EdgeV-SE: Self-Reflective Fine-tuning Framework for Edge-Deployable Vision-Language Models

Yoonmo Jeon , Seunghun Lee , Woongsup Kim *

Posted Date: 24 December 2025

doi: 10.20944/preprints202512.2100.v1

Keywords: Vision-Language Model (VLM); edge computing; self-reflective learning; consistency regularization; mutual learning; satellite iot; nvidia jetson; disaster analysis

 doi:10.20944/preprints202512.2100.v1

Article

EdgeV-SE: Self-Reflective Fine-tuning Framework for Edge-Deployable Vision-Language Models

Yoonmo Jeon, Seunghun Lee and Woongsup Kim *

Department of Information and Communication Engineering, Dongguk University, Seoul 04620, Republic of Korea

* Correspondence: woongsup@dongguk.edu

Featured Application

The proposed framework enables the deployment of robust Vision-Language Models on resource-constrained off-the-shelf edge devices, such as the NVIDIA Jetson series. Its primary application is real-time disaster damage assessment using satellite imagery in communication-denied environments, facilitating immediate decision-making for first responders.

Abstract

The deployment of Vision-Language Models (VLMs) in Satellite IoT scenarios is critical for real-time disaster assessment but is often hindered by the substantial memory and compute requirements of state-of-the-art models. While parameter-efficient fine-tuning (PEFT) enables adaptation, with minimal computational overhead, standard supervised methods often fail to ensure robustness and reliability on resource-constrained edge devices. To address this, we propose EdgeV-SE, a self-reflective fine-tuning framework that significantly enhances the performance of VLM without introducing any inference-time overhead. Our framework incorporates an uncertainty-aware self-reflection mechanism with asymmetric dual pathways: a generative linguistic pathway and an auxiliary discriminative visual pathway. By estimating uncertainty from the linguistic pathway using a log-likelihood margin between class verbalizers, EdgeV-SE identifies ambiguous samples and refines its decision boundaries via consistency regularization and cross-pathway mutual learning. Experimental results on hurricane damage assessment demonstrate that our approach improves image classification accuracy, enhances image–text semantic alignment, and achieves superior caption quality. Notably, our work achieves these gains while maintaining practical deployment on a commercial off-the-shelf edge device such as NVIDIA Jetson Orin Nano, preserving the inference latency and memory footprint. Overall, our work contributes a unified self-reflective fine-tuning framework that improves robustness, calibration, and deployability of VLMs on edge devices.

Keywords: Vision-Language Model (VLM); edge computing; self-reflective learning; consistency regularization; mutual learning; satellite iot; nvidia jetson; disaster analysis

1. Introduction

With the increasing frequency and intensity of natural disasters due to climate change [1], rapid and accurate analysis of satellite imagery for damage assessment has become critically important. In large-scale hurricane or flood events, communication infrastructure is often damaged or unavailable, and network connectivity to ground stations is frequently intermittent. These disaster-zone conditions pose significant challenges for cloud-based AI, particularly in terms of latency, bandwidth, and availability. As a result, Intelligent satellite IoT [2], where satellites, high-altitude platforms, or unmanned aerial vehicles are paired with on-board or nearby edge devices that perform local inference, has emerged as a practical alternative for real-time disaster response.

 doi:10.20944/preprints202512.2100.v1

In this context, vision–language models (VLMs) are particularly attractive. A single VLM can both classify damage severity and generate human-readable descriptions of the scene, providing interpretable evidence (e.g., "standing water covering roads", "intact roofs and dry ground") that is valuable for analysts and first responders. However, most state-of-the-art VLMs are designed for server-class GPUs with abundant memory, rendering them too resource-intensive for direct deployment on small edge devices in Satellite IoT scenarios. For example, popular open-source models such as LLaVA-1.5-7B **[3]** require server-class GPUs with substantial memory footprints, particularly in FP16 settings. In addition, top-tier commercial models like GPT-4V **[4]** have undisclosed architectures and model scales and are deployed via cloud-based inference, which precludes practical on-device deployment.

The primary goal of our research is to develop a high-performance VLM that can efficiently run on commercial off-the-shelf edge devices, such as the NVIDIA Jetson Orin Nano **[5]**, which has been widely adopted for on-device deep learning applications **[6]**. Jetson Orin Nano provides only 8 GB of unified LPDDR5 memory and modest compute throughput, even though it is specifically designed as an edge-inference module. This fundamental resource mismatch makes it infeasible to deploy such large VLMs directly on edge devices, motivating the use of more compact yet expressive architectures together with intelligent fine-tuning strategies. In this work, we focus on how to fine-tune a mid-sized VLM so that it becomes both reliable and edge-deployable, rather than further compressing the architecture itself. Among existing models, we choose BLIP-Large **[7]** as a representative encoder–decoder VLM that is close to the upper limit of what can be deployed on commercial off-the-shelf edge devices, such as the NVIDIA Jetson series, in FP16.

For complex and specialized tasks such as hurricane damage assessment from satellite imagery, characterized by scarce labeled data and subtle visual differences between damage and no-damage classes, standard supervised fine-tuning (SFT) alone is often data-inefficient and insufficient to deliver the robustness and reliability required in real-world settings. SFT optimizes a token-level or label-level cross-entropy objective, which primarily rewards alignment with the ground-truth target. As training progresses, samples that the model already fits well quickly contribute diminishing gradients [8], even though their predictions can remain fragile under minor distribution shifts or visually subtle variations. Moreover, SFT does not explicitly encourage the model to assess the reliability of its own decisions or to focus learning on borderline cases that matter most in low-label regimes. These characteristics motivate training strategies that complement scarce labels with additional, self-generated signals such as consistency regularization and agreement-based learning.

To address these challenges, we adopt a self-reflective fine-tuning strategy that augments SFT with uncertainty-aware, self-generated learning signals, without changing the model size or adding inference-time overhead. In contrast to recent self-correction approaches that rely on iterative prompting at inference time, which incurs additional latency, our method embeds the reflective process directly into the model parameters during training. This approach allows the model to benefit from self-evaluation without incurring any runtime computational overhead on resource-constrained edge devices. To this end, we propose EdgeV-SE (Edge-deployable Vision–Language Models using Self Evaluation and Self Enhancement), a self-reflective fine-tuning framework designed to enable models to identify their own uncertainty and reconcile internal inconsistencies between complementary pathways. Conceptually, EdgeV-SE operationalizes self-reflection [9,10] within an encoder–decoder VLM by turning the discrepancy between a generative linguistic pathway and a discriminative visual pathway into explicit learning signals.

Empirically, our results show that the proposed framework substantially improves both classification accuracy and caption quality without introducing any inference-time overhead on commercial edge devices such as the Jetson Orin Nano [5].

The main contributions of our work are as follows:

- We propose a self-reflective fine-tuning framework, EdgeV-SE, that enables the model to recognize its own uncertainty and learn by resolving internal inconsistencies.

 doi:10.20944/preprints202512.2100.v1

- We introduce an efficient mechanism that enhances prediction reliability with minimal overhead by designing asymmetric dual pathways, a generative linguistic path and a discriminative visual path, within the VLM and performing internal cross-validation through mutual learning [11].
- We demonstrate the practical effectiveness of our approach by empirically validating the proposed model's superior classification accuracy and inference efficiency on an actual edge device [5].

The remainder of this paper is organized as follows. Section 2 reviews related work on (i) efficient vision–language models for edge environments, (ii) semi-/self-supervised learning, (iii) self-reflective learning, and (iv) uncertainty-aware learning and calibration. Section 3 presents the proposed self-reflective fine-tuning framework, EdgeV-SE, detailing its asymmetric dual pathways, margin-based uncertainty diagnosis, and the consistency and mutual-learning objectives. Section 4 describes the experimental setup and implementation, including dataset construction, training configuration, evaluation protocol, and on-device benchmarking. Section 5 reports the experimental results and analysis, including ablations and comparisons with representative fine-tuning baselines. Section 6 discusses practical considerations for deployment, robustness and reliability under domain shift, and remaining limitations. Finally, Section 7 concludes the paper and outlines future research directions.

2. Related Works

2.1. Efficient VLMs for Edge Environments

Research on deploying VLMs on edge devices has primarily focused on model compression techniques such as Quantization [12], semantic-aware token pruning [13], and Knowledge Distillation [14]. More recently, efforts have been made to design lightweight architectures specifically for the edge, such as MobileVLM [15] and edge-cloud collaborative approaches for vision-language models [16]. While effective, these approaches often involve a significant trade-off in accuracy or require redesigning and pre-training models from scratch. An alternative is Parameter-Efficient Fine-Tuning (PEFT) methods like LoRA [17] and its variant QLoRA [18], which freeze the pre-trained weights and train only a small number of adaptive modules to improve memory efficiency. Our work adopts QLoRA [18] for training efficiency, but our core contribution lies not in optimizing trainable parameters but in enhancing the sample efficiency and robustness of the fine-tuning process itself.

2.2. Semi-Supervised and Self-Supervised Learning

Semi-Supervised Learning (SSL) aims to improve model performance by leveraging a small amount of labeled data along with a large amount of unlabeled data [19]. Consistency regularization, a key principle in SSL, posits that a model's prediction should remain invariant to non-essential perturbations of its input, such as data augmentation. The effectiveness of this principle has been demonstrated in seminal works like FixMatch [20] and Unsupervised Data Augmentation (UDA) [21], as well as in methods like Mean Teacher [22].

Meanwhile, Self-Supervised Learning generates supervisory signals from the data itself to guide the learning process, as seen in contrastive methods like MoCo [23] and SimCLR [24]. Our work applies these principles to the fine-tuning stage, aligning with the philosophy of self-supervision by using internally generated feedback as a learning signal without requiring extra labeled samples beyond the SFT dataset.

 doi:10.20944/preprints202512.2100.v1

2.3. Self-Reflective Learning

In recent deep learning research, there has been a growing interest in the ability of models to review and revise their own predictions, a concept known as 'self-reflection' or 'self-correction' [9,10]. This is often studied in the context of Large Language Models (LLMs) identifying logical fallacies in their generated text or improving conclusions through multi-step reasoning processes like Chain-of-Thought [25] and Self-Consistency [26]. Furthermore, Self-Distillation [27], where a model uses its own past predictions (soft labels) as targets for the next generation of learning, shares a similar philosophy with our mutual learning mechanism [11].

EdgeV-SE introduces a novel approach to applying self-reflection within the vision-language fine-tuning process. While contemporary LLMs often perform self-correction using iterative linguistic prompts, this method is less feasible for encoder-decoder architectures like BLIP-Large [7], which are not inherently designed for complex instructional feedback loops.

2.4. Uncertainty-Aware Learning and Calibration

Reliable deployment in resource-constrained or safety-critical environments depends not only on accuracy but also on how well a model's confidence reflects its probability of correctness. Prior work has studied predictive uncertainty in deep networks through approximate Bayesian approaches such as Monte Carlo dropout, which interprets dropout as approximate Bayesian inference and enables uncertainty estimation without modifying the architecture [28]. Ensemble-based methods provide another practical alternative, showing that independently trained model ensembles yield strong uncertainty estimates and often improve calibration, including under distribution shift [29]. In addition, confidence calibration has been systematically analyzed in modern neural networks, where post-hoc methods such as temperature scaling were shown to be simple yet highly effective, and Expected Calibration Error (ECE) has become a standard diagnostic metric [30].

Orthogonally, a related line of research focuses on emphasizing hard or ambiguous samples during training. Online Hard Example Mining (OHEM) selects high-loss examples to concentrate optimization on difficult cases [31], while focal loss reshapes cross-entropy to down-weight easy samples and focus gradients on hard negatives [8]. While aligned with prior directions, EdgeV-SE differs in that it internally estimates uncertainty and selectively gates self-reflective objectives between two pathways, without requiring inference-time sampling or model duplication.

3. Self-Reflective Fine-Tuning (EdgeV-SE)

EdgeV-SE augments standard supervised fine-tuning (SFT) with a training-time self-reflective mechanism derived from the model's internal disagreement. The training update for each mini-batch proceeds in four phases: (1) discrepancy induction, (2) uncertainty diagnosis, (3) discrepancy resolution, and (4) supervised consolidation (Figure 1; Algorithm 1).

Phase 1 — Discrepancy induction creates complementary predictions from asymmetric linguistic vs. visual pathways to expose internal disagreement.

Phase 2 — Uncertainty diagnosis estimates sample uncertainty via a margin-based self-assessment and computes an uncertainty weight.

Phase 3 — Discrepancy resolution focuses learning on uncertain samples by enforcing augmentation consistency and cross-pathway agreement.

Phase 4 — Supervised consolidation aggregates supervised losses and reflection losses to update parameters.

The first three phases generate and refine self-produced training signals, while the final phase performs the standard supervised update to preserve captioning ability and align the model with the downstream damage classification objective. Phase 4 supervised consolidation is not merely a collection of independent regularization terms. Instead, phase 4 treats internal linguistic–visual disagreement as a learnable uncertainty signal that directly guides optimization.

 doi:10.20944/preprints202512.2100.v1

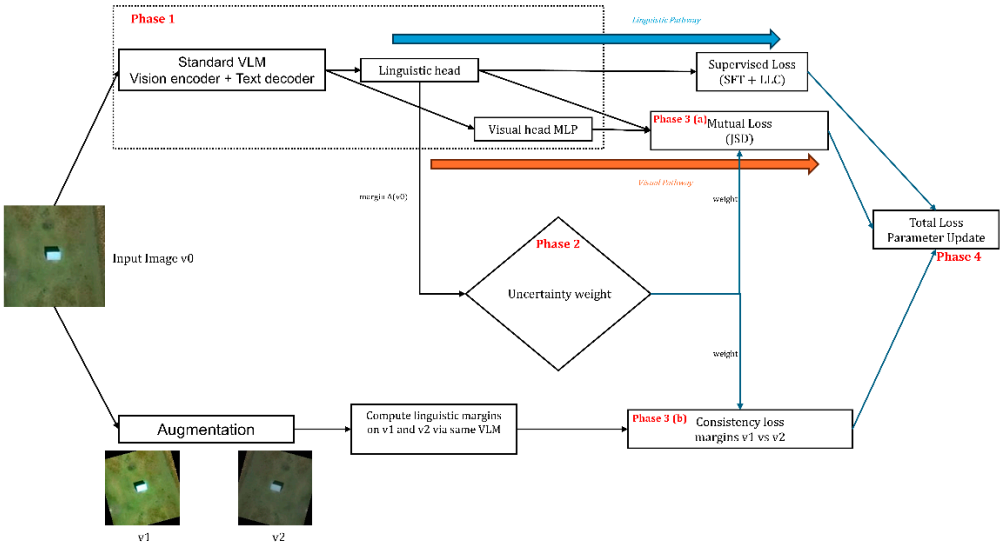

Figure 1. EdgeV-SE: Overall architecture of EdgeV-SE, showing the dual-path (Generative vs. Discriminative) and the consistency-guided mutual reflection process.

Importantly, the visual pathway is employed only during training and is not activated during inference. Inference relies solely on the original VLM pathway, incurring no additional runtime overhead on edge devices.

3.1. Discrepancy Induction: Asymmetric Dual Pathways

Standard vision–language models often conflate semantic plausibility with visual evidence, which can lead to overconfident predictions in visually ambiguous scenarios [32]. To explicitly disentangle these factors, we decompose the model into two asymmetric prediction pathways from distinct viewpoints, a generative viewpoint and a discriminative viewpoint, representing complementary but potentially conflicting perspectives.

The *linguistic pathway*, corresponding to a generative view, produces class evidence through the text decoder conditioned on the image, capturing high-level semantics. In parallel, a lightweight *visual pathway*, corresponding to discriminative view, predicts the same classes directly from vision encoder features, emphasizing low-level visual cues. Since these pathways emphasize complementary cues, their inevitable disagreement on ambiguous samples serves as an intrinsic signal for self-reflective learning.

(a) Linguistic Pathway (Generative Viewpoint)

The linguistic pathway functions as the Generative Branch (acting as a 'Theorist'), interpreting high-level semantics such as debris, flooding, or roof collapse. Given an input image, it evaluates the likelihood of each class by conditioning on prompt-based class descriptions and subsequently applies Softmax normalization to obtain image-conditioned class probabilities.

Given an original image v_0, we score each class verbalizer/prompt y_t using the decoder likelihood:

$$l_c(v_0) = logP_\theta(y_t \mid v_0), \tag{1}$$

where K is the number of classes to identify.

We convert these scores to class probabilities using a temperature-scaled Softmax normalization:

$$p_L(c \mid v_0) = \frac{\exp\left(\frac{l_c(v_0)}{T}\right)}{\sum_{k=0}^{K-1} \exp\left(\frac{l_k(v_0)}{T}\right)}, \tag{2}$$

 doi:10.20944/preprints202512.2100.v1

where T is temperature, $l_c(v_0)$ is the decoder log-likelihood for class c, and $p_L(c \mid v_0)$ is the image-conditioned class probability estimated from the linguistic pathway.

(b) Visual Pathway (Discriminative Viewpoint)

The visual pathway functions as the Discriminative Branch (acting as an "Empiricist"), predicting class probabilities directly from visual encoder, focusing on visual details.

Let $h(v_0)$ be the vision encoder representation of the original image. A lightweight head produces logits and probabilities:

$$z_V(v_0) = \mathrm{MLP}\big(h(v_0)\big), \quad p_V(k|v_0) = \mathrm{softmax}\big(z_V(v_0)\big)_k \tag{3}$$

3.2. Recognizing Uncertainty: Identifying 'Uncertain' Samples via Self-Diagnosis

This phase quantifies the uncertainty of the class prediction for a given image by measuring the margin between the highest and second-highest class probabilities. To quantify its own uncertainty, the model measures a prediction margin from the linguistic pathway:

$$\Delta(v_0) = l_c(v_0) - max_{k \neq c} l_k(v_0) \tag{4}$$

where c denotes the class with the maximum predicted probability for image v_0.

In our main experiments, the downstream task is binary (damage vs. no-damage) and the linguistic pathway uses two class verbalizers c_1 and c_0. In this case, the top-1 vs. runner-up margin in Eq. (4) reduces to the two-way log-likelihood gap $| l_1(v_0) - l_0(v_0) |$. For implementation convenience (Algorithm 1; Figure 2), we compute the signed margin $\Delta(v_0) = l_1(v_0) - l_0(v_0)$ and apply the uncertainty gate using its magnitude $| \Delta(v_0) |$ in Eq. (5), while the sign of $\Delta(v_0)$ is later used as decision evidence in the LL-margin classifier (Section 4.2).

Furthermore, uncertainty diagnosis is performed on the original view v_0, whereas $\Delta(v_1)$ and $\Delta(v_2)$ are computed only for the augmentation-consistency objective in Eq. (6).

A small absolute margin $\Delta(v_0)$ implies indecision or conflict between the two outcomes, corresponding to an ambiguous or high-uncertainty sample. This margin explicitly identifies samples where semantic priors and visual evidence diverge, which are precisely the cases where standard supervised fine-tuning becomes unreliable under limited supervision. Uncertainty diagnosis is based on the margin from the linguistic pathway because it directly reflects inference-time decision confidence, whereas the visual pathway is employed solely as an auxiliary during training.

The model assigns a higher learning weight to such uncertainty based on a threshold τ, as schematically illustrated in Figure 2:

$$w_{\text{uncert}}(v_0) = \begin{cases} \lambda_{\text{uncert}}, & if |\Delta(v_0)| < \tau \\ 1, & if |\Delta(v_0)| \geq \tau \end{cases} \tag{5}$$

where λ_{uncert} is a tunable hyperparameter that amplifies the learning signal for "uncertain" samples identified by the model itself, while confident samples receive a standard weight of 1. The concrete values of τ and λ_{uncert} are selected on a validation set and reported in Section 4.2. This discrete weighting scheme, aligned with Algorithm 1, encourages the model to focus on its own weaknesses.

 doi:10.20944/preprints202512.2100.v1

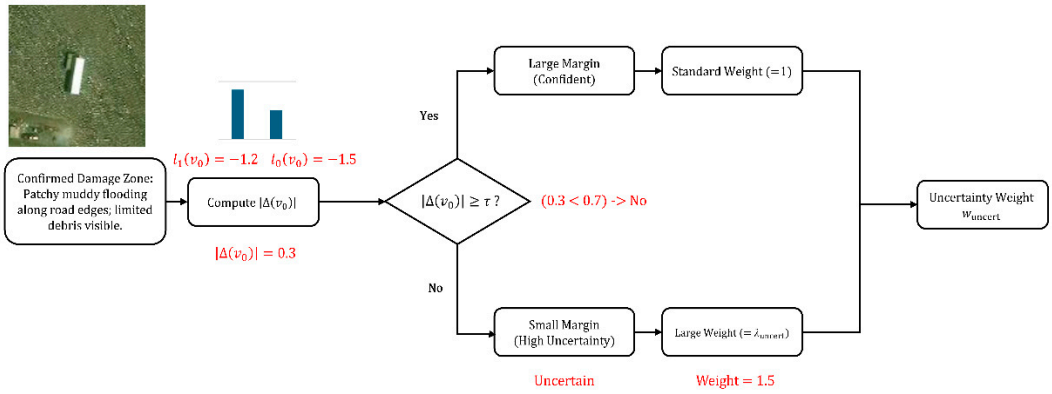

Figure 2. Schematic illustration of the linguistic margin-based self-diagnosis mechanism. The model calculates the margin $\Delta(v_0)$ between the damage verbalizer (l_1) and no-damage verbalizer (l_0) predictions. If the margin is below the threshold τ (e.g., 0.7), the sample is diagnosed as uncertain, and a boosted weight(λ_{uncert}) is assigned for training.

3.3. Discrepancy Resolution: Internalizing Knowledge through Consistency and Mutual Consensus

Once samples are identified as uncertain, EdgeV-SE resolves the internal discrepancies by applying weighted Consistency Regularization and Mutual Learning. These objectives are weighted by w_{uncert} calculated from the original view (v_0).

(a) Consistency Regularization

Given an input image v_0, EdgeV-SE constructs two stochastically augmented views $v_1 = aug_1(v_0)$ and $v_2 = aug_2(v_0)$ using mild spatial/photometric transformations. Then, EdgeV-SE enforces consistency based on their prediction margins, weighted by the uncertainty of v_0 :

$$\mathcal{L}^{w}_{\text{consist}}(x) = w_{\text{uncert}}(v_0) \cdot \big(\Delta(v_1) - \Delta(v_2)\big)^2 \tag{6}$$

This regularization encourages the decision boundary to remain stable under perturbations such as changes in viewpoint or illumination. While consistency regularization promotes invariance, it alone does not determine which prediction pathway should be trusted when their outputs disagree.

To address this limitation, EdgeV-SE couples consistency with margin-based disagreement and mutual learning, enabling consistency to be selectively enforced on uncertain samples. Consistency is defined on the logit margin rather than on raw probability outputs, since the margin serves as a more direct proxy for decision boundary stability and prediction confidence, whereas probability vectors are more susceptible to temperature scaling and calibration effects.

(b) Mutual Learning ($\mathcal{L}^{w}_{\text{mutual}}(x)$)

Here, the Linguistic Pathway ("Theorist") and the Visual Pathway ("Empiricist") act as soft targets for each other. This objective regularizes the linguistic pathway to remain visually grounded (reducing semantic drift or hallucination) while distilling high-level semantics into the visual pathway.

We quantify the cross-pathway agreement using the Jensen–Shannon divergence (JSD) between p_L and p_V **[33]**. JSD is symmetric and bounded, so $0 \leq \text{JSD}(p_L, p_V) \leq \ln 2$, and remains finite even when supports differ, yielding stable gradients for alignment. We adopt JSD because it is symmetric, bounded, and numerically stable when the two pathways assign divergent support, unlike KL which can diverge. For readability, qualitative figures report normalized values $\text{JSD}/\ln 2 \in [0,1]$.

We define the mutual loss as a weighted JSD between p_L and p_V (Eq. (7)), implemented as the mean of two KL terms to the mixture m:

$$\mathcal{L}^{w}_{\text{mutual}}(x) = w_{\text{uncert}}(v_0) \cdot \tfrac{1}{2}[KL(p_L \parallel m) + KL(p_V \parallel m)], m = \tfrac{1}{2}(p_L + p_V). \tag{7}$$

 doi:10.20944/preprints202512.2100.v1

Through this cooperative process, linguistic semantics are distilled into visual recognition, while visual grounding prevents linguistic hallucinations.

3.4. Supervised Consolidation: Aggregation of Supervised and Self-Reflective Losses to Update Parameters

The total objective integrates four components, standard supervised fine-tuning, auxiliary classification, consistency, and mutual learning, applied to the same mini-batch in each iteration (Algorithm 1). The uncertainty-weighted consistency and mutual terms act as auxiliary stabilizers that selectively reinforce reflection on ambiguous samples, thereby improving both convergence and interpretability.

(a) Supervised fine-tuning loss from generated caption

$$\mathcal{L}_{\mathrm{SFT}}(x) = CE(\hat{y}(v_0), y^{cap}), \tag{8}$$

where $\hat{y}(v_0)$ denotes the model's predicted token distribution (logits) under teacher forcing for the primary view v_0, and y^{cap} is the ground-truth caption. This is the standard token-level cross-entropy loss that preserves the model's image–text generation ability.

(b) Auxiliary classification loss from linguistic pathway

$$\mathcal{L}_{\mathrm{LLC}}(x) = CE(l(v_0), y^{cls}), \tag{9}$$

where $y^{cls} \in$ {damage,no-damage} is the binary class label (denoted as c_i in Algorithm 1). This loss directly supervises the two-way log-likelihoods from the linguistic pathway, sharpening its decision margin and tying the generative branch to the downstream damage-classification objective.

(c) Self-Reflective objectives

The self-reflective terms are given by Eqs. (6)–(7):

$$\mathcal{L}^{w}_{\mathrm{consist}}(x) = w_{\mathrm{uncert}}(v_0)\,(\Delta(v_1) - \Delta(v_2))^2, \tag{10}$$

$$\mathcal{L}^{w}_{mutual}(x) = w_{\mathrm{uncert}}(v_0) \cdot \frac{1}{2}[KL(p_L \parallel m) + KL(p_V \parallel m)], m = \frac{1}{2}(p_L + p_V). \tag{11}$$

The consistency term enforces a stable linguistic margin across augmentations, ensuring that the decision boundary is robust to benign perturbations. The mutual learning term encourages agreement between the linguistic and visual predictions, keeping the linguistic theorist visually grounded while injecting high-level semantics into the visual empiricist.

In both cases, the uncertainty weight $w_{\mathrm{uncert}}(v_0)$ amplifies gradients on uncertain samples with small $|\Delta(v_0)|$, so the model spends more capacity on introspectively correcting its own mistakes.

(d) Overall Training Objective

We combine all components into the final training objective:

$$\boxed{\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{SFT}} + \lambda_{\mathrm{LLC}}\,\mathcal{L}_{\mathrm{LLC}} + \lambda_{\mathrm{consist}}\,\mathcal{L}^{w}_{\mathrm{consist}} + \lambda_{\mathrm{mutual}}\,\mathcal{L}^{w}_{\mathrm{mutual}}} \tag{12}$$

In practice, each term is averaged over the mini-batch as in Algorithm 1. The uncertainty weight is computed per sample from the original view v_0 and applied consistently across both self-reflective objectives.

We set the weighting coefficients $\lambda_{\mathrm{consist}}$, $\lambda_{\mathrm{mutual}}$, λ_{LLC}, as well as the uncertainty parameters $\lambda_{\mathrm{uncert}}$, τ, based on validation performance and stability; their concrete values are summarized in Section 4.2. The coefficients are chosen to ensure that no individual objective dominates the training process, while appropriately amplifying the contribution of uncertain samples.

The complete training procedure is summarized in Algorithm 1. It details each mini-batch update, including uncertainty weighting, consistency enforcement, mutual agreement between the two pathways, and the final supervised caption and classification steps.

 doi:10.20944/preprints202512.2100.v1

Algorithm 1: EdgeV-SE Training Procedure (per mini-batch)

Input: Model parameters θ, visual head H, batch $\mathcal{B} = \{(v_{0,i}, y_i, c_i)\}_{i=1}^{B}$
Input: Hyperparameters $\tau, T, \lambda_{\text{uncert}}, \lambda_{\text{consist}}, \lambda_{\text{mutual}}, \lambda_{\text{LLC}}$
Output: Updated parameters θ, H

for $i = 1$ to B **do**
  1. Discrepancy Induction (Dual Pathways) on original $v_{0,i}$
  $h_i \leftarrow \text{VisionEncoder}(v_{0,i})$
  Compute linguistic logits $\{l_0(v_{0,i}), l_1(v_{0,i})\}$ ▷ *Eq. (1)*
  $p_i^L \leftarrow \text{Softmax}([l_0(v_{0,i}), l_1(v_{0,i})]/T)$ ▷ *Theorist, Eq. (2)*
  $\Delta_i \leftarrow l_1(v_{0,i}) - l_0(v_{0,i})$ ▷ *Margin, Eq. (4)*
  $p_i^V \leftarrow \text{Softmax}(H(h_i))$ ▷ *Empiricist, Eq. (3)*
  2. Uncertainty Diagnosis (Self-Diagnosis) on original $v_{0,i}$
  $w_i^{\text{uncert}} \leftarrow \mathbb{I}[|\Delta_i| < \tau] \cdot \lambda_{\text{uncert}} + \mathbb{I}[|\Delta_i| \geq \tau] \cdot 1$ ▷ *Eq. (5)*
  3. Discrepancy Resolution via Augmentation Consistency & Consensus
  $v_{1,i} \leftarrow \text{Augment}_1(v_{0,i}),\ v_{2,i} \leftarrow \text{Augment}_2(v_{0,i})$ ▷ *two stochastic views*
  Compute linguistic logits $\{l_0(v_{1,i}), l_1(v_{1,i})\}$
  $\Delta_i^{(1)} \leftarrow l_1(v_{1,i}) - l_0(v_{1,i})$
  Compute linguistic logits $\{l_0(v_{2,i}), l_1(v_{2,i})\}$
  $\Delta_i^{(2)} \leftarrow l_1(v_{2,i}) - l_0(v_{2,i})$
  $L_{\text{consist},i} \leftarrow w_i^{\text{uncert}} \cdot (\Delta_i^{(1)} - \Delta_i^{(2)})^2$ ▷ *Eq. (6)*
  $m_i \leftarrow 0.5 \cdot (p_i^L + p_i^V)$
  $L_{\text{mutual},i} \leftarrow w_i^{\text{uncert}} \cdot \frac{1}{2}[\text{KL}(p_i^L \| m_i) + \text{KL}(p_i^V \| m_i)]$ ▷ *Eq. (7)*
end for
4. Parameter Update
$L_{\text{SFT}} \leftarrow \frac{1}{B}\sum_i \text{CrossEntropy}(\text{Model}(v_{0,i}), y_i)$ ▷ *y_i: caption tokens, Eq. (8)*
$L_{\text{LLC}} \leftarrow \frac{1}{B}\sum_i \text{CrossEntropy}([l_0(v_{0,i}), l_1(v_{0,i})], c_i)$ ▷ *Eq. (9)*
$L_{\text{total}} \leftarrow L_{\text{SFT}} + \lambda_{\text{LLC}} L_{\text{LLC}} + \lambda_{\text{consist}} \frac{1}{B}\sum_i L_{\text{consist},i} + \lambda_{\text{mutual}} \frac{1}{B}\sum_i L_{\text{mutual},i}$
Update θ, H by minimizing L_{total}

4. Experimental Setup and Implementation

4.1. Dataset and Preprocessing

We conducted a comprehensive set of experiments to validate the proposed EdgeV-SE framework in comparison with both standard and self-improving baselines.

We employed a balanced subset of a hurricane damage assessment dataset consisting of satellite images categorized into two classes: *damage* and *no_damage* [34]. Since the original dataset provides only image-level labels (damage vs. no_damage) and lacks human-written captions, we generated domain-specific pseudo-reference captions using LLaVA-1.5-7B [3]. These captions serve as weak supervision for caption fine-tuning and as a standardized reference for text-metric reporting. They should not be interpreted as human-verified ground truth.

We used two classes of prompts (Table 1), one for damage and one for no-damage, to encourage factual, concise descriptions in a consistent style. To estimate the pseudo-caption noise rate, we manually audited ≈200 randomly sampled images; ≈8% were judged noisy. These samples were excluded from the dataset prior to the train/validation/test split, and inter-annotator agreement exceeded 90%, with disagreements resolved by discussion.

To reduce the risk of over-interpreting style-matching as "factual improvement," we report (i) discriminative classification metrics (Accuracy/F1) and (ii) a reference-free image–text alignment proxy (CLIPScore), in addition to overlap-based caption metrics (CIDEr-D, BERTScore).

The final curated dataset contains 5,000 damage/5,000 no-damage images for training, 1,000/1,000 images for validation, and 1,000/1,000 images for testing, resulting in 14,000 images in total balanced across the two classes.

 doi:10.20944/preprints202512.2100.v1

Table 1. Prompts with a structured instruction.

Category	Prompt Template
Damage Zone	SYSTEM: You are a remote sensing analyst for a disaster relief agency. Your task is to describe satellite images with factual, concise language for an automated damage assessment system. CONSTRAINTS 1. Your entire response MUST start exactly with "Confirmed Damage Zone:". 2. Include specific evidence of damage, such as "visible flooding", "damaged buildings", or "scattered debris". 3. Describe the overall state of the area, focusing on the extent of impact. 4. The response must be a single, coherent paragraph. USER: <image>Analyze the provided satellite image for clear signs of hurricane damage.
No Damage Zone	SYSTEM: You are a remote sensing analyst for a disaster relief agency. Your task is to describe satellite images with factual, concise language for an automated damage assessment system. CONSTRAINTS 1. Your entire response MUST start exactly with "No Damage Zone:". 2. Confirm the absence of damage by describing visible signs of normalcy, such as "intact roofs", "clear streets", or "dry ground". 3. Describe the overall state of the area, confirming stability and functionality. 4. The response must be a single, coherent paragraph. USER: <image>Analyze the provided satellite image to confirm the absence of hurricane damage.

For all caption-quality metrics, we remove the leading class header phrase (e.g., "Confirmed Damage Zone:" or "No Damage Zone:") to avoid coupling the scores to fixed prompt formatting.

LLaVA-1.5 is used solely as an offline annotation tool to provide weak caption supervision and standardized textual references for evaluation. Accordingly, classification metrics (Accuracy and F1) are treated as the primary indicators of task performance, while caption-related metrics (CIDEr-D, BERTScore, and CLIPScore) are reported as complementary measures of descriptive consistency and image-text alignment rather than as human-verified ground truth.

4.2. Model Configuration and Training Details

We use the Salesforce/blip-image-captioning-large [7] model as the base Vision-Language Model (VLM). For parameter-efficient fine-tuning, all methods except Standard SFT used the same QLoRA [18] setting (r = 16, α = 32, dropout = 0.05) applied to selected text layers and the last two vision transformer blocks. Standard SFT uses a text-only LoRA configuration with a higher-rank adapter (r = 128, α = 256), while keeping the same base model, data, optimizer, and decoding settings for a controlled comparison. The optimization was performed using AdamW8bit with a cosine learning-rate scheduler and a warm-up ratio of 0.1. Learning rates were set separately for text and vision modules (1×10^{-4} and 3×10^{-5}, respectively), with weight decay = 0.01. The model was trained for 12 epochs with a batch size of 8 and gradient accumulation of 4. Models were trained on an NVIDIA RTX 4090 GPU, while all on-device performance benchmarks (e.g., inference speed) were conducted on an NVIDIA Jetson Orin Nano (8 GB) device [5] to verify edge-level deployability.

The main hyperparameters for the self-reflective fine-tuning were empirically set as $\tau = 0.7$, $\lambda_{\text{uncert}} = 1.5$, $\lambda_{\text{consist}} = 0.3$, $\lambda_{\text{mutual}} = 0.2$, $\lambda_{\text{LLC}} = 0.3$. For the temperature scaling in Eq. (2), we fix the temperature to T = 2.0 across all experiments. For caption metrics we use beam search (num_beams=4, do_sample=False, max_new_tokens=120). For on-device benchmarks (Section 5.3), we report caption-generation latency using num_beams=1 and max_new_tokens=40. For classification, we use the

 doi:10.20944/preprints202512.2100.v1

validation-selected LL-margin threshold (maximizing Macro-F1), stored as the optimal threshold parameter.

For the linguistic pathway, we compute length-normalized log-likelihoods for two class verbalizers c_1(damage) and c_0(no-damage) (set to the header strings in Table 1, i.e., "Confirmed Damage Zone:" and "No Damage Zone:") using the decoder likelihood $l_k(v) = \log P_\theta(c_k \mid v)$. Following prior prompt-based scoring practice, we normalize by the number of tokens in c_k (excluding special tokens) to avoid length bias. The log-likelihood margin $\Delta(v) = l_1(v) - l_0(v)$ is used as decision evidence, and the binary decision threshold is selected on the validation set (maximizing Macro-F1) and stored as the optimal threshold parameter. Unless stated otherwise, the temperature in Eq. (2) is fixed to $T = 2.0$.

The discriminative pathway is implemented as a lightweight head on top of the vision encoder output. We intentionally designed this head as a shallow MLP to minimize the training-time memory footprint, ensuring that the self-reflective mechanism remains computationally efficient even during fine-tuning. Concretely, we apply (average) pooling over the final-layer vision tokens to obtain a single feature vector, followed by a two-layer MLP $(d_{vis} \to d_h \to 2)$with ReLU to produce logits $z_V(v)$and probabilities p_V. This head is used only during training to generate self-reflective signals and is removed at inference, so runtime latency/memory remain identical to the base VLM.

We generate two independent augmented views using mild spatial/photometric transforms: random horizontal flip (p=0.5), random rotation (±15 degrees), and color jitter (brightness/contrast/saturation = 0.2/0.2/0.2). These augmentations are applied independently to v_1 and v_2.

4.3. Evaluation Metrics

To comprehensively assess both visual and linguistic performance, we evaluate the model using classification and captioning metrics. For classification, we report Accuracy, class-wise F1, and Macro-F1, along with Expected Calibration Error (ECE) for calibration analysis (Section 5.5). Caption quality is assessed using three complementary metrics: (1) CIDEr-D [35], which evaluates caption fluency and consensus; (2) BERTScore [36] (F1 x100), which measures semantic alignment with the ground-truth text; and (3) CLIPScore [37], computed as cosine similarity ×100 using CLIP ViT-B/32, which evaluates the factual consistency (groundedness) of generated captions with respect to the source image. To ensure statistical reliability, all reported metrics are accompanied by 95% confidence intervals, estimated via 1,000 bootstrap iterations on the test set.

5. Experimental Results and Analysis

5.1. Component-Wise Analysis and Ablation Study

To investigate how self-reflective mechanisms progressively enhance the model's reasoning ability, we conducted a detailed study of the internal variants of EdgeV-SE. Tables 2 and 3 summarize the performance at each phase of this developmental trajectory.

5.1.1. Self-Reflection Variants

During the initial exploration stage, we sequentially introduced two versions of Self-Reflection to encourage autonomous improvement:

- **Self-Reflection type 1 Generate-then-Correct:** This variant adopted an iterative mechanism where the model first produced preliminary captions and then re-generated refined outputs conditioned on its own previous text. While it aimed to simulate self-revision, this approach yielded only modest improvement (F1 = 0.918) due to the lack of explicit numerical feedback and stability control.
- **Self-Reflection type 2 Direct LL + Multitask:** To address the limitations of v1, this version incorporated direct log-likelihood (LL) supervision and a multitask alignment loss. This

 doi:10.20944/preprints202512.2100.v1

architectural modification replaces heuristic regeneration with a differentiable JSD-based coupling between the linguistic distribution p_L and visual distribution p_V. This shift yielded a significant performance jump (Macro-F1 0.964), establishing the foundational backbone for the final EdgeV-SE framework.

5.1.2. Objective-Level Ablation Study

Building on the foundation of Self-Reflection type 2, we conducted an ablation study to isolate the contributions of the specific reflective components, Consistency Regularization and Mutual Learning. We evaluate two ablated variants (Model A: SFT + Consistency; Model B: SFT + Mutual Learning) and the full EdgeV-SE model to isolate the contribution of each component:

- **Model A SFT + Consistency**: This variant enforces prediction invariance across multiple augmented visual views (e.g., random flip, rotation and photometric jitter). By regularizing the decision boundary against distributional noise, Model A achieved an F1 score of 0.966, demonstrating that consistency regularization effectively mitigates overfitting and improves spatial robustness. This result indicates that consistency regularization alone stabilizes training, but it does not fully resolve cross-modal disagreement without the mutual-learning component.
- **Model B SFT + Mutual Learning:** This variant introduces a cross-pathway mutual agreement loss between the vision encoder and text decoder representations. This loss encourages both branches to produce congruent feature semantics, enabling uncertainty-aware refinement during back-propagation. Model B achieved an F1 score of 0.981, outperforming consistency-only training and underscoring the role of cross-pathway agreement.
- **EdgeV-SE:** EdgeV-SE framework combines both Consistency and Mutual Learning under a unified reflective optimization. This configuration achieves the best overall results (0.985 F1 for Damage, 0.986 F1 for No-Damage), demonstrating a clear synergistic effect that enhances not only classification precision but also caption coherence and factual grounding.

Table 2. Evolution and Ablation (Class-wise F1).

Phase	Model	Damage F1	No-Damage F1	Notes
Variants	Self-Reflection type 1	0.918 ± 0.004	0.918 ± 0.004	Generate-then-Correct
Variants	Self-Reflection type 2	0.964 ± 0.003	0.964 ± 0.003	Direct LL + Multi-task
Ablation	Model A	0.966 ± 0.003	0.966 ± 0.003	SFT + Consistency
Ablation	Model B	0.981 ± 0.002	0.980 ± 0.002	SFT + Mutual Learning
Ours	**EdgeV-SE**	**0.985 ± 0.004**	**0.986 ± 0.002**	Combine all

Note: Values represent Mean ± 95% Confidence Interval.

A similar trend was observed in caption generation performance, highlighting that the proposed self-reflective mechanisms improve not only classification accuracy but also descriptive expressiveness.

As shown in Table 3, a significant performance leap occurs with Self-Reflection type 2 (CIDEr-D 37.49, BERTScore 90.77), which incorporated direct log-likelihood supervision and a multi-task alignment loss (JSD-based) that couples $p_L$and p_V. This model establishes a strong foundation for both high-accuracy classification (F1 = 0.964) and high-quality caption generation.

The subsequent ablation models (Model A, B) and the final EdgeV-SE model primarily enhance classification F1 accuracy (Table 2) while sustaining this high level of caption quality. For instance, while Model B achieves a major classification jump to 0.981 F1, it maintains a high-quality caption profile (CIDEr-D 38.00, BERTScore 90.60), demonstrating that the mutual learning component enhances discriminative reasoning without causing linguistic degradation.

EdgeV-SE model achieves the highest traditional fluency (CIDEr-D 38.37) and semantic alignment (BERTScore 90.82). While Self-Reflection type 2 showed the highest image-text consistency (CLIPScore 28.79), our model's optimization toward the classification F1 score (0.985) appears to

 doi:10.20944/preprints202512.2100.v1

create a slight trade-off, marginally reducing this groundedness metric in favor of superior discriminative accuracy.

In summary, EdgeV-SE not only achieves the highest classification accuracy but also demonstrates superior language generation ability, producing domain-consistent, and context-aware captions even in edge-constrained environments.

Building upon these internal developments, we next position EdgeV-SE against strong fine-tuning and preference-optimization baselines in Section 5.2.

Table 3. Captioning Performance for Progressive Model Variants.

Phase	Model	CIDEr-D (↑)	BERTScore (↑)	CLIPScore (↑)	Notes
Variants	Self-Reflection type 1	26.08 ± 0.62	88.05 ± 0.15	28.46 ± 0.18	Generate-then-Correct
Variants	Self-Reflection type 2	37.49 ± 0.55	90.77 ± 0.11	28.79 ± 0.14	Direct LL + Multi-task
Ablation	Model A	38.14 ± 0.48	90.79 ± 0.09	28.51 ± 0.13	SFT + Consistency
Ablation	Model B	38.00 ± 0.50	90.60 ± 0.10	28.55 ± 0.12	SFT + Mutual Learning
Ours	EdgeV-SE	38.37 **± 0.42**	90.82 **± 0.08**	28.21 **± 0.11**	Combine all

5.2. Comparison with Advanced Fine-Tuning Baselines

To comprehensively evaluate the effectiveness of the proposed EdgeV-SE, we compared it against four key baselines representing distinct learning paradigms: (1) Standard SFT, (2) Controlled SFT (a LoRA-optimized variant), (3) Self-Rewarding VLM **[38]**, and (4) Iterative Self-Correction **[9,10]**. Our goal is to compare EdgeV-SE against methods that similarly aim to improve the model through self-generated signals within an efficient fine-tuning process. We do not include direct preference optimization **[39]** methods in the main comparison because they require additional preference data (e.g., preference pairs) and typically introduce a separate alignment stage beyond the single-stage PEFT fine-tuning budget considered in this work. Our focus is on training-time, self-generated learning signals under a comparable supervision and compute budget for models intended for edge inference; preference-optimization pipelines are complementary and left for future work.

We first established strong supervised PEFT baselines under the same data split, optimizer, and decoding settings. Standard SFT is a supervised baseline that trains text-only LoRA adapters with a higher-rank setting (r = 128, α = 256) on the BLIP text decoder, while keeping the base model frozen. Controlled SFT uses the same supervised objective and identical data/decoding setup, but employs a lower-rank QLoRA setting (r = 16, α = 32) and expands the trainable scope to include selected text layers plus the last two vision transformer blocks, yielding a more balanced parameter update.

Importantly, the Trainable Params (%) in Table 4 reflects both (i) the adapter scope and (ii) the LoRA rank; therefore, despite covering both modalities, Controlled SFT can have a smaller trainable-parameter fraction due to its substantially lower rank. As shown in Table 4, Controlled SFT improves Macro-F1 from 0.911 to 0.930 compared to Standard SFT. However, its gains plateau under purely supervised PEFT, motivating mechanisms that explicitly handle uncertainty and self-correction.

 doi:10.20944/preprints202512.2100.v1

Table 4. Baseline Classification Performance (Standard SFT vs. Controlled SFT).

Method	LoRA Scope	LoRA (r, α)	Trainable Params (%)	Damage F1	No-Damage F1	Macro-F1	Notes
Standard SFT	Text layers only	(128,256)	14.860	0.912 ± 0.006	0.910 ± 0.007	0.911 ± 0.007	High-rank only LoRA
Controlled SFT	Text + last 2 Vision Blocks	(16,32)	1.330	0.932 ± 0.004	0.928 ± 0.005	0.930 ± 0.005	Low-rank cross-modal LoRA

Note: Trainable Params (%) depends on both the adapter scope and the LoRA rank; Standard SFT uses a higher rank than Controlled SFT. All results are reported with 95% bootstrap CIs (1,000 iterations).

Next, we compared EdgeV-SE against two advanced self-improvement paradigms. Since original works like Self-Rewarding [38] and Reflexion [9] target decoder-only LLMs via prompting, we implemented custom adaptations suitable for the BLIP encoder-decoder architecture:

- **Self-Rewarding VLM:** This method generates multiple candidate captions via stochastic sampling and selects the one with the highest internal generative confidence (log-likelihood) as a pseudo-label for training.
- **Iterative Self-Correction:** This method performs a second "correction" forward pass to refine the initial output, explicitly training the model to produce a higher-confidence version of its first prediction.

Table 5. and 6 present the comparative results. While both self-improvement baselines (Macro-F1 0.979 and 0.972, respectively) significantly outperform the SFT baselines, EdgeV-SE achieves the best Macro-F1 (0.985). This demonstrates that our dual-pathway reflective mechanism is more effective than single-pathway confidence maximization.

Table 5. Comparative Classification Performance.

Method	Evaluation Mechanism	Macro-F1	Remarks
Standard SFT	Cross-Entropy	0.911 ± 0.007	Weak SFT Baseline
Controlled SFT	Cross-Entropy	0.930 ± 0.005	Strong SFT Baseline
Self-Rewarding VLM	LL-based Reward Selection	0.979 ± 0.003	Internal Reward Optimization
Iterative Self-Correction	Uncertainty-based Refinement (Adapted)	0.972 ± 0.004	Progressive Self-Correction
EdgeV-SE	Dual Pathways + Consistency + Mutual Learning	0.985 **± 0.002**	Self-Reflective Optimization

A critical distinction emerges in caption quality (Table 6). While the advanced baselines successfully improved classification, their gains in caption quality were limited. Notably, their CLIPScores (26.28 and 24.30) are the lowest among all methods, suggesting that optimizing a single internal confidence signal may not translate to better image–text alignment, and can be associated with semantic drift. In sharp contrast, EdgeV-SE achieves the highest scores across CIDEr-D (38.37), BERTScore (90.82), while maintaining a competitive CLIPScore (28.21) comparable to the strong SFT baseline. Crucially, while other methods tend to optimize for a single internal reward signal, EdgeV-SE's dual-pathway consistency enforces a robust consensus between linguistic and visual modalities, preventing semantic drift.

Table 6. Comparison of Captioning Performance (Across Methods).

Method	CIDEr-D (↑)	BERTScore (↑)	CLIPScore (↑)	Key Characteristics
Standard SFT	29.71 ± 0.58	86.18 ± 0.18	27.94 ± 0.20	Basic Supervised Learning
Controlled SFT	31.77 ± 0.52	87.50 ± 0.14	28.24 ± 0.16	LoRA Optimization
Self-Rewarding VLM	32.60 ± 0.65	88.23 ± 0.12	26.28 ± 0.22	Internal Reward Selection
Iterative Self-Correction	29.58 ± 0.61	88.19 ± 0.15	24.30 ± 0.25	Progressive Refinement

 doi:10.20944/preprints202512.2100.v1

EdgeV-SE (Ours)	38.37 **± 0.42**	90.82 **± 0.08**	28.21 **± 0.11**	Consistency + Mutual Learning

Finally, regarding computational efficiency, EdgeV-SE achieves these performance gains without iterative generate-and-select loops or multi-round correction at training time. Instead, EdgeV-SE relies on lightweight self-reflective losses, margin consistency and cross-pathway agreement, under the same PEFT training schedule. These results highlight a fundamental advantage of our approach: while other methods tend to over-optimize a single internal reward signal, improving classification at the expense of caption quality, EdgeV-SE's dual-pathway consistency forces a robust consensus between linguistic and visual modalities.

Consequently, EdgeV-SE provides a practical single-pass alternative that delivers strong classification performance while maintaining competitive caption quality, without the additional training-time compute required by iterative generation or evaluation schemes.

5.3. On-Device Performance

We verified that the performance gains of EdgeV-SE are obtained without additional inference-time overhead compared to a parameter-efficient baseline (Controlled SFT) on the target edge device. We benchmarked our model on an NVIDIA Jetson Orin Nano (8 GB) in FP16 precision with batch = 1. To reflect deployment practice, inputs are resized to a short-side of 384, decoding uses num_beams = 1 and max_new_tokens = 40, and we report per-image latency after a 3-step warm-up over a 10-image set (balanced 5/5 damage vs. no-damage).

Table 7 summarizes latency, throughput, and peak unified memory usage. EdgeV-SE runs at 1696.9 ms per image (≈ 1.70 s), 0.589 FPS, and with a peak memory allocation of 0.915 GB. Because EdgeV-SE leaves the inference stack unchanged (BLIP-Large + LoRA) and activates all reflective mechanisms only during training (consistency and mutual-agreement losses; see Eq. (6)–(7)), the Controlled SFT baseline—sharing the identical architecture at inference—exhibits virtually identical latency and memory footprint (differences within run-to-run noise). Reported latency corresponds to caption generation under the stated decoding settings (num_beams=1, max_new_tokens=40). We report latency for caption generation, which is computationally more demanding than classification-only inference, to validate the model's efficiency under maximum load.

These results corroborate that our method introduces no additional inference overhead beyond the base BLIP model. The self-reflective losses are injected exclusively at training time, and inference uses the same BLIP-Large + LoRA network without the auxiliary visual head. This makes EdgeV-SE directly suitable for Satellite IoT deployments where Jetson-class devices serve as on-site processing nodes under strict power and latency constraints.

Table 7. On-Device Performance Benchmark (NVIDIA Jetson Orin Nano 8 GB).

Method	Latency (ms/image)	Throughput (FPS)	Peak Memory (GB)	Remarks
Controlled SFT	1697.0	0.589	0.915	Strong SFT Baseline
EdgeV-SE (Ours)	1696.9	0.589	0.915	No Additional Overhead

5.4. Qualitative Analysis

To provide empirical insight into how EdgeV-SE outperforms the standard SFT baseline, we visualize representative classification scenarios in Figures 3 and 4. For each case, we report the log-likelihood margin Δ (Eq. 4) and the linguistic probability p^L (Eq. 2), which serve as indicators of the model's internal confidence. We emphasize that the uncertainty threshold used during training (Figure 2) serves only as a diagnostic gate for identifying ambiguous samples, whereas the large margins observed here emerge naturally after convergence and reflect the model's final decision confidence.

 doi:10.20944/preprints202512.2100.v1

5.4.1. Analysis of Damage Detection (Sensitivity & Calibration)

Figure 3 illustrates the models' responses to disaster scenes. For a clear-cut flood case (Figure 3a), both models correctly identify damage. However, EdgeV-SE yields a substantially larger positive margin (Δ = +8.50) with a high linguistic probability ($p(\text{damage}|v)$ = 0.972), reflecting stronger internal agreement on salient evidence such as extensive standing water inundating road segments and surrounding building footprints. In contrast, in a challenging scenario where muddy floodwater visually resembles exposed soil (Figure 3b), the SFT baseline fails (False Negative) by interpreting the scene as accessible open terrain. EdgeV-SE correctly recognizes the subtle boundaries of turbid floodwater and its interaction with road edges and nearby structures. Notably, EdgeV-SE outputs a moderated margin for this hard sample (Δ = +4.20; $p(\text{damage}|v)$ = 0.887) relative to the easy case, indicating calibrated confidence under visual ambiguity rather than blind overconfidence.

5.4.2. Analysis of False Positive Reduction (Specificity & Hallucination)

Figure 4 examines the models' robustness against visually confusing artifacts in non-damaged areas. For an easy no-damage case (Figure 4a), both models correctly predict normalcy, while EdgeV-SE exhibits strong negative evidence ($\Delta = -7.80$; $p(\text{damage}|v) = 0.018$), consistent with high confidence in the no-damage decision. In the hard no-damage scenario (Figure 4b), the SFT baseline produces a critical hallucination-driven false positive by mistaking a rectangular swimming pool for floodwater. EdgeV-SE suppresses this semantic drift by explicitly distinguishing the pool from flooding and maintains a confident negative margin (Δ = −5.10; $p(\text{damage}|v)$ = 0.065). The reduced magnitude compared to the easy case ($|\Delta|$: 7.80 → 5.10) also reflects appropriate uncertainty in the presence of water-like visual patterns while preserving correct classification.

5.4.3. Summary of Qualitative Findings

These qualitative results corroborate the quantitative improvements reported in Section 5.1. While the standard SFT baseline exhibits failure modes under visual ambiguity—missing muddy flooding (false negatives) and hallucinating damage in the presence of water-like objects such as swimming pools (false positives)—EdgeV-SE achieves more reliable visual grounding. Importantly, EdgeV-SE's confidence signals are better calibrated: the margins are more extreme for easy cases (Figure 3a: Δ = +8.50; Figure 4a: Δ = −7.80) and appropriately moderated for hard cases (Figure 3b: Δ = +4.20; Figure 4b: Δ = −5.10), aligning confidence with scene difficulty. Crucially, these behaviors are internalized during training, enabling precise and context-aware decisions even under severe visual ambiguity.

 doi:10.20944/preprints202512.2100.v1

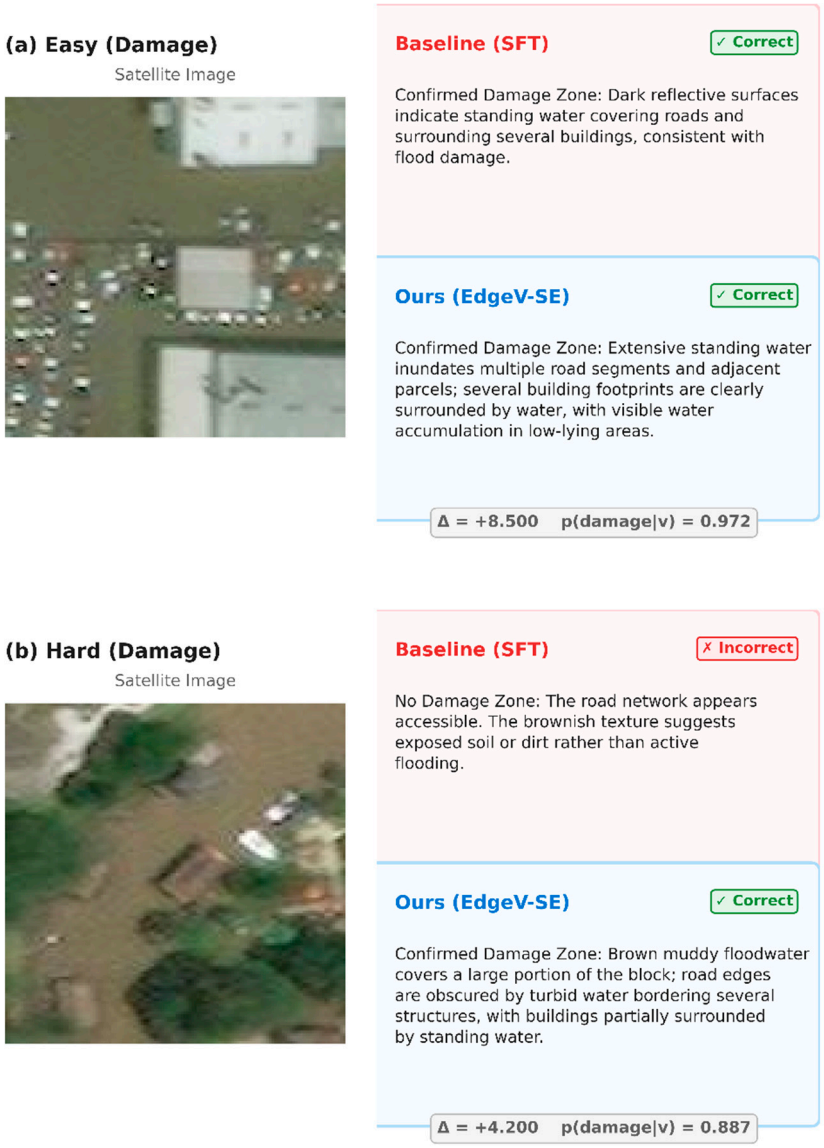

Figure 3. Qualitative comparison on Damage examples (Sensitivity Analysis). In each panel, the upper box shows the Baseline (SFT) output and the lower box shows EdgeV-SE. (a) In a clear flood scenario, both models detect damage, while EdgeV-SE provides a more precise grounded description with high confidence ($\Delta = +8.50$; $p(\text{damage} \mid v) = 0.972$). (b) In a hard sample where muddy floodwater resembles exposed soil, the SFT baseline produces a false negative, whereas EdgeV-SE correctly detects flooding with calibrated confidence ($\Delta = +4.20$; $p(\text{damage} \mid v) = 0.887$).

 doi:10.20944/preprints202512.2100.v1

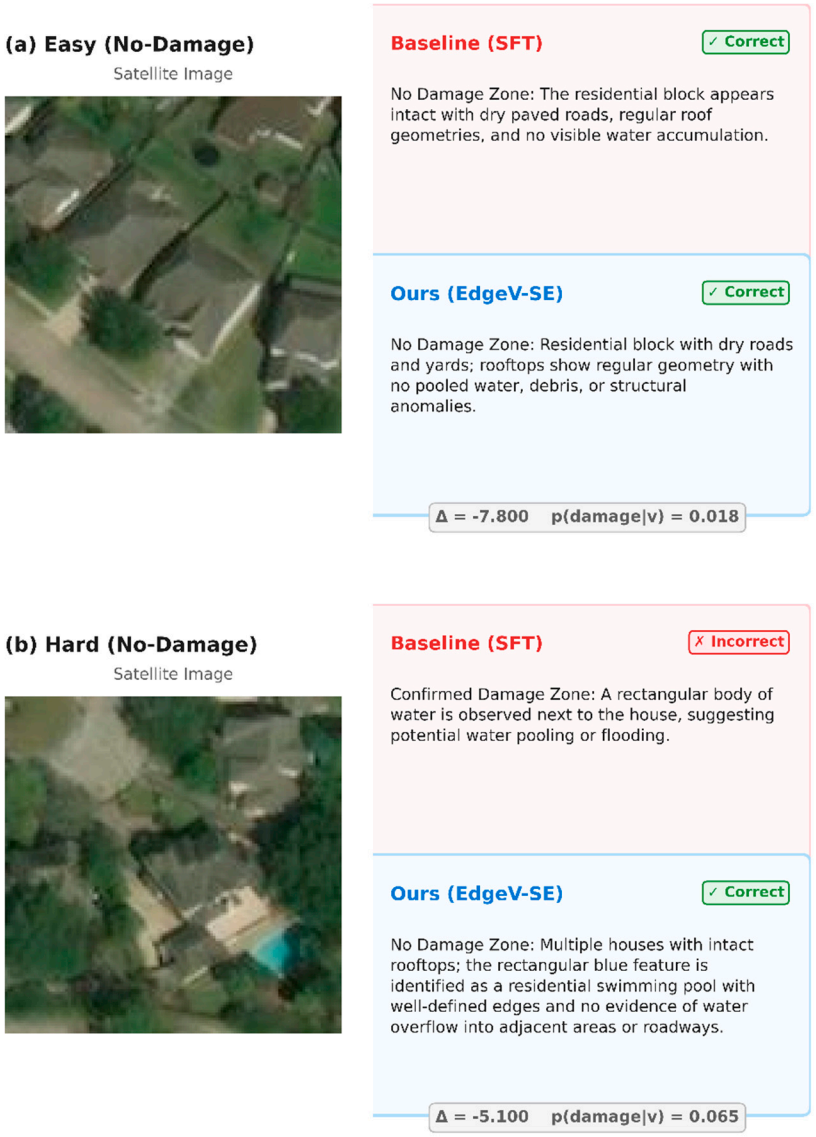

Figure 4. Qualitative comparison on No-Damage examples (Specificity Analysis). In each panel, the upper box shows the Baseline (SFT) output and the lower box shows EdgeV-SE. (a) In an easy residential scene, both models correctly predict no damage, with EdgeV-SE yielding strong negative evidence ($\Delta = -7.80$; $p(\text{damage} \mid v) = 0.018$). (b) In a hard scene containing a rectangular swimming pool, the SFT baseline produces a false positive by mistaking the pool for floodwater, while EdgeV-SE correctly distinguishes it from flooding ($\Delta = -5.10$; $p(\text{damage} \mid v) = 0.065$), demonstrating robust visual grounding and resistance to object hallucination.

5.5. Robustness under Common Corruptions and Calibration

Beyond clean test images, Satellite IoT deployments must contend with various distortions introduced by sensors, atmospheric conditions, and transmission pipelines. To examine whether EdgeV-SE's gains translate into such settings, we evaluate robustness and calibration under seven common corruptions—rotation, brightness, contrast, Gaussian blur, Gaussian noise, JPEG compression, and occlusion—each at five severities (1–5). For every condition we keep the classifier unchanged, using the LL-margin decision rule with the validation-selected threshold stored as the optimal threshold parameter, and compute Accuracy, class-wise F1, Macro-F1, and Expected Calibration Error (ECE; 15 bins) from the same temperature-scaled probability in Eq. (2). We present a reliability diagram on clean data in Figure 5 and plot severity–performance curves for Macro-F1 (Figure 6) and ECE (Figure 7).

 doi:10.20944/preprints202512.2100.v1

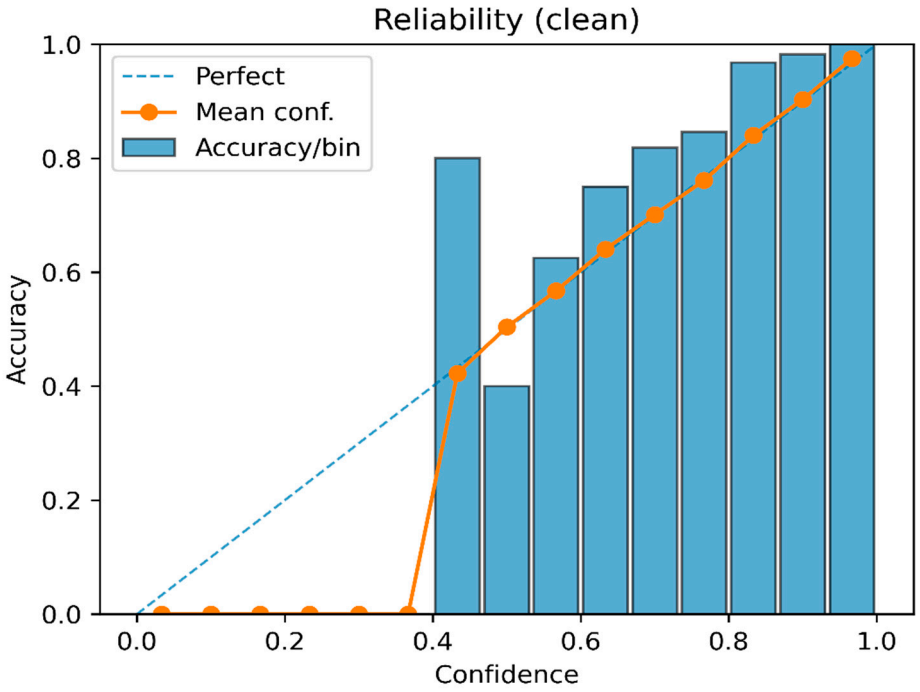

Figure 5. Reliability diagram.

On clean inputs, the model achieves high performance (Acc = 0.990, Macro-F1 = 0.985) and is well-calibrated (ECE = 0.036; Figure 5). Aggregated over all 35 corruption–severity settings by pooling all corrupted test images, performance remains high (Acc = 0.875, Macro-F1 = 0.875) with low ECE (0.057), indicating that confidence generally tracks accuracy. The severity curves show graceful degradation for brightness, blur, and rotation; JPEG exhibits a moderate drop only at the highest severity; in contrast, severe contrast corruption (s = 5) and Gaussian noise (s ≥ 3) are the principal failure modes where accuracy approaches chance and ECE rises sharply (Figures 6–7). Occlusion has minimal effect across severities in our data, likely because the synthetic central patch seldom masks key evidence.

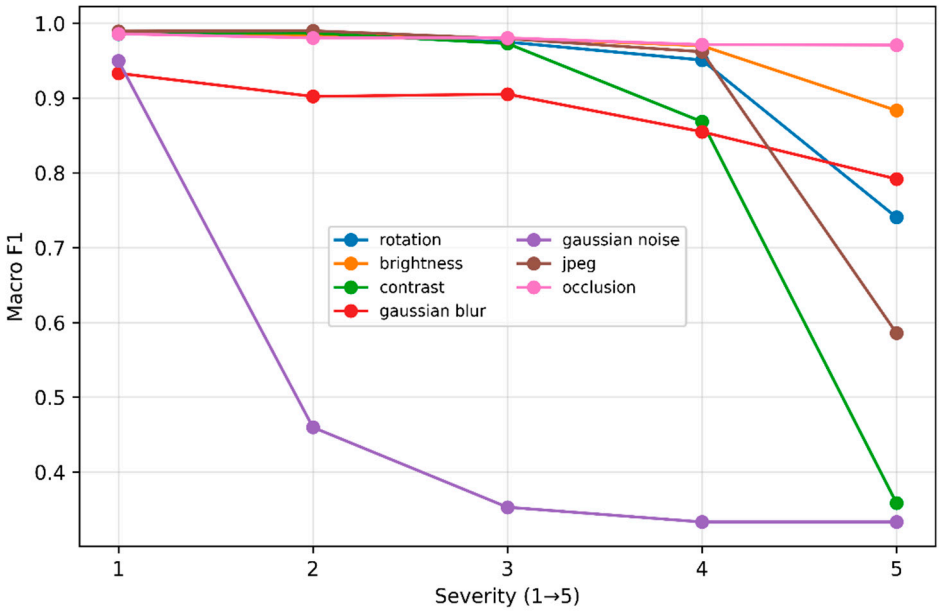

Figure 6. Severity curves (Macro-F1 vs. severity).

 doi:10.20944/preprints202512.2100.v1

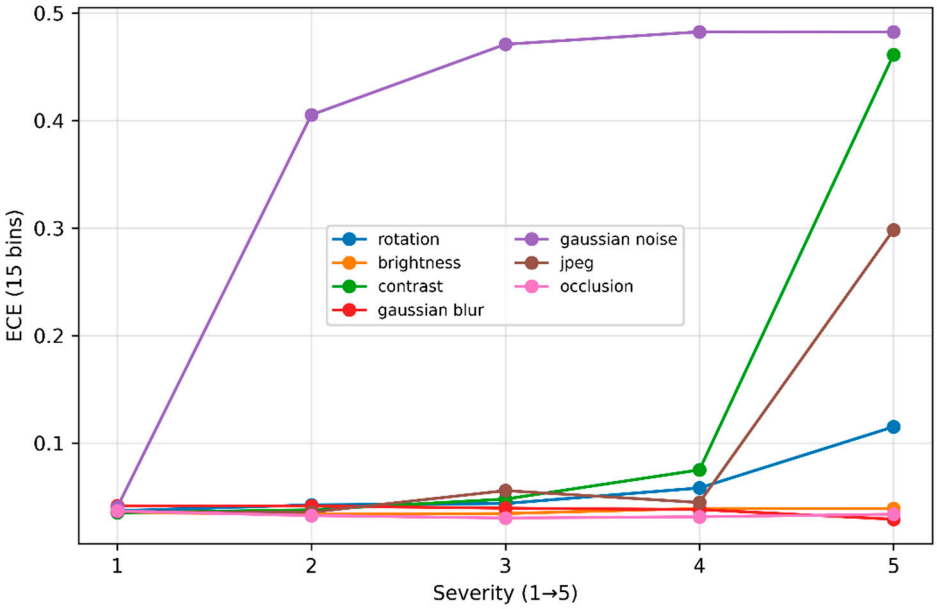

Figure 7. Severity curves (ECE vs. severity).

We attribute the observed robustness and calibration to two training-time signals that incur no inference-time overhead: (i) a Δ-consistency term that stabilizes the LL-margin decision under view changes, and (ii) mutual alignment/learning between the language and visual pathways, which reduces both over- and under-confidence on atypical inputs.

5.6. Robustness to Prompt Verbalizers

To examine whether the classification gains of EdgeV-SE are driven by memorizing specific label tokens, we evaluate the trained model using alternative class verbalizers at test time. While training uses the original class-conditioned prompts, we replace the default verbalizers ("Confirmed Damage Zone"/"No Damage Zone") with multiple semantically equivalent phrasings that do not share lexical overlap with the training prefixes.

As shown in Table 8, EdgeV-SE maintains consistent classification performance across different verbalizer choices, with only marginal variation in Accuracy and Macro-F1. This indicates that the proposed framework does not rely on specific label tokens, but instead learns robust visual-linguistic associations for damage discrimination.

Table 8. Classification performance of EdgeV-SE under different class verbalizers. The default setting uses the original class headers employed during training, while Sets A–C use semantically equivalent alternative phrasings without lexical overlap with the training prefixes.

Verbalizer Set	Accuracy	Macro-F1
Default	0.990 ± 0.004	0.985 ± 0.002
Alternative Set A	0.988 ± 0.005	0.983 ± 0.003
Alternative Set B	0.991 ± 0.004	0.984 ± 0.002
Alternative Set C	0.987 ± 0.005	0.982 ± 0.004

Specifically, alternative Sets A-C replace the default header with semantically equivalent but lexically distinct phrasings (e.g., "Severe Structural Damage Detected"/"No Visible Damage Detected". "Damage Evident in Area"/"Area Appears Undamaged"; etc.).

 doi:10.20944/preprints202512.2100.v1

6. Discussion

6.1. Synergistic Mechanisms of Self-Reflection

The significant performance gains of EdgeV-SE, improving Macro-F1 from 0.911 to 0.985 , stem from the mechanism's ability to selectively weigh gradient updates based on internal uncertainty. Unlike standard SFT which treats all samples equally, our margin-based weighting mechanism (Eq. 5) effectively acts as an internal curriculum, forcing the model to allocate more gradient capacity to ambiguous samples (e.g., subtle roof damage or occluded debris) that usually fall into the long tail of the error distribution. Furthermore, the mutual learning objective (Eq. 7) serves as a regularization term that prevents the "semantic drift" often observed in VLMs, where the model generates plausible captions that are visually unsupported. By forcing the linguistic theorist to agree with the visual empiricist, EdgeV-SE ensures that generated captions remain grounded in pixel-level evidence.

Unlike conventional consistency regularization that enforces invariance between multiple predictions, EdgeV-SE explicitly treats internal disagreement between generative (linguistic) and discriminative (visual) pathways as a diagnostic signal, and resolves it through uncertainty-aware weighting rather than uniform agreement enforcement.

6.2. Operational Feasibility vs. Hard Real-Time Constraints

A critical consideration for edge deployment is the definition of "real-time." Our benchmark on the Jetson Orin Nano shows an inference speed of approximately 0.59 FPS (1.70 s/image). While this does not meet the standard for "video-rate" real-time (e.g., >30 FPS) required for autonomous driving, it satisfies the "operational real-time" constraints of Satellite IoT disaster response. In typical satellite-to-ground scenarios, data transmission bandwidth is the primary bottleneck, often taking minutes per high-resolution image block. Consequently, an inference latency of 1.7 s is negligible compared to transmission latency. Therefore, EdgeV-SE provides a viable solution for on-site filtering, where the device processes images locally and transmits only prioritized "Damage" alerts with text descriptions, drastically reducing the bandwidth requirement compared to raw image transmission.

6.3. Justification for VLM over Lightweight Classifiers

One might question whether a lightweight image-only classifier (e.g., ResNet/MobileNet) could offer lower latency for binary damage detection. While such models can be attractive for throughput, they do not provide natural-language explanations. In disaster assessment workflows, practitioners often require interpretable evidence (e.g., flooding vs. debris vs. roof damage) rather than an opaque binary label. EdgeV-SE targets this decision-support setting by producing both a damage decision and an evidence-oriented caption. A direct speed/accuracy comparison to lightweight classifiers is left to future work.

6.4. Robustness and Reliability under Domain Shift

The bootstrap analysis and corruption tests indicate that EdgeV-SE is well-calibrated on clean inputs and degrades gracefully under several common perturbations (e.g., mild rotation/brightness/blur). However, the controlled corruption suite also reveals clear failure modes under extreme contrast collapse and stronger Gaussian noise. Therefore, robustness to real deployment conditions may require sensor- and pipeline-specific augmentation and calibration strategies beyond modest perturbations used in our self-reflective training.

6.5. Limitations and Future Directions

Despite the promising results, this study has limitations that warrant discussion.

- **Pseudo-Ground Truth.** We relied on LLaVA-1.5-generated captions due to the lack of dense human annotations. While we filtered noisy samples, some residual hallucinations from the

 doi:10.20944/preprints202512.2100.v1

teacher model may remain. Future work should incorporate a human-in-the-loop verification stage for the validation set to further guarantee semantic precision.

- **Domain Specificity.** Our experiments focused on hurricane damage. Extending this framework to other disaster types (e.g., wildfires, earthquakes) or diverse geographic landscapes is a necessary step to validate generalizability.
- **Comparison scope.** This work focuses on algorithmic fine-tuning improvements for a fixed backbone (BLIP-Large). We did not perform an exhaustive comparison against emerging edge-specific VLM architectures (e.g., MobileVLM, TinyLLaVA). Evaluating EdgeV-SE as a plug-in tuning strategy across lighter backbones is an important direction for future work. Since EdgeV-SE is an architecture-agnostic training framework, it can theoretically be applied to any encoder-decoder or decoder-only VLM (e.g., MobileVLM, TinyLLaVA) to enhance their edge reliability.

Although our experiments focus on hurricane damage assessment, the proposed framework is domain agnostic, as it relies only on internal log-likelihood margins and cross-pathway agreement. We therefore expect EdgeV-SE to generalize to other satellite-based disaster scenarios (e.g., floods or wildfires) and to different encoder-decoder VLM backbones, which we leave for future work.

7. Conclusions

We proposed a self-reflective fine-tuning framework, EdgeV-SE, for edge-deployable VLMs. Conceptually, EdgeV-SE turns the model's own uncertainty and internal disagreement into learning signals, without extra labels, external oracles, or runtime overhead. EdgeV-SE integrates uncertainty-aware weighting based on linguistic margins, margin-level multi-view semantic consistency, and dual-pathway mutual learning between linguistic and visual routes to deliver robust and reliable performance for edge-deployable vision–language models.

Our proposed model yields a substantial improvement in classification performance, increasing Macro-F1 from 0.911 to 0.985 over standard supervised fine-tuning without introducing any additional inference-time overhead. Moreover, EdgeV-SE consistently improves caption quality across key metrics, including CIDEr-D, BERTScore, and CLIPScore, resulting in concise, factual, and context-aware descriptions. Crucially, these gains are achieved while preserving edge feasibility, maintaining low latency and high throughput on resource-constrained platforms such as the Jetson Orin Nano.

In future work, we will focus on strengthening both generalization and real-world deployability of EdgeV-SE. First, we plan to construct and release a human-verified, multi-hazard benchmark (e.g., hurricanes, floods, wildfires, earthquakes) with structured damage attributes to replace or complement pseudo-caption supervision, enabling more rigorous evaluation of calibration and robustness under distribution shift. Second, we will extend EdgeV-SE to a continual/active learning setting, where low-margin (high-uncertainty) samples encountered on the edge can be flagged for optional human review and then incorporated for incremental adaptation to new geographies, sensors, and acquisition conditions without full retraining. Third, we will pursue deployment-aware optimization by combining EdgeV-SE with compression techniques (e.g., quantization, structured pruning, and distillation) and by validating the framework across lighter VLM backbones and multimodal inputs, aiming to further reduce latency and memory usage while improving reliability in time-critical disaster response workflows.

Author Contributions: Conceptualization, Y.J. and W.K.; methodology, Y.J.; software, Y.J. and S.L.; validation, Y.J. and S.L.; formal analysis, Y.J.; investigation, Y.J.; resources, W.K.; data curation, Y.J. and S.L.; writing—original draft preparation, Y.J.; writing—review and editing, W.K.; visualization, Y.J.; supervision, W.K.; project administration, W.K.; funding acquisition, W.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by IITP(Institute of Information and communications Technology Planning and Evaluation)-ICAN(ICT Challenge and Advanced Network of HRD) grant funded by the Korea

 doi:10.20944/preprints202512.2100.v1

government(Ministry of Science and ICT)(IITP-2024-00436744). This work was supported by the Dongguk University Research Fund of 2015.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author(s).

Acknowledgments: The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Masson-Delmotte, V.; Zhai, P.; Pirani, A.; Connors, S.L.; Péan, C.; Berger, S.; Caud, N.; Chen, Y.; Goldfarb, L.; Gomis, M.I.; et al. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2021. https://doi.org/10.1017/9781009157896.
2. Liu, Z.; Jiang, Y.; Rong, J. Resource Allocation Strategy for Satellite Edge Computing Based on Task Dependency. Appl. Sci. 2023, 13, 10027. https://doi.org/10.3390/app131810027.
3. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 34892–34916. https://doi.org/10.48550/arXiv.2304.08485.
4. OpenAI. *GPT-4V(ision) System Card*; OpenAI: San Francisco, CA, USA, 2023. Available online: https://cdn.openai.com/papers/GPTV_System_Card.pdf (accessed on 20 December 2025).
5. NVIDIA Corporation. Jetson Orin Nano Developer Kit User Guide. Available online: https://developer.nvidia.com/embedded/learn/jetson-orin-nano-devkit-user-guide/ (accessed on 20 December 2025).
6. Shin, D.-J.; Kim, J.-J. A Deep Learning Framework Performance Evaluation to Use YOLO in Nvidia Jetson Platform. Appl. Sci. 2022, 12, 3734. https://doi.org/10.3390/app12083734.
7. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv 2022, arXiv:2201.12086. https://doi.org/10.48550/arXiv.2201.12086.
8. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 22–29 October 2017; pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.324.
9. Shinn, N.; Cassano, F.; Berman, E.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv 2023, arXiv:2303.11366. https://doi.org/10.48550/arXiv.2303.11366.
10. Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. Self-Refine: Iterative Refinement with Self-Feedback. arXiv 2023, arXiv:2303.17651. https://doi.org/10.48550/arXiv.2303.17651.
11. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep Mutual Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4320–4328. https://doi.org/10.1109/CVPR.2018.00454.
12. Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H.; Kalenichenko, D. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2704–2713. https://doi.org/10.1109/CVPR.2018.00286.
13. Seo, H.; Choi, Y.S. V-PRUNE: Semantic-Aware Patch Pruning Before Tokenization in Vision-Language Model Inference. *Appl. Sci.* 2025, *15*, 9463. https://doi.org/10.3390/app15179463.
14. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. arXiv 2015, arXiv:1503.02531. https://doi.org/10.48550/arXiv.1503.02531.

 doi:10.20944/preprints202512.2100.v1

15. Chu, X.; Qiao, L.; Lin, X.; Xu, S.; Yang, Y.; Hu, Y.; Wei, F.; Zhang, X.; Zhang, B.; Wei, X.; et al. MobileVLM: A Fast, Strong and Open Vision Language Assistant for Mobile Devices. arXiv 2023, arXiv:2312.16886. https://doi.org/10.48550/arXiv.2312.16886.
16. Liu, X.; Zhang, Y.; Wang, Y.; Wang, Y. Aligned Vector Quantization for Edge-Cloud Collaborative Vision-Language Models. arXiv 2024, arXiv:2411.05961. https://doi.org/10.48550/arXiv.2411.05961.
17. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022. https://doi.org/10.48550/arXiv.2106.09685.
18. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 10088–10115. https://doi.org/10.48550/arXiv.2305.14314.
19. Van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. *Mach. Learn.* 2020, 109, 373–440. https://doi.org/10.1007/s10994-019-05855-6.
20. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.A.; Cubuk, E.D.; Kurakin, A.; Li, C.L. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020; Volume 33, pp. 596–608. https://doi.org/10.48550/arXiv.2001.07685.
21. Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; Le, Q. Unsupervised Data Augmentation for Consistency Training. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020; Volume 33, pp. 6256–6268. https://doi.org/10.48550/arXiv.1904.12848.
22. Tarvainen, A.; Valpola, H. Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30. https://doi.org/10.48550/arXiv.1703.01780.
23. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738. https://doi.org/10.1109/CVPR42600.2020.00975.
24. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning (ICML), Virtual, 13–18 July 2020; pp. 1597–1607. https://doi.org/10.48550/arXiv.2002.05709.
25. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 24824–24837. https://doi.org/10.48550/arXiv.2201.11903.
26. Wang, X.; Wei, J.; Schuurmans, D.; Wu, Q.; Ma, T.; Le, Q. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023. https://doi.org/10.48550/arXiv.2203.11171.
27. Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; Ma, K. Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3712–3721. https://doi.org/10.1109/ICCV.2019.00381.
28. Gal, Y.; Ghahramani, Z. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.* arXiv 2016, arXiv:1506.02142. https://doi.org/10.48550/arXiv.1506.02142.
29. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles.* arXiv 2017, arXiv:1612.01474. https://doi.org/10.48550/arXiv.1612.01474.
30. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. *On Calibration of Modern Neural Networks.* arXiv 2017, arXiv:1706.04599. https://doi.org/10.48550/arXiv.1706.04599.
31. Shrivastava, A.; Gupta, A.; Girshick, R. *Training Region-based Object Detectors with Online Hard Example Mining.* arXiv 2016, arXiv:1604.03540. https://doi.org/10.48550/arXiv.1604.03540.
32. Vo, A.; Nguyen, K.-N.; Taesiri, M. R.; Dang, V. T.; Nguyen, A. T.; Kim, D. Vision Language Models are Biased. arXiv:2505.23941. https://doi.org/10.48550/arXiv.2505.23941.

33. Lin, J. Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory 1991, 37, 145–151. https://doi.org/10.1109/18.61115.
34. Mader, K. Satellite Images of Hurricane Damage. Available online: https://www.kaggle.com/datasets/kmader/satellite-images-of-hurricane-damage (accessed on 24 November 2025). https://doi.org/10.21227/sdad-1e56.
35. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based Image Description Evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087.
36. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020. https://doi.org/10.48550/arXiv.1904.09675.
37. Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; Choi, Y. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 7–11 November 2021; pp. 7514–7528. https://doi.org/10.18653/v1/2021.emnlp-main.595.
38. Yuan, W.; Pang, R.Y.; Cho, K.; Sukhbaatar, S.; Xu, J.; Weston, J. Self-Rewarding Language Models. arXiv 2024, arXiv:2401.10020. https://doi.org/10.48550/arXiv.2401.10020.
39. Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C.D.; Finn, C. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. In Proceedings of the *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 53728–53741. https://doi.org/10.48550/arXiv.2305.18290