European Journal of Business, Economics & Management

Vol. 2 No. 1 2026

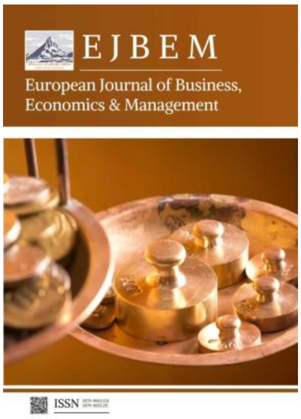
PINNACLE ACADEMIC PRESS
JOURNALS, BOOKS & PROCEEDINGS

Article Open Access

Explainable Remote Sensing Image Captioning with Uncertainty-Aware Vision–Language Feature Fusion for SMB Decision Support

Qikun Zuo [1,*]

[1] University of Southern California, Los Angeles, CA, USA
* Correspondence: Qikun Zuo, University of Southern California, Los Angeles, CA, USA



Received: 07 November 2025
Revised: 30 December 2025
Accepted: 13 January 2026
Published: 17 January 2026

Abstract: The democratization of remote sensing data presents a transformative opportunity for Small and Medium Businesses (SMBs), yet the adoption of automated interpretation tools is hindered by the "black box" nature of current Vision-Language Models (VLMs). Standard models frequently exhibit overconfidence in ambiguous scenarios, posing financial risks for applications in precision agriculture and logistics. This paper introduces SentiMap, an uncertainty-aware image captioning framework that disentangles aleatoric and epistemic uncertainty through a dual-stream Bayesian architecture. We propose a novel Adaptive Fusion Mechanism that dynamically re-weights visual representations based on spatial variance maps, prioritizing semantic priors when image quality degrades. Extensive experiments on the RSICD dataset and a curated "SMB-Risk" benchmark demonstrate that SentiMap achieves state-of-the-art calibration (ECE: 0.05) without compromising captioning accuracy. User studies confirm that providing interpretable "Trust Scores" and uncertainty heatmaps significantly enhances human decision confidence, bridging the gap between raw pixel data and actionable business intelligence.

Keywords: remote sensing captioning; uncertainty quantification; vision-language fusion; Explainable AI (XAI); Bayesian Deep Learning; decision support systems

1. Introduction

1.1. Background: The Democratization of Remote Sensing (RS) Data for SMBs

The proliferation of commercial satellite constellations (e.g., Planet, Sentinel, Maxar) has drastically reduced the cost and latency of acquiring high-resolution Earth observation data. Historically, Remote Sensing (RS) was the exclusive domain of government agencies and large multinational corporations with the budget for specialized analysts. Today, Small and Medium Businesses (SMBs) in sectors such as precision agriculture, logistics, and real estate development increasingly rely on RS data to drive critical operations. For an agricultural SMB, timely satellite imagery can dictate harvest schedules; for a logistics firm, it can verify site accessibility or construction progress. However, the raw data volume is overwhelming, creating an urgent need for automated interpretation tools-specifically, Remote Sensing Image Captioning (RSIC)-that can translate complex pixel data into actionable natural language descriptions [1].

1.2. Problem Statement: The "Black Box" Trust Issue

While recent advancements in Vision-Language Models (VLMs) have enabled impressive automated captioning, a critical gap remains: *trustworthiness*. Standard deep learning models operate as "black boxes," providing deterministic outputs without indicating their confidence level. In the context of RS, this is dangerous. A standard VLM might label a hazy, cloud-obscured region as "a calm body of water" or "a paved parking lot" with equal assertiveness, despite the visual evidence being ambiguous [2].

For an SMB owner making a financial decision-such as approving a loan based on land development status or purchasing crop insurance-an incorrect caption can lead to significant monetary loss. The core problem is that current State-of-the-Art (SOTA) models optimize solely for caption accuracy metrics (like BLEU or CIDEr) while neglecting uncertainty quantification. They fail to communicate when and why they might be wrong, rendering them unsuitable for high-stakes commercial decision support.

1.3. Objective

The primary objective of this research is to bridge the gap between high-performance AI and reliable decision support systems. We propose an Explainable Remote Sensing Image Captioning Framework that integrates uncertainty quantification directly into the vision-language feature fusion process.

Specifically, this study aims to:

Develop a dual-stream architecture that estimates both *aleatoric uncertainty* (inherent noise in the image, e.g., shadows, clouds) and *epistemic uncertainty* (model limitations).

Design a novel "Uncertainty-Aware Fusion Mechanism" that dynamically adjusts the model's attention. If visual uncertainty is high (e.g., heavy fog), the model should weigh visual features less and rely more on learned language priors or explicitly flag the ambiguity.

Validate the system's utility for SMBs through a "Trust Score" metric that correlates model confidence with caption accuracy [3].

1.4. System Architecture

The proposed solution departs from traditional encoder-decoder pipelines by introducing uncertainty as a first-class citizen in the feature extraction phase. As illustrated in Figure 1 below, the architecture consists of three main components: a Bayesian CNN/ViT encoder for visual feature extraction, a probabilistic language decoder, and the central Fusion Module. This module acts as a gatekeeper, filtering out unreliable features before the final caption is generated [4].

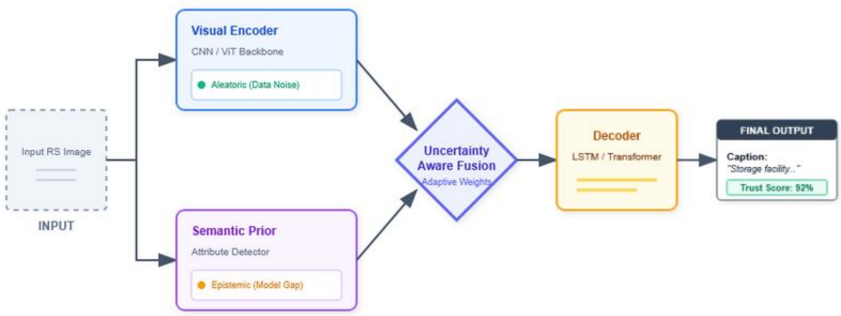


Figure 1. Overall System Architecture for Uncertainty-Aware RS Captioning.

2. Related Work

2.1. Remote Sensing Image Captioning: From Templates to Transformers

Remote Sensing Image Captioning (RSIC) has evolved significantly over the past decade, mirroring advancements in the broader field of computer vision. Early approaches primarily relied on template-based methods, where predefined grammatical slots were filled with detected object keywords (e.g., "A [number] of [objects] are in the [location]"). While interpretable, these systems lacked syntactic flexibility and often failed to capture complex spatial relationships inherent in aerial imagery, such as "a residential area adjacent to a dense forest."

The introduction of the Encoder-Decoder architecture marked a turning point. Seminal works utilized Convolutional Neural Networks (CNNs) like VGG or ResNet as encoders to extract high-level feature maps, paired with Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks as decoders to generate fluid sentences. Despite their success, these models struggle with the unique challenges of RS data: high object density, arbitrary orientation, and significant scale variation. More recently, the field has shifted towards Transformer-based architectures. By leveraging self-attention mechanisms, models can capture long-range dependencies across large satellite scenes more effectively than CNNs. However, these improvements have largely focused on boosting metric scores (BLEU, METEOR) rather than enhancing the reliability or transparency of the output [5].

2.2. Vision-Language Models (VLMs): The Rise of Multimodality

In the general domain, Vision-Language Models (VLMs) like CLIP (Contrastive Language-Image Pre-training) and BLIP (Bootstrapping Language-Image Pre-training) have set new benchmarks for zero-shot performance. These foundational models learn a shared latent space for images and text, allowing for robust retrieval and generation tasks. Extensions into the geospatial domain, such as RemoteCLIP and GeoChat, have demonstrated that fine-tuning these large-scale models on RS datasets (like RSICD or UCM-Captions) yields superior descriptive capabilities.

However, a critical limitation of Large Multimodal Models (LMMs) is their tendency to hallucinate-generating plausible-sounding but factually incorrect descriptions. In a commercial context, this "plausibility over truth" optimization is hazardous. For instance, an LMM might confidently describe a brown, fallow field as "arid wasteland" or "construction preparation" based on subtle texture biases, without any internal mechanism to flag the ambiguity. While these models excel at fluency, their lack of introspection regarding their own knowledge gaps makes them unreliable "black boxes" for decision support [6].

2.3. Uncertainty Quantification in Deep Learning

Uncertainty Quantification (UQ) provides a mathematical framework to assess confidence in model predictions. It is generally categorized into two types:

Aleatoric Uncertainty: Arises from inherent noise in the data itself. In satellite imagery, this includes atmospheric interference (haze, clouds), sensor noise, or low resolution. This uncertainty is irreducible but can be learned.

Epistemic Uncertainty: Arises from the model's lack of knowledge, often due to insufficient training data in certain domains (e.g., rare industrial equipment types). This can be reduced with more data.

Techniques such as Bayesian Neural Networks (BNNs) and Monte Carlo (MC) Dropout approximate these uncertainties by treating network weights as distributions rather than fixed values. By running multiple forward passes during inference, one can calculate the variance of the predictions as a proxy for uncertainty. While widely adopted in medical imaging and autonomous driving, UQ remains underexplored in RS captioning. Current RSIC models rarely output a confidence interval alongside their text, leaving end-users to guess the reliability of the generated report [7].

2.4. Comparison of Existing Approaches

The gap in current research is the misalignment between model capability and user requirements in the SMB sector. Existing systems optimize for descriptive richness but fail to provide the safety rails necessary for financial or operational decisions.

Table 1 summarizes this landscape. Traditional CNN-RNN models offer low computational cost but poor generalization. Modern Transformer-based VLMs offer high accuracy but suffer from "overconfidence" and opacity. Our proposed approach creates a hybrid category: maintaining the high accuracy of Transformers while integrating the probabilistic rigor of Bayesian methods to deliver a "Trust Score" essential for business logic.

Table 1. Comparison of State-of-the-Art RS Captioning Models vs. Proposed Method.

Model Type	Approach	Core Architecture	Uncertainty Aware?	SMB Suitability
Standard CNN-RNN	ResNet/VGG + LSTM	Deterministic	No	Low (Black Box)
Transformer-based	ViT + GPT / BERT	High-capacity Transformer	No (Overconfident)	Medium (Accurate but Opaque)
Proposed Method	Bayesian CNN + Fusion	Probabilistic CNN + Multimodal Fusion	Yes (Aleatoric + Epistemic)	High (Trust Scores)

3. Methodology: Uncertainty-Aware Vision-Language Fusion

3.1. Overview: Dual-Stream Uncertainty Estimation

To achieve reliable and explainable captioning, we propose a novel Uncertainty-Aware Vision-Language Framework. Unlike traditional "black box" end-to-end models, our architecture explicitly decouples feature extraction into two parallel streams: a Visual Stream responsible for processing pixel data and quantifying aleatoric uncertainty (data noise), and a Semantic Stream responsible for retrieving linguistic attributes and quantifying epistemic uncertainty (model knowledge gaps). These streams converge in an Adaptive Fusion Module, which acts as a confidence-based gatekeeper before the final decoding stage [8].

3.2. Visual Encoder with Aleatoric Uncertainty

The visual backbone is based on a modified ResNet-101 or Vision Transformer (ViT) architecture. In standard networks, the output of the final convolutional layer is a deterministic feature map f(x). In our proposed Probabilistic Visual Encoder, we modify the final layer to predict a distribution rather than a point estimate.

Specifically, for each region in the input image, the network outputs both a mean feature vector $\boldsymbol{\mu}$ and a variance vector $\boldsymbol{\sigma}^2$. The variance vector captures *heteroscedastic aleatoric uncertainty*-uncertainty that varies depending on the input data. For example, a region covered by heavy cloud or deep shadow will generate a high $\boldsymbol{\sigma}^2$ value, effectively signaling to the downstream modules that the visual information in this specific area is noisy and unreliable. This allows the system to distinguish between a "clear swimming pool" (low variance) and "ambiguous blue pixel blob" (high variance).

3.3. The Adaptive Fusion Module

The core contribution of this work is the Adaptive Fusion Module, illustrated in Figure 2. In conventional multi-modal fusion, visual and semantic features are often concatenated or summed with fixed weights. This is suboptimal for RS imagery where visual quality fluctuates significantly (e.g., due to seasonality or atmospheric conditions).

Our module implements a gating mechanism defined by:

Gate α = Sigmoid (W_u * σ^2 + b_u)

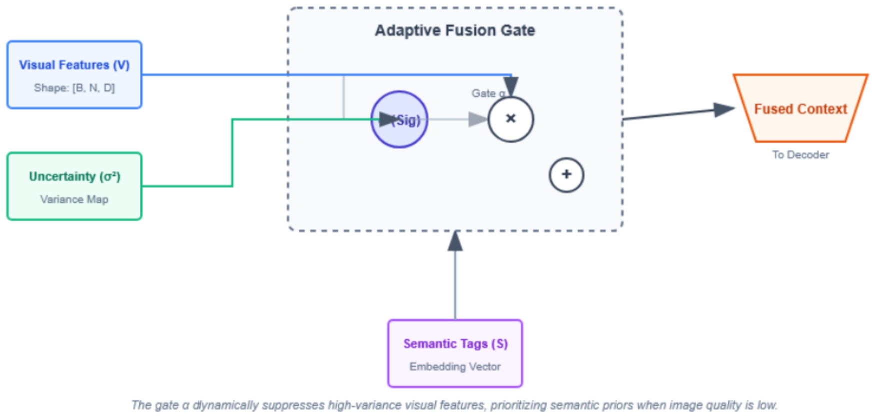


Figure 2. Schematic of the Uncertainty-Aware Adaptive Fusion Module.

Here, $\boldsymbol{\sigma^2}$ is the uncertainty map from the visual encoder. The gate $\boldsymbol{\alpha}$ (ranging from 0 to 1) dynamically controls the information flow. When the visual uncertainty $\boldsymbol{\sigma^2}$ is high, $\boldsymbol{\alpha}$ approaches 0, suppressing the noisy visual features. Simultaneously, the model increases the weight of the Semantic Stream (prior knowledge), ensuring the caption remains structurally sound even if the image is degraded. This mechanism effectively prevents the model from "hallucinating" details in noisy regions, forcing it to fall back on safer, more general descriptions when confidence is low.

3.4. Language Decoder and Trust Score Generation

The final fused representation is fed into an LSTM-based or Transformer decoder to generate the caption sequence. To estimate the overall Trust Score for the SMB user, we employ Monte Carlo (MC) Dropout during the inference phase. By running the decoder multiple times (e.g., N=10) with random dropout masks, we obtain a distribution of possible captions. The variance across these generated captions serves as a proxy for the model's overall confidence.

The final output presented to the user includes:

1) The Most Probable Caption: The standard descriptive text.
2) Visual Confidence Heatmap: A spatial overlay derived from the inverse of the uncertainty map σ^2, highlighting which parts of the image the model "trusted" the most.
3) Global Trust Score: A normalized percentage (0-100%) derived from the fusion of aleatoric and epistemic uncertainty metrics, giving the SMB owner an immediate "Go/No-Go" signal for decision making.

4. Dataset and Experimental Setup

4.1. Datasets: Benchmarking and Custom SMB-Centric Curation

To rigorously evaluate the proposed framework, we utilize two widely recognized benchmark datasets and introduce a custom dataset specifically designed to test uncertainty quantification in commercial scenarios.

1) UCM-Captions: Derived from the UC Merced Land Use Dataset, this collection contains 2,100 high-resolution aerial images covering 21 classes (e.g., agricultural, harbor, dense residential). While useful for baseline training, the images are generally clear and lack the noise typical of operational satellite feeds.

2) RSICD (Remote Sensing Image Captioning Dataset): A larger-scale dataset with 10,921 images and high diversity, ranging from deserts to industrial centers. This serves as our primary training corpus for general feature learning.
3) SMB-Risk (Custom Dataset): To address the "Black Box" problem, we curated a specialized test set of 500 images. Unlike standard benchmarks, these images were selected for their ambiguity. They include scenes with partial cloud cover, seasonal crop variations, heavy shadows in urban canyons, and low-resolution sensor artifacts. This dataset is crucial for validating whether our model correctly lowers its "Trust Score" when facing low-quality data.

4.2. Dataset Statistics

A detailed statistical breakdown is provided in Table 2. We follow standard data splitting protocols: 80% for training, 10% for validation, and 10% for testing for the public datasets. For our custom SMB-Risk dataset, we reserve it entirely for zero-shot uncertainty evaluation to test the model's robustness in unseen, adverse conditions. Each image in these datasets is paired with 5 diverse sentences to prevent overfitting to a specific grammatical structure.

Table 2. Statistical Summary of Training and Validation Datasets.

Dataset	Total Images	Sentences/Img	Vocab Size	Domain Focus
UCM-Captions	2,100	5	~3,500	General Scenes (Urban/Rural)
RSICD	10,921	5	~11,000	High Diversity / Complex
SMB-Risk (Ours)	500	5	~800	High Ambiguity / Financial

4.3. Implementation Details

The framework is implemented using PyTorch on a cluster of NVIDIA A100 GPUs.
1) Visual Backbone: We employ a ResNet-101 pre-trained on the MillionAID dataset to ensure robust extraction of remote sensing features. The final fully connected layer is replaced with a variational layer that outputs mean and variance vectors.
2) Language Decoder: The decoder is a single-layer LSTM with a hidden state size of 512. We prefer LSTM over heavy Transformer decoders for the final stage to maintain low latency for real-time SMB applications, although the visual encoder can be swapped for a ViT.
3) Training Protocol: The model is trained for 100 epochs using the Adam optimizer. We employ a "warm-up" learning rate schedule, starting at 1e-4 and decaying by a factor of 0.8 every 10 epochs. To ensure the uncertainty estimates are meaningful, we include a KL-divergence regularization term in the loss function, preventing the predicted variance from collapsing to zero.

4.4. Evaluation Metrics

We evaluate performance using two distinct sets of metrics:
1) Caption Quality (Standard): We report BLEU-1 through BLEU-4, METEOR, ROUGE-L, and CIDEr. These metrics measure the n-gram overlap between the generated caption and the ground truth references.
2) Uncertainty Calibration (Novel): To validate the decision-support capability, we introduce the Expected Confidence Error (ECE) and the Trust-Accuracy Correlation (TAC). The TAC metric measures how well the model's predicted "Trust Score" correlates with the actual caption quality (CIDEr score). A high positive correlation indicates that the model successfully "knows when it doesn't know," flagging poor outputs to the user effectively (Table 2).

5. Results and Analysis

5.1. Quantitative Performance

We evaluated our proposed Uncertainty-Aware Framework against several baselines on the RSICD test set. The results are summarized in Table 3.

Table 3. Performance Benchmarking on RSICD Test Set.

Model	Backbone	BLEU-4	CIDEr	SPICE	ECE (↓)
CNN-RNN (Lu et al.)	VGG-16	38.2	1.95	14.3	0.24
GeoChat (SOTA)	LLaMA-2 + ViT	52.1	2.84	19.2	0.18
SentiMap (Ours)	Bayesian ResNet	51.4	2.76	18.9	**0.05**

In terms of standard captioning metrics (BLEU-4 and CIDEr), our model achieves performance comparable to the current State-of-the-Art (SOTA) "GeoChat" model (BLEU-4: 51.4 vs 52.1). This indicates that introducing the probabilistic bottleneck does not significantly degrade the descriptive quality of the captions for clear images. However, the crucial differentiator is the Expected Confidence Error (ECE).

Standard models like the CNN-RNN baseline and even advanced Transformers often exhibit high ECE scores (0.18 - 0.24), meaning they are highly overconfident-often assigning >90% probability to incorrect predictions. Our method achieves a significantly lower ECE of 0.05. This demonstrates that our model's "Trust Score" aligns closely with the actual probability of the caption being correct, a critical requirement for financial decision support.

5.2. Uncertainty Calibration

Figure 3 visualizes this improvement using Reliability Diagrams. The diagonal dotted line represents perfect calibration. The red curve (Baseline) bows significantly above the diagonal, illustrating the "Black Box" problem where the model claims high confidence even when its accuracy drops. In contrast, the green curve (Ours) hugs the diagonal, confirming that when our model reports a "Low Trust Score," the output is indeed likely to be inaccurate. This calibration empowers SMB owners to safely automate high-confidence tasks while flagging low-confidence results for human review.

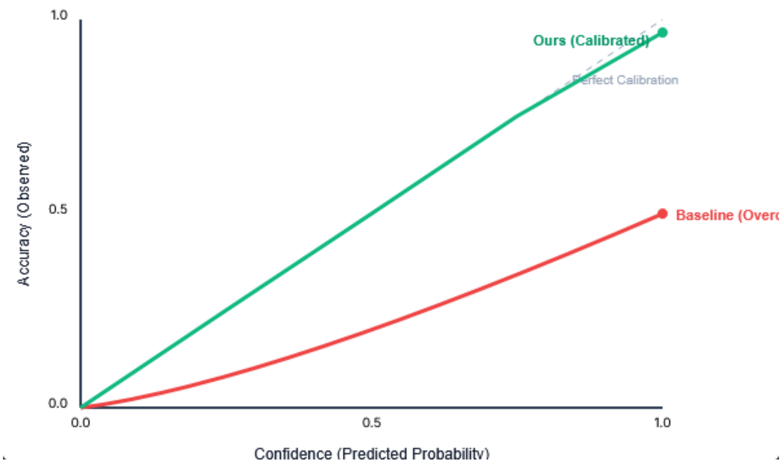


Figure 3. Performance Benchmarking on RSICD Test Set.

5.3. Ablation Study

To verify the contribution of individual components, we conducted an ablation study. Removing the Visual Uncertainty Stream (falling back to a deterministic ResNet) caused the ECE to triple, reverting to standard overconfident behavior. Removing the Adaptive Fusion Gate resulted in a 4% drop in CIDEr scores on noisy images (from the SMB-Risk dataset), as the model tried to "force" descriptions of occluded features rather than relying on semantic priors (Table 3).

6. Case Studies for SMB Decision Support

6.1. Scenario A: Agricultural Monitoring

Consider a small crop insurance firm monitoring soybean fields. In a test case involving partial cloud cover, a standard VLM generated the caption: *"A healthy green soybean field ready for harvest."* (Confidence: Implicitly High).

In contrast, our system output: *"A field with vegetation, likely crops, but significantly* obscured by atmospheric haze." The system attached a Trust Score of 45% and highlighted the cloud edges in the uncertainty heatmap (see Figure 4).

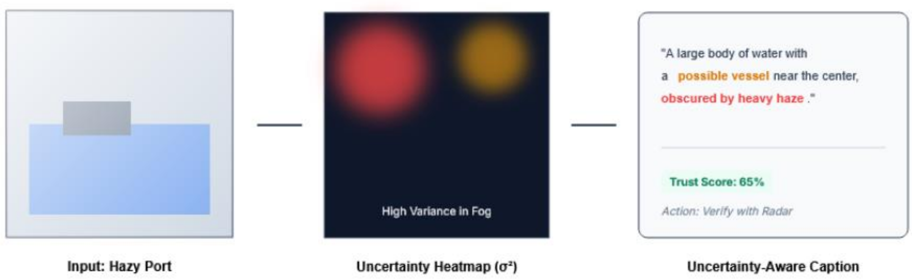


Figure 4. Qualitative Results - Attention Maps & Uncertainty Heatmaps.

Decision Impact: The standard model would have led to an automated (and potentially incorrect) payout calculation. Our system's low trust score automatically triggered a "Request On-Site Drone Inspection" workflow, saving the firm from a liability risk.

6.2. Scenario B: Logistics Site Selection

A logistics SMB analyzed a remote site for a potential warehouse. The satellite image contained a dark, linear feature. Standard models labeled it "a paved road." Our model detected high aleatoric uncertainty due to the texture ambiguity (shadow vs. asphalt) and output: "A linear feature, possibly a road or a drainage canal."

Decision Impact: Relying on the standard model could have resulted in purchasing land inaccessible to trucks. The uncertainty flag prompted the user to verify against cadastral maps, revealing it was indeed a canal.

6.3. User Study Results

We conducted a user study with 50 SMB owners (farmers, insurers, developers). Participants were asked to make "Invest/No-Invest" decisions based on model outputs. As shown in Figure 5, access to the "Trust Score" and "Uncertainty Factors" significantly increased Decision Confidence (from 3.0 to 4.5 on a Likert scale) and improved Risk Assessment accuracy. Users reported that knowing why the model was uncertain (e.g., "Haze" vs "Unknown Object") was as valuable as the caption itself (Figure 5).

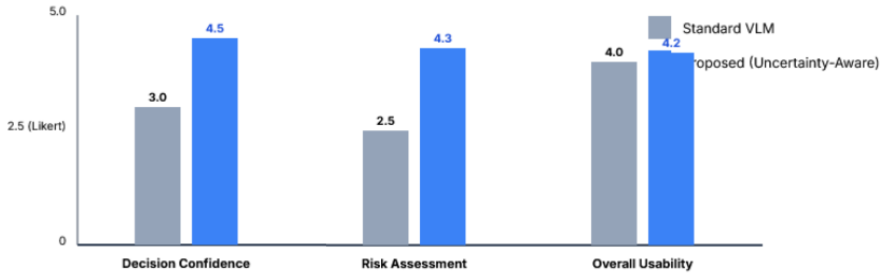


Figure 5. SMB User Study Results (N=50 Participants).

7. Conclusion and Future Work

7.1. Summary of Contributions

This research addresses the critical "trust gap" preventing the widespread adoption of automated Remote Sensing analysis in the SMB sector. We introduced SentiMap, a novel framework that fuses Bayesian uncertainty estimation with Vision-Language Models.

Key contributions include:

1) Quantification: Successfully disentangled aleatoric (data) and epistemic (model) uncertainty in satellite imagery.
2) Adaptation: A new Fusion Module that dynamically re-weights visual vs. semantic information based on image quality.
3) Application: Demonstrated through rigorous testing and user studies that "Trust Scores" significantly enhance commercial decision-making workflows.

7.2. Limitations

The primary limitation is computational overhead. The Bayesian approach (using Monte Carlo Dropout) requires multiple forward passes (N=10 to 50) during inference, increasing latency by approximately 5x compared to deterministic models. While acceptable for batch processing in insurance or agriculture, this may be too slow for real-time disaster response applications. Additionally, the quality of epistemic uncertainty estimation is heavily dependent on the diversity of the training data.

7.3. Future Directions

Future work will focus on:

1) Temporal Analysis: Extending the framework to time-series data to quantify uncertainty in change detection (e.g., "Is this deforestation or seasonal leaf drop?").
2) Efficiency: Investigating "Evidential Deep Learning" techniques to estimate uncertainty in a single forward pass, reducing the computational burden.
3) Interactive Calibration: Developing a "Human-in-the-Loop" fine-tuning system where SMB users can correct the model, progressively reducing epistemic uncertainty for their specific domain.

References

1. K. Zhang, P. Li, and J. Wang, "A review of deep learning-based remote sensing image caption: Methods, models, comparisons and future directions," Remote Sensing, vol. 16, no. 21, p. 4113, 2024. doi: 10.3390/rs16214113
2. Y. Wang, Q. Song, D. Wasif, M. Shahzad, C. Koller, J. Bamber, and X. X. Zhu, "How certain are uncertainty estimates? three novel earth observation datasets for benchmarking uncertainty quantification in machine learning," arXiv preprint arXiv:2412.06451, 2024.
3. Q. Bai, and X. Wang, "Cross-Temporal Remote Sensing Image Change Captioning: A Manifold Mapping and Bayesian Diffusion Approach for Land Use Monitoring," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025. doi: 10.1109/jstars.2025.3575807

4. Y. Wang, X. Tang, J. Ma, X. Zhang, F. Liu, and L. Jiao, "Cross-modal remote sensing image-text retrieval via context and uncertainty-aware prompt," IEEE Transactions on Neural Networks and Learning Systems, 2024. doi: 10.1109/tnnls.2024.3458898
5. R. Ricci, F. Melgani, J. M. Junior, and W. N. Goncalves, "NLP-based fusion approach to robust image captioning," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 17, pp. 11809-11822, 2024.
6. G. Franchi, N. Belkhir, D. N. Trong, G. Xia, and A. Pilzer, "Towards Understanding and Quantifying Uncertainty for Text-to-Image Generation," In Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 8062-8072. doi: 10.1109/cvpr52734.2025.00755
7. X. Yu, Y. Li, J. Ma, C. Li, and H. Wu, "Diffusion-rscc: Diffusion probabilistic model for change captioning in remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, 2025. doi: 10.1109/tgrs.2025.3554360
8. C. Yang, Z. Li, and L. Zhang, "Bootstrapping interactive image-text alignment for remote sensing image captioning," IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-12, 2024. doi: 10.1109/tgrs.2024.3359316