98.71	7.35	3.99	99.36	1.68	5.94	98.54	5.25	3.99	97.75	11.99	3.49	96.79	18.53	2.76	96.62	15.10	2.84	92.95	32.97	2.19
SWNC	99.50	0.93	**7.01**	99.21	2.20	5.60	98.71	5.59	**4.64**	98.70	6.56	**4.36**	98.15	10.02	3.90	96.80	13.34	3.34	95.97	20.71	**3.05**
Ensemble	**99.75**	**0.87**	**7.01**	**99.62**	**1.17**	**6.67**	**99.05**	**4.07**	**4.64**	**98.97**	**5.69**	**4.36**	**98.69**	**6.97**	**4.38**	**97.73**	**11.10**	**3.38**	**96.73**	**18.27**	**3.05**

Table 2: Failure detection on Medium tasks ($0.935 \leq$ Ensemble AUROC ≤ 0.96). Target model: CLIP (ViT-B/32). Best **bold**; second-best underlined. "Acc" = zero-shot top-1 accuracy. AUC = AUROC, FPR = FPR@95, d = Cohen's d. ASONAM17 dam implies ASONAM17 damage task, where the classes are no, mild and severe damage.

Method	UCF-101 (44) *Acc: 63.86%*			ASONAM17 dam *Acc: 58.12%*			DTD (7) *Acc: 42.04%*			Food-101 (4) *Acc: 79.02%*			Oxford Pets (32) *Acc: 83.15%*			MHII (53) *Acc: 26.44%*			CIFAR-100 (21) *Acc: 64.65%*		
	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑
Baselines																					
MCM (27)	83.07	49.18	1.39	71.10	82.83	0.77	74.00	69.67	0.94	88.02	40.06	1.70	89.11	39.94	1.80	62.27	82.76	0.44	83.32	52.79	1.43
LVU (8; 52; 20)	81.11	53.97	1.27	92.46	30.08	2.32	67.22	84.21	0.65	79.59	60.60	1.16	79.43	65.41	1.17	72.19	76.05	0.86	81.39	58.42	1.29
ViLU (22)	88.11	36.92	1.79	95.43	18.63	2.90	68.66	79.45	0.72	91.48	37.74	2.04	86.64	54.63	1.60	78.02	51.34	1.17	90.88	38.06	1.98
Proposed Detectors																					
ATLAS	92.38	28.01	2.25	95.50	17.99	**2.95**	82.14	52.26	1.33	92.94	28.00	2.35	93.24	27.36	2.34	90.93	36.97	2.12	89.64	40.06	1.91
SOARER	88.61	35.71	1.82	95.52	18.97	2.91	79.58	55.35	1.20	92.11	30.39	2.23	92.84	28.60	2.27	90.40	45.40	2.04	88.56	42.69	1.81
QDA	95.40	19.36	2.62	93.20	31.27	2.42	94.24	27.69	2.32	92.78	28.59	2.32	89.15	42.84	1.80	93.28	29.89	2.51	91.50	32.99	2.11
ICS	92.61	28.73	2.01	94.18	26.80	2.58	90.43	38.10	1.78	90.08	37.30	1.88	88.11	47.30	1.66	85.97	49.62	1.57	89.60	41.67	1.81
SWNC	94.95	20.59	2.71	95.33	19.50	2.91	93.65	29.45	2.43	93.04	28.92	**2.39**	91.01	34.82	2.06	93.95	29.31	**2.58**	92.63	32.01	2.30
Ensemble	**96.22**	**16.38**	**2.74**	**95.83**	**17.09**	**2.95**	**94.94**	**26.27**	**2.50**	**94.85**	**22.07**	**2.39**	**94.59**	**22.51**	**2.60**	**94.58**	**20.11**	2.41	**93.85**	**27.24**	**2.31**

Tables 1–3 report classification failure detection using CLIP as the target backbone over 20 dataset–task settings grouped by difficulty (Easy: 7 tasks, Medium: 7 tasks, Hard: 6 tasks; with difficulty being a function of AUC as defined in table captions). Across these 20 tasks, the mean AUROC values are: MCM 76.87%, LVU 79.79%,

ViLU 86.70%; SOARER 89.34%, ATLAS 90.72%, ICS 90.92%, AugLikelihoodQDA 92.02%, SWNC 93.17%; and Ensemble 94.24%. The average AUROC gain of the best single detector (SWNC) over the strongest baseline (ViLU) is 6.47 percentage points, and the ensemble improves over ViLU by 7.54 percentage points.

The difficulty grouping highlights how performance degrades as failures become less separable. On Easy tasks (Table 1), Ensemble AUROC ranges from 96.73 to 99.75. It's a well-separated regime, and with the SigLIP backbone ICS and AugLikelihoodQDA approach saturation. On Medium tasks (Table 2), it ranges from 93.85 to 96.22. On Hard tasks (Table 3), it ranges from 85.59 to 93.43. On very hard tasks like FGVC Aircraft baselines visibly break down yet the proposed detectors deliver large improvements. The ensemble stays substantially above baselines on the most label-noisy setting (CrisisMMD Informativeness: 86.30 versus 52.08 for MCM). A single-task exception for individual detectors occurs on CIFAR-10 (Table 1), where ViLU reaches 97.56 AUROC and the best single proposed detector reaches 97.27 AUROC, while the ensemble reaches 97.73 AUROC.

Table 3: Failure detection on Hard tasks (Ensemble AUROC < 0.935). Target model: CLIP (ViT-B/32). Best **bold**; second-best underlined. "Acc" = zero-shot top-1 accuracy. AUC = AUROC, FPR = FPR@95, d = Cohen's d. CrisisMMD Info implies Informativness task, CrisisMMD Dmg implies Damage severity task and CrisisMMD Hum implies the Humanitarian categories task on CrisisMMD dataset (2).

Method	Stanford Cars (51) *Acc: 57.18%*			SUN 397 (48) *Acc: 60.81%*			FGVC Aircraft (25) *Acc: 18.15%*			CrisisMMD Info *Acc: 52.44%*			CrisisMMD Dmg *Acc: 43.48%*			CrisisMMD Hum *Acc: 32.38%*		
	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑
Baselines																		
MCM (27)	79.58	56.96	1.23	77.56	65.63	1.09	73.21	69.48	0.89	52.08	92.16	0.08	60.37	92.17	0.37	65.41	85.50	0.55
LVU (8, 52, 20)	64.83	88.60	0.53	70.97	76.49	0.79	67.18	75.65	0.71	73.21	76.07	0.91	67.00	81.88	0.64	78.37	70.72	1.07
ViLU (22)	71.90	73.86	0.84	79.35	60.35	1.16	69.99	71.16	0.82	75.28	74.40	1.02	74.82	68.12	0.94	84.67	66.48	1.47
Proposed Methods																		
ATLAS	83.14	50.88	1.43	84.80	50.46	1.52	78.63	61.26	1.08	85.72	53.71	1.57	85.07	**50.87**	1.51	84.87	65.47	1.50
SOARER	80.38	54.76	1.26	82.04	55.70	1.34	75.92	64.72	1.00	85.88	54.59	1.57	83.79	54.06	1.43	84.76	64.78	1.52
QDA	91.32	34.34	1.91	89.70	41.91	1.78	84.51	46.10	1.38	85.29	56.44	1.54	82.47	57.97	1.36	79.93	70.30	1.22
ICS	87.57	49.26	1.56	85.65	53.15	1.46	80.63	59.15	1.12	85.32	57.77	1.54	81.02	62.32	1.19	81.57	69.61	1.29
SWNC	91.84	31.82	2.14	90.10	40.79	1.97	85.81	47.40	1.60	85.51	56.52	1.57	84.62	58.26	1.50	84.68	**62.85**	1.53
Ensemble	**93.43**	**29.16**	**2.26**	**91.68**	**35.61**	**2.06**	**87.00**	**43.78**	**1.67**	**86.30**	**52.09**	**1.61**	**85.80**	54.35	**1.57**	**85.59**	64.64	**1.60**

We measured per-backbone mean AUROC for each detector over the full dataset–task grid. The SigLIP minus CLIP difference is 1.10 points and OpenCLIP minus CLIP is 0.20 points. Baselines show larger changes across backbones (for example LVU increases from 79.90 to 84.60). High baseline variance indicates that single-perspective baseline detectors are more sensitive to specific embedding geometries and logit scales of the underlying VLM backbones and MUSE's low fluctuation (≤ 0.76 points) demonstrates its backbone-agnostic robustness. The per-backbone mean AUROC table can be found in the supplementary material under Additional Results.

5.2 Cross-modal pair retrieval failure detection

Table 4: Cross-modal retrieval failure detection results with **CLIP** as the target retrieval model. Each block reports one $^{\text{Dataset}}_{split}$ configuration together with the corresponding zero-shot top-1 retrieval accuracy (*Acc*). AUC and FPR@95 are reported in %; Cohen's d is unitless. Best per column is **bold**, second-best is underlined. Zero-shot accuracy is reported over the whole dataset. CrisisMMD official split implies CrisisMMD Informativeness official split.

Method	CrisisMMD *official* *Acc: 18.04*			CrisisMMD *92.5/5.0/2.5* *Acc: 18.04*			CrisisMMD *99/0.5/0.5* *Acc: 18.04*			UC Irvine *70/15/15* *Acc: 25.37*			UC Irvine *92.5/5.0/2.5* *Acc: 25.37*			MHII *official* *Acc: 80.29*			MHII *92.5/5.0/2.5* *Acc: 80.29*		
	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑	AUC↑	FPR↓	d↑
MCM (27)	62.29	81.11	0.498	60.28	82.78	0.422	69.14	59.65	0.713	70.83	80.30	0.795	74.68	84.44	0.973	57.02	86.30	0.240	60.38	80.00	0.386
LVU (8, 52, 20)	77.59	72.91	1.082	75.87	73.63	1.009	67.89	92.98	0.634	60.50	89.91	0.394	62.21	90.54	0.466	75.23	78.48	0.967	78.47	56.60	1.173
ViLU (22)	87.15	45.60	1.681	83.28	47.43	1.469	85.79	54.97	1.516	82.92	53.95	1.350	81.17	59.91	1.236	85.54	49.40	1.520	87.64	42.36	1.531
SOARER	87.38	46.05	1.667	84.23	57.38	1.493	88.73	36.84	1.742	81.53	53.51	1.323	85.21	36.04	**1.642**	83.52	60.51	1.438	84.18	44.79	1.411
ATLAS	87.77	43.38	**1.712**	86.23	48.92	1.613	**90.88**	**30.99**	1.812	82.72	53.62	**1.374**	85.02	**35.59**	1.313	85.39	48.44	1.554	87.36	43.75	1.541
Ensemble	**87.79**	**43.18**	1.695	**86.97**	**47.09**	**1.689**	90.69	35.67	**1.860**	**83.31**	**53.44**	1.357	**85.53**	37.84	1.565	**87.80**	**39.11**	**1.711**	**88.92**	**31.94**	**1.555**

Table 4 reports cross-modal retrieval failure detection results using CLIP. On the CrisisMMD *99/0.5/0.5* split with CLIP, ATLAS improves AUROC from 69.14% to 90.88% and reduces FPR95 from 59.65% to 30.99% relative to MCM (27) (Table 4). While ViLU (22) serves as a strong baseline, our ensemble edges out the best AUROC on both MHII splits with 87.80% and 88.92%, respectively. MCM (27) and LVU (8, 52, 20) alternate as the lowest-performing methods depending on the retrieval configuration. The results show that in low-retrieval-accuracy regimes, the similarity distribution still contains a usable signal that separates correct from incorrect retrievals, and that modeling cross-modal alignment helps extract it.

Multiple splits were tried because tweet image–text retrieval is highly sensitive to which posts land in train vs. test, real-world human colloquialisms (slang, abbreviations, sarcasm), missing context, and event-specific phrasing can change difficulty a lot across partitions. Multiple crisis datasets have frequent modality mismatch (image and text only loosely related, or one modality dominates), so a single split can over- or under-estimate reliability depending on how many mismatched pairs end up in the test set. Using multiple splits may be able to check that failure prediction is robust to these real-world variations rather than tuned to one particular partition.

5.3 Qualitative Results

We compute Spearman rank correlation (ρ) to assess agreement between detector failure rankings. It is observed that ATLAS and SOARER form a neural block that with a relatively higher correlation, while AugLikelihoodQDA, SWNC, and ICS exhibit lower or intermediate correlations with this block ($\rho = 0.632$–0.839), indicating complementary failure signals. These correlation patterns prove heterogeneity of the detectors and supports their use in the ensemble. Detailed correlation results are provided in Subsection 10.3 of the Supplementary Material.

Further qualitative results include concrete instances (see Figure 2) where MUSE succeeded in classifying as CLIP failure but baselines could not.



(a) True class: American Pit Bull Terrier



(b) True class: American Pit Bull Terrier



(c) True Class: Beagle



(d) True class: Bengal

Figure 2: Some images where CLIP fails in zero-shot settings. MUSE flags the failures correctly, but baselines fail.

5.4 Ablation Studies

We perform component-wise ablations for the five detectors, with all deltas computed against each model's Full Reference result. The results reveal distinct sensitivity profiles. SWNC is robust to most single-component removals (ΔAUC between -0.11% and -0.70%), but degrades sharply when restricted to semantic-only inputs (ΔAUC $= -5.43\%$, ΔFPR $= +19.54\%$), showing that visual and semantic neighborhoods act as complementary signals. AugLikelihoodQDA depends critically on covariance regularization: most component removals induce moderate drops (ΔAUC $\in [-3.6\%, -4.3\%]$), but forcing full shrinkage (α=1) collapses the likelihood ratio and reduces performance to chance (AUC $= 50\%$, FPR@95 $= 100\%$). ATLAS shows a narrow, distributed sensitivity (most ΔAUC $\in [-2.0\%, -2.5\%]$), indicating that its scoring, gating, freezing, and auxiliary caption loss contribute jointly rather than through any dominant component; notably, removing CLIP meta-features has a small AUC effect (-0.97%) but a large FPR@95 effect ($+13.03\%$), affecting the high-recall operating point more than global ranking.

SOARER is moderately sensitive to most components (ΔAUC up to -2.9%) but degrades substantially when the predicted-class text embedding is removed (ΔAUC $= -8.70\%$, ΔFPR $= +14.94\%$), identifying it as SOARER's central semantic anchor. ICS is the most sensitive detector overall: removing the balanced sampler, intra-class cosine signal, caption auxiliary loss, KNN features, subcluster representation, or freezing strategy produces ΔAUC between -4.4% and -11.2%, with the largest FPR@95 increase ($+26.24\%$) from removing subcluster information, indicating strong reliance on the joint interaction of local structure, auxiliary supervision, and training schedule. Detailed ablation study results can be found in Subsection 10.7 in the Supplementary Material.

6 Discussion

Foundational vision–language datasets such as CC3M (41), CC12M (5), LAION-400B (38), and LAION-5B (37) are heavily processed for machine learning and primarily emphasize literal image–text alignment, whereas multimodal tweet datasets such as CrisisMMD, UC Irvine Multimodal, and MHII reflect more contextual and colloquial human communication. Thus, when tweet text and images are not directly aligned, zero-shot retrieval becomes a pragmatic grounding problem, often producing low top-1 accuracy and diffuse similarity distributions in which failure information is not captured by the maximum similarity alone. This behavior is reflected in CrisisMMD, where CLIP under the 99/0.5/0.5 split improves from 85.90% AUROC and 50.42% FPR95 with MCM to 90.88% AUROC and 30.99%

FPR95 with ATLAS, indicating that distributional and cross-modal alignment cues can provide substantial information beyond top-1 confidence (Table 4). In contrast, on MHII, where retrieval accuracy is approximately 80% across backbones, MCM performs strongly in several settings, suggesting that maximum similarity is more informative when image–text correspondence is closer to literal description. No individual detector dominates all datasets, splits, and backbones. This reflects substantial heterogeneity in tweet-level image–text relationships. Some configurations seem to favor ATLAS or SOARER, while others favor MCM or the Ensemble. Across all 21 retrieval configurations, the Ensemble achieves the strongest aggregate performance with 85.87% mean AUROC, 46.19% mean FPR95, and $d = 1.575$, while ATLAS provides the strongest mean performance among individual detectors on CLIP and SigLIP. These results suggest that retrieval failure prediction benefits from combining complementary confidence, distributional, and cross-modal alignment signals rather than relying exclusively on the top-1 similarity score.

This discussion has been covered in greater detail in Section 11 in the Supplementary Material.

6.1 Limitations

MUSE is not universally applicable across task formats. Three of our classification detectors, AugLikelihoodQDA, ICS, and SWNC, are fundamentally class-conditional and do not directly transfer to the class-less cross-modal retrieval setting. MUSE still requires a labeled training pool with observed correctness/failure outcomes to fit discriminative detectors and calibrators, so its reliability can degrade under very severe domain shift when the failure mechanisms differ from those seen during detector training. Some components, such as covariance estimation for AugLikelihoodQDA, may introduce some non-trivial overhead, and the ensemble can, particularly on retrieval tasks, occasionally underperform the best constituent on a specific split because fusion trades peak performance for average robustness. In retrieval settings, although ATLAS and SOARER achieve stronger overall failure detection, they require backbone- and dataset-specific training on success/failure labels induced from zero-shot predictions. When successes or failures are scarce, these induced labels can produce a highly unilaterally dominated training distribution, increasing the risk of overfitting and limiting generalization. In such settings, MCM (27), which is entirely training-free, may sometimes achieve stronger performance by avoiding this label-induced training bias while directly exploiting the model's confidence structure. MCM thus provides a deterministic, parameter-free fallback when failure labels or retraining are unavailable, whereas trained detectors can provide additional failure-discriminative capacity when sufficient labeled data are available. MUSE is designed for failure detection, not failure correction. In safety-critical settings (disaster triage, medical imaging), the primary operational need is reliable flagging of untrustworthy predictions for human-in-the-loop escalation or abstention. Failure correction is a valuable but distinct research direction (17), left to future work.

7 Future Directions

Despite our extensive study, MUSE can be potentially explored across a wider spectrum of architectures for exhaustiveness. These include efficiency-oriented backbones like TinyCLIP (47), scaling-optimized models like MetaCLIP (49), and long-context variants like LongCLIP (55), as well as additional fine-grained visual benchmarks such as Stanford Dogs (19) and NABirds (46). Beyond general visual recognition, extending MUSE to domain-specific, safety-critical distributions, such as aerial disaster assessment (e.g., xBD (14), FloodNet (35)), multimodal emergency dispatch (12), and clinical medical imaging (45), presents a promising frontier for estimating reliability assurance with MUSE. Further, extending MUSE to generative and instruction-tuned VLMs would test whether the same multi-perspective principle transfers to open-ended outputs such as captions or answers, where failure is not a single label mismatch. Moving from static post hoc scoring to online, streaming reliability estimation (similar to autonomous vehicles) may enable real-time use in disaster-response pipelines, including adaptive triage and human-in-the-loop escalation. Richer fusion mechanisms, such as calibrated meta-learners(40) and uncertainty-aware routing (36), may further exploit the complementarity between detectors. Coupling MUSE with active data collection and abstention policies could translate improved failure prediction into measurable downstream gains in decision quality for high-stakes visual reasoning.

8 Reproducibility

All experiments use five fixed seeds ($\{42–46\}$) with deterministic PyTorch, NumPy, and cuDNN settings. Frozen backbones (CLIP, SigLIP, OpenCLIP) use publicly released weights, and features are pre-extracted with content-hashed caches. Splits are held constant across seeds and backbones; hyperparameter search is confined to train/validation splits, with the uncontaminated test set reserved for reporting. Full reproducibility details, including details on datasets, training, hyperparameters, ensembling, and compute, are provided in Supplementary Material (Section 9, specifically subsections 9.14 and 9.15). The code is available at https://anonymous.4open.science/r/MUSE-EB9D/.

REFERENCES

[1] Lemar Abdi, Amaan Valiuddin, Francisco Caetano, Christiaan Viviers, and Fons van der Sommen. Zero-shot image anomaly detection using generative foundation models. In *Workshop on Anomaly Detection with Foundation Models, ICCV*, 2025. arXiv:2507.22692.

[2] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2018. arXiv:1805.00713.

[3] Anton Baumann, Rui Li, Marcus Klasson, Santeri Mentu, Shyamgopal Karthik, Zeynep Akata, Arno Solin, and Martin Trapp. Post-hoc probabilistic vision-language models. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2412.06014.

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3568, 2021. arXiv:2102.08981.

[6] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, 2023. arXiv:2212.07143.

[7] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[8] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[9] Eduardo Dadalto, Marco Romanelli, Georg Pichler, and Pablo Piantanida. A data-driven measure of relative uncertainty for misclassification detection. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.01710.

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

[11] European Commission. Article 14: Human oversight | eu artificial intelligence act. EU Artificial Intelligence Act, July 2024.

[12] Pablo Ferri, Carlos Sáez, Antonio Félix-De Castro, Purificación Sánchez-Cuesta, and Juan M. García-Gómez. An end-to-end solution for out-of-hospital emergency medical dispatch triage based on multimodal and continual deep learning. *Artificial Intelligence in Medicine*, 170:103264, 2025.

[13] Federica Granese, Marco Romanelli, Daniele Gorla, Catuscia Palamidessi, and Pablo Piantanida. Doctor: A simple method for detecting misclassification errors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.02395.

[14] R. Gupta, R. Hosfelt, S. Sajeev, N. Patel, B. Goodman, J. Doshi, E. Heim, H. Choset, and M. Gaston. xbd: A dataset for assessing building damage from satellite imagery. 2019. 10.48550/arXiv.1911.09296.

[15] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. Zenodo, 2018.

[16] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017. arXiv:1610.02136.

[17] Zhuo Huang, Chang Liu, Yinpeng Dong, Hang Su, Shibao Zheng, and Tongliang Liu. Machine vision therapy: Multimodal large language models can enhance visual robustness via denoising in-context learning. In *International Conference on Machine Learning (ICML)*, 2024.

[18] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, 2016.

[19] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization: Stanford dogs. 2012.

[20] Michael Kirchhof, Mark Collier, Seong Joon Oh, and Enkelejda Kasneci. Pretrained visual uncertainties. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25950–25960, 2024. arXiv:2402.16569.

[21] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[22] Marc Lafon, Yannis Karmim, Julio Silva-Rodríguez, Paul Couairon, Clément Rambour, Raphaël Fournier-Sniehotta, Ismail Ben Ayed, Jose Dolz, and Nicolas Thome. Vilu: Learning vision-language uncertainties for failure prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17807–17817, 2025.

[23] Fei-Fei Li, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 178, 2004.

[24] Jie Ma, Zongxi Liu, and Yi Zhu. Failsae: Towards interpretable failure prediction for vision-language models via sparse autoencoders, 2026.

[25] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[26] A. Mashrur, W. Luo, N. A. Zaidi, and A. Robles-Kelly. Robust quantification of prediction uncertainty by inducing heterogeneity in deep ensembles. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2023.

[27] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. arXiv preprint, 2022.

[28] Hussein Mouzannar, Yara Rizk, and Mariette Awad. Multimodal damage identification for humanitarian computing. In *UCI Machine Learning Repository*, 2018.

[29] National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). NIST Artificial Intelligence Framework AI 100-1, National Institute of Standards and Technology, 2023.

[30] Dat T. Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit Mitra. Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 569–576, 2017.

[31] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[32] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3498–3505, 2012.

[33] V. Penza, X. Du, D. Stoyanov, L. S. Mattos, and E. De Momi. Long term safety area tracking (lt-sat) with online failure detection and recovery for robotic minimally invasive surgery. *Medical Image Analysis*, 45, 2018.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. arXiv:2103.00020.

[35] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. 2020. 10.48550/arXiv.2012.02951.

[36] Ralf Römer, Adrian Kobras, Luca Worbis, and Angela P. Schoellig. Failure prediction at runtime for generative robot policies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 25278–25294, 2022. arXiv:2210.08402.

[38] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. In *NeurIPS Workshop on Data-Centric AI*, 2021. arXiv:2111.02114.

[39] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[40] T. V. Shareena and S. Padmavathi. Deep meta-learner stacking with explainable ai for vision-based plastic waste classification. *SN Computer Science*, 7(7):761, 2026.

[41] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2556–2565, 2018.

[42] Cedric A. B. Smith. Some examples of discrimination. *Annals of Eugenics*, 13(1):272–282, 1946.

[43] Elena Battini Sönmez, Sefer Memiş, Berker Arslan, and Okan Zafer Batur. The segmented UEC food-100 dataset with benchmark experiment on food detection. *Multimedia Systems*, 29(4):2049–2057, 2023.

[44] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. Technical Report CRCV-TR-12-01, Center for Research in Computer Vision, University of Central Florida, 2012. arXiv:1212.0402 [cs.CV].

[45] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6:1399–1406, 2022. 10.1038/s41551-022-00936-9.

[46] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. G. Ipeirotis, P. Perona, and S. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. pages 595–604, 2015.

[47] K. Wu, H. Peng, Z. Zhou, B. Xiao, M. Liu, L. Yuan, H. Xuan, M. Valenzuela, X. Chen, X. Wang, H. Chao, and H. Hu. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. 2023. 10.48550/arXiv.2309.12314.

[48] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492, 2010.

[49] H. Xu, S. Xie, X. E. Tan, P. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, and C. Feichtenhofer. Demystifying clip data. 2023. 10.48550/arXiv.2309.16671.

[50] Jie Yan, Xiaohang Guan, Wei-Shi Zheng, Hong Chen, and Rui Wang. Global and local vision-language alignment for few-shot learning and few-shot ood detection. In *Lecture Notes in Computer Science (LNCS)*, volume 15964. Springer, 2026.

[51] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3973–3981, 2015. arXiv:1506.08959.

[52] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[53] C. Yu, B. Hu, and Z. Wang. Open-world disaster information identification from multimodal social media. *Complex & Intelligent Systems*, 11(1), 2024.

[54] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023. arXiv:2303.15343.

[55] B. Zhang, P. Zhang, X. Dong, Y. Zang, and J. Wang. Long-clip: Unlocking the long-text capability of clip. 2024. 10.48550/arXiv.2403.15378.

[56] Shu Zou, Xin Tian, Qi Zhao, Zhen Yang, and Jie Zhang. Simlabel: Consistency-guided ood detection with pretrained vision-language models. In *Lecture Notes in Computer Science (LNAI)*, volume 16370. Springer, 2026.