Look, Reason, Defuse: Bridging Perception and Domain Knowledge for Real-World Unexploded Ordnance Identification

Gheorghe Marian Craioveanu
University Politehnica of Bucharest, Romania
gheorghe.craioveanu@upb.ro

Grigore Stamatescu
University Politehnica of Bucharest, Romania
grigore.stamatescu@upb.ro

Olga Saukh
Graz University of Technology, Austria
saukh@tugraz.at

Abstract

We introduce DEFUSAL*, a neuro-symbolic framework that narrows the semantic gap between the perceptual flexibility of Vision-Language Models (VLMs) and the strict safety requirements of Explosive Ordnance Disposal (EOD) procedures, by enabling the identification of unexploded ordnance types in real-world conditions characterised by long-tailed distributions. We bring together core theoretical foundations and extensive domain expertise through a tightly integrated set of novel components, including a custom, validated EOD Knowledge Graph, a purpose-built safety mechanism, all embedded within a specialised, real-world operational framework. We evaluate the proposed architecture across 13 configurations, including a four-stage ablation study and a direct comparison with a neuro-symbolic baseline for logical reasoning. All experiments are performed on a real-world dataset comprising 13,648 individual bombs, which is also analysed considering the challenges involved, offering a practical resource for UXO identification. We show that the* DEFUSAL *framework surpasses the traditional zero-shot approach, improving the F1-Score by 32.7% when the system makes the decision, with the safety mechanism enabled. This work acts as a roadmap for trustworthy, real-world deployment of humanitarian UXO clearance tools that go beyond simple label prediction, instead reasoning about the underlying physical reality to offer reliable, interpretable, and safety-aware decision support in post-conflict environments, even when dealing with rare UXO types.*

1. Introduction

According to the North Atlantic Treaty Organization (NATO) Standardization Agreement NATO STANAG [34],

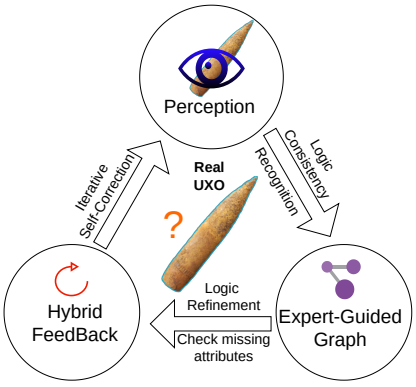


Figure 1. Logic-guided architecture of the proposed DEFUSAL framework for unexploded ordnance (UXO) identification.

unexploded ordnance (UXO) is defined as explosive ordnance that has been primed, fused, armed, or otherwise prepared for action, and that has been fired, dropped, launched, projected, or placed but remains unexploded due to malfunction, design, or any other cause. The removal of Unexploded Ordnance (UXO) in regions emerging from armed conflicts constitutes both a major humanitarian imperative and a complex engineering problem [20, 51]. Although a precise worldwide estimate is unavailable, existing regional assessments [20] indicate that the total number of UXO items remaining in the environment reaches hundreds of millions. As an illustrative case, UXOs Lao reports [24, 52] that of the approximately 270 million submunitions deployed over Laos, a significant proportion did not explode as intended, with failure rates commonly estimated in the range of 10-30 %, resulting in up to 81 million undiscovered UXOs. These remnants pose enduring threats to public safety, impede socio-economic progress, and impose severe constraints on land and infrastructure management. Although supervised approaches can perform well on fixed,

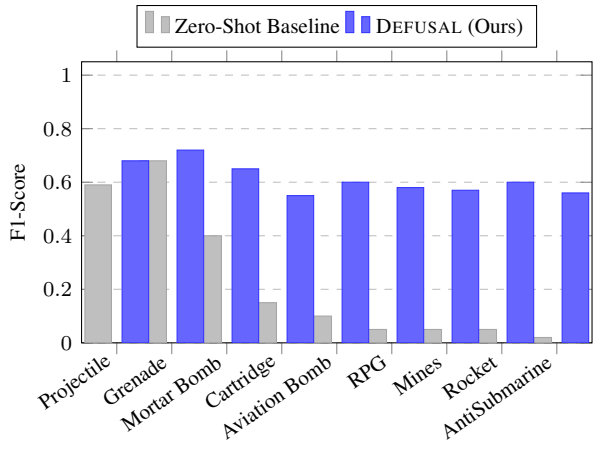


Figure 2. **Summary of results (per-class F1-score).** DEFUSAL is the proposed framework, designed as a neurosymbolic system guided by logical reasoning. In the zero-shot setting, the model shows a bias toward the more frequent class in the *training corpus*. The real-world UXO task involves a multi-level, long-tailed distribution (also see Appendix E) with challenging instances (also see Figure 4 and Appendix H), while requiring outputs that can be logically verified (see Section 3.1) and validated (see Appendix C). The framework further incorporates safety mechanisms (see Section 3.4 and Figure 3) reflecting the safety-critical nature of the domain—key requirements for trustworthy AI deployment in humanitarian applications.

closed datasets, they do not readily accommodate the open-set characteristics of UXO identification without ongoing retraining. In practice, new and rare types of ammunition often emerge in the field, for which training data are initially not available. Our approach is generalised to foundation-model settings that require adaptive and interpretable reasoning under open-set conditions. To address the challenges, a neuro-symbolic adaptive framework that seeks to infer contextual meaning from low-level visual attributes, exemplified in Figure 1, can improve robustness and reduce False Negative cases. As a first common challenge in the current state of UXO identification, ordnance datasets exhibit long-tail behaviour. Thousands of rare or historical variants of UXOs exist for which labelled visual data are scarce or non-existent. Deep learning models struggle to generalise to these rare classes, often overfitting to background textures [41] rather than learning object visual characteristics or semantics. In the zero-shot paradigm, foundation models are prone to hallucination, inevitably leading to elevated risks to human life. Although Chain-of-Thought [57] and reasoning capabilities in foundation models have made great strides, generating hypotheses and intermediate steps does not ensure factual precision [5, 64], especially due to the under-representation of certain classes, making it difficult to generalise to previously unseen or semantically undescribed classes.

Our neuro-symbolic method closes this adaptability gap by enabling the recognition of rare categories through logical specifications, even before adequate or balanced training data for Unexploded Ordnance (UXO) is present. Based on the limits and risks posed by UXOs, to address the inherent unreliability of black-box foundation models in safety-critical environments, we introduce a unified study for Explosive Ordnance Disposal (EOD) operations, a system inspired by real-world operational needs, designed to enforce physical and logical consistency in Vision-Language Models (VLMs). Our key contributions are as follows:

- **Neuro-Symbolic Framework for Real-World UXO Identification:** We introduce a closed-loop multimodal framework, with a safety mechanism, that couples a Vision-Language Model with a symbolic reasoning engine based on Łukasiewicz fuzzy logic and Probabilistic Soft Logic. Operating in a zero-shot regime, the system autonomously detects and corrects perceptual errors (e.g., hallucinations) through iterative logic-guided self-correction, reducing false negatives in a safety-critical, real-world setting.
- **EOD Knowledge Graph for Long-Tail Generalisation:** We introduce a domain-specific knowledge graph constructed and validated by an EOD specialist against technical manuals [17, 18, 43], allowing zero-shot generalisation to rare and unseen ordnance types without retraining, directly addressing the scarcity and heavy class imbalance inherent to real-world UXO data. Experiments are conducted on 13,648 real-world images from active demining interventions.
- **Benchmark, Ablation, and Reasoning Analysis:** We conduct 16 types of experiments, with 10 runs per setup, demonstrating consistent improvements over baselines with fully auditable decision traces. Figure 2 presents a summary of the key results per-class of the F1-Score.

This work is intended to contribute to the saving of human lives and the stabilisation of communities affected by UXO. Consequently, we provide full access to our code, including the custom knowledge graph and implementation functions. We also provide access to our dataset gathered in 4 years with unexploded bombs in real-world emergency situations. See Appendix D for more details.

2. Related Work

Current State-of-the-Art methods in image-based UXO assessment primarily rely on CNNs or Vision Transformers for object localisation and classification, without in-depth identification or explainability. To our knowledge, VLM-based approaches have received no attention in UXO mitigation, as deep learning for UXO tasks is still in its early stages, as a result of high risks and lack of representative datasets. The highly relevant nature of this field is underscored by recent *Horizon EU* research initiatives [2, 15] that

focus on the integration of machine learning and UXO detection. Specifically, these initiatives seek to explore machine learning techniques to reduce false cases and operational costs, with the overarching objective of developing clearance methodologies that are more efficient, more economical, and demonstrably safer than current practices. For this purpose, we adopt a modular design accessible without heavy computational resources.

In the past, researchers Blagojević et al. [7] used image processing techniques based on HSV analysis and thermal mapping to find unexploded munitions hidden in vegetation. The authors proposed an approach that uses the HSV colour space to generate a "green mask" that filters out grass, bushes, and trees from aerial optical imagery, pinpointing terrain areas more likely to contain UXO. Following this, researchers Dodić et al. [14] fine-tuned YOLO V5 [40] to find unexploded bombs, reporting over 90% mAP in controlled environments, but without explicit class identification in the field or interpretability. Although these methods offer theoretically promising metrics, actual UXOs are frequently encountered in far more challenging environments within former conflict zones (e.g., mountain trenches obscured by dense vegetation), not in controlled settings. Nevertheless, their findings demonstrate that UXOs consist of composable visual primitives, including artificial shape, particular texture, and characteristic elements, which together provide the theoretical foundation for our knowledge graph in the zero shot neuro symbolic framework [49]. The researchers Begkas et al. [6] introduced UXORD-10K, a dataset of over 10,000 images in 8 UXO categories, sourced from CAT-UXO [10] and relying on technical drawings and EOD manual images rather than field photography. Benchmarking CNNs and ViTs on this data achieved a Top-1 classification accuracy of 76.2%. In the context of autonomous robotic demining, Mishchuk et al. [32] conducted a comparative analysis between YOLOv8 [21] and RT-DETR [29, 63], reporting 78.32% mAP and 84.61% F1, modelling their dataset on MS COCO [28], using objects that share similar shapes with UXOs.

Further related work and its extended analysis can be found in the Appendix G, where we detailed neuro-symbolic approaches in similar safety-critical application [48, 60–62], emphasising the role of logical operators [22, 35, 45, 56] and UXO limitations. Regarding the real-world dataset used in the present study and the challenges related to its distribution for classical supervised methods, we have dedicated a separate section in Appendix E. It is worth noting that, in addition to the previously mentioned deep learning technical validations, disposal methodologies based on understanding primitives (shape, features, colours, condition, etc.) of munitions for identification are also referenced in specialised EOD manuals for human-based identification [17, 18, 43].

3. Methods

We introduce DEFUSAL, a neuro-symbolic framework that takes a UXO image as input and produces an interpretable ordnance identification accompanied by auditable rationales. We begin with the problem formulation and notation in Section 3.1, followed by the ontologically structured Knowledge Graph in Section 3.2, the PSL-based Consistency Validator in Section 3.3, the hybrid safety mechanism in Section 3.4, and finally the safeguards and uncertainty management procedures in Section 3.5.

3.1. Problem definition and notation

In this section, we formally define the safety-critical ordnance identification problem and introduce the necessary notation. Let $\mathcal{D} = \{(x_i, s_i)\}_{i=1}^{N}$ denote a dataset of N paired samples. The visual input space is defined as $\mathcal{X} \subset \mathbb{R}_{+}^{H\times W\times C}$, representing RGB images of height H, width W, and C channels. For each image $x \in \mathcal{X}$, there exists an associated element from the semantic space $\mathcal{S}$, which encodes in natural language either the initial inference prompt or the output generated by the hybrid feedback (safety) mechanism. The framework additionally incorporates an extensible Knowledge Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which captures relationships between visual evidence extracted from x and the corresponding UXO categories. We define the label space $\mathcal{Y} = \{y_1, \ldots, y_K\} \subset \mathcal{V}$, representing K distinct types of ordnance. Furthermore, we introduce a logic-agnostic attribute space $\mathcal{A} = \{a_1, \ldots, a_M\} \subset \mathcal{V}$, where each attribute corresponds to observable characteristics associated with the ordnance.

The system implements a composite mapping

$$f_\theta : \mathcal{X} \times \mathcal{S} \to \Delta^{K-1} \times [0,1]^M \times \Delta^4,$$

where Δ^n denotes the $(n+1)$-dimensional probability simplex. The semantic context $s \in \mathcal{S}$ guides the extraction of visual features from the input image $x \in \mathcal{X}$. The Vision–Language Model (VLM) produces confidence scores over the attribute space, while class distributions are derived through symbolic reasoning over the Knowledge Graph. The UXO type is inferred via a neuro-symbolic reasoning process that exploits structural dependencies in the Knowledge Graph to combine soft prediction scores from the identity space $\mathcal{Y}$ and the attribute space $\mathcal{A}$. To formalise UXO identification, we employ Probabilistic Soft Logic (PSL) [37], which defines a joint probability distribution over variables in the graph. Logical rules linking attributes to classes are expressed using Łukasiewicz t-norm logic. This relaxation allows reasoning over continuous truth values in the interval $[0, 1]$, derived directly from the confidence scores produced by the VLM, rather than relying on binary logic (see Appendix G). The final class prediction is obtained by solving a convex optimisation problem that computes the

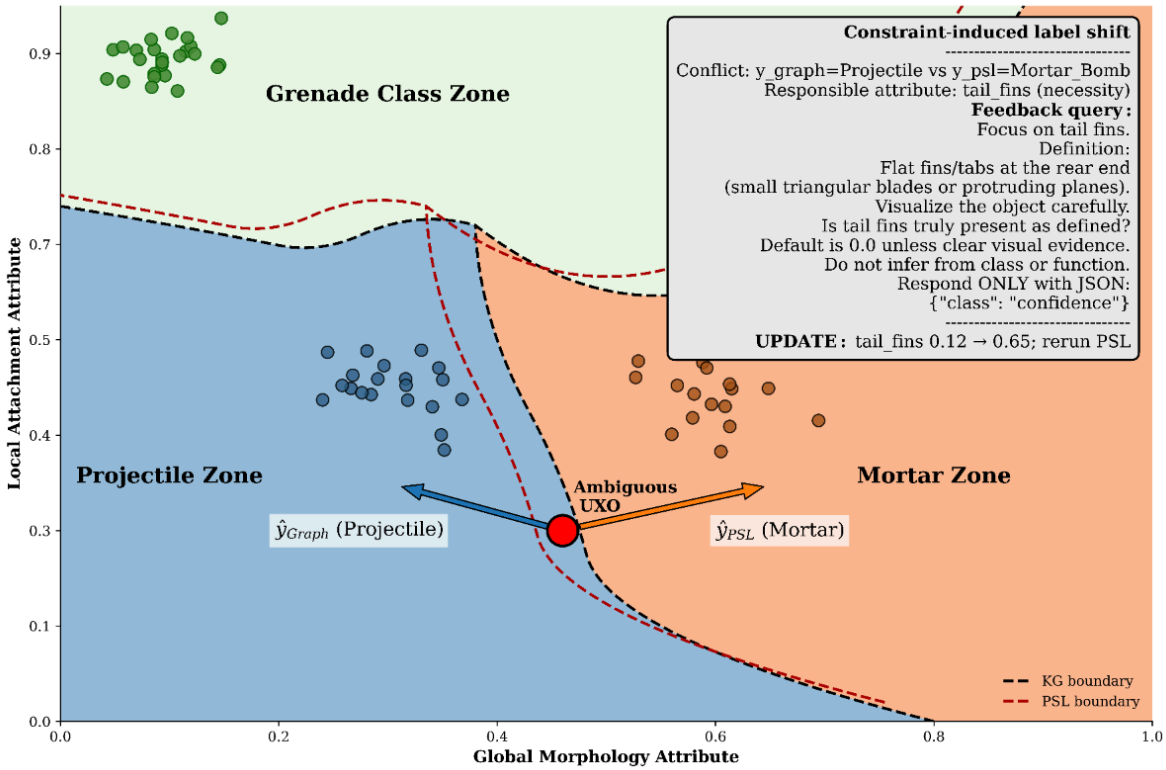


Figure 3. **Dimensionality reduced view to a space defined by 2 attributes.** The decision boundaries separate the graph-hypothesis space into distinct ordnance classes ($\mathcal{Y}$). We have contradictory hypotheses, $\hat{y}_{graph} = \textit{Projectile}$ and $\hat{y}_{PSL} = \textit{Mortar Bomb}$. We re-query and introduce new constraints $\hat{a}_{logic}$ to resolve the label shift and force convergence toward the correct class, such that $\hat{y}_{PSL2} = \hat{y}_{PSL1}$ or $\hat{y}_{PSL2} = \hat{y}_{graph}$.

Maximum a Posteriori (MAP) assignment while minimising violations of the semantic constraints. When the optimisation landscape exhibits multiple competing optima, the hybrid feedback mechanism refines the semantic space $\mathcal{S}$ by introducing auxiliary constraints that reduce ambiguity and enforce convergence.

A safety-critical condition, referred to as the *Uncertainty State*, is triggered when any of the following conditions occur: (i) the VLM fails to produce reliable attribute detections; (ii) the Knowledge Graph provides insufficient evidence for all candidate classes; (iii) the system oscillates between multiple hypotheses without converging to $\hat{y}_{\text{graph}} = \hat{y}_{\text{psl}}$; or (iv) the logical constraint requiring the cumulative confidence of required attributes to exceed that of forbidden attributes is violated. A qualitative illustration of this safety mechanism and the transition to the *Uncertainty State* is shown in Figure 3.

By adopting *Probabilistic Soft Logic (PSL)* as the reasoning backbone [1, 37], the framework preserves the fine-grained confidence scores obtained when parsing the VLM output through the Knowledge Graph and quantifies logical violations via "distance-to-satisfaction" penalties, thereby enabling the detection of *constraint-induced label changes*. Instead of directly predicting classification labels, the visual input x is processed to produce probabilities over primitive attributes relevant to UXO assessment. By evaluating the *inconsistency energy* of the logical constraints, the system detects misalignments between visual evidence and domain knowledge. Operating within an attribute-based zero-shot regime [3, 31, 38], the architecture illustrated in Figure 1 addresses the need for interpretability and safety-critical deployment through five integrated components: (1) a VLM Inspector acting as a knowledge-enhanced zero-shot primitive attribute extractor; (2) an ontologically structured Knowledge Graph; (3) a PSL-based consistency validator; (4) a hybrid safety mechanism; and (5) stability safeguards.

3.2. Ontologically-Structured Knowledge Graph

To represent the probabilistic output distributions of the Vision-Language Model (VLM) and the invariant physical laws governing Explosive Ordnance Disposal (EOD), we employ a Symbolic Knowledge Graph (KG) as a deterministic regularizer. For the graph $\mathcal{G}$, the set of vertices $\mathcal{V}$ is divided into disjoint subsets: Class nodes $\mathcal{V}_y \subset \mathcal{Y}$ (e.g.,

Mortar), attributes nodes $\mathcal{V}_a \subset \mathcal{A}$, For attribute space, we decomposed $\mathcal{A}$ into three topological categories essential for UXO generalisation: *Global Morphology* (e.g., elongated_cylindrical), *Local Attachments* (e.g., *tail_fins*), *Surface Topology* (e.g., segmented_body_pattern). Unlike standard classification paradigms where a neural network approximates a mapping $f : \mathcal{X} \to \mathcal{Y}$, our VLM functions as a primitive feature extractor that estimates a probability vector $\hat{\mathbf{a}} \in [0, 1]$ over the attribute space $\mathcal{A}$, treating class identities as latent variables constrained by a graph structure where the edge set $\mathcal{E}$ encodes domain expertise through specific ontological predicates. First, the HAS_PART or the necessity constraint encodes the implication $y \implies a$ to populate the set of required attributes $\mathcal{N}_{req}(y) = \{a \in \mathcal{A} \mid (y, a) \in \mathcal{E}_{\text{nec}}\}$, establishing, for example, that the existence of a *Mortar_Bomb* (y) logically requires the presence of a *teardrop_shape* attribute (a), where a violation occurs if the hypothesis is true (selected class y) while evidence is missing. Simultaneously, EXCLUDES or the disjointness constraint encodes the implication $y \implies \neg a$ to define the set of physically impossible configurations $\mathcal{N}_{forbid}(y) = \{a \in \mathcal{A} \mid (y, a) \in \mathcal{E}_{\text{exc}}\}$, such as *Projectiles* that do not possess structural *tail_fins*, which serves to mitigate visual hallucinations by penalising high confidence in incompatible features. To quantify the adherence of the VLM predictions (continuous probability distributions) to these rigid ontological constraints, we generalise the Boolean operators to a continuous interval by formulating them based on the Lukasiewicz implications, T-norms [25]. We define the implication operator $A \to B$ as $\min(1, 1 - A + B)$. For mutual exclusion constraints (C $\to \neg$A), the distance becomes $d_{\text{sat}}(C \to \neg A) = \max(0, C + A - 1)$, penalising configurations where both the class and the forbidden attribute are simultaneously highly confident. Consequently, the *distance to satisfaction* or the cost of violation is defined as the negation of the truth value: $d_{sat}(A \to B) = 1 - \min(1, 1 - A + B) = \max(0, A - B)$. We avoid imposing a hard decision boundary within the energy formulation itself. We define the graph-induced energy as follows:

$$\mathcal{J}_{KG}(y, \mathbf{a}) = \sum_{a_i \in \mathcal{N}_{req}(y)} (1 - I(a_i)) + \sum_{a_j \in \mathcal{N}_{forbid}(y)} I(a_j) \tag{1}$$

In Equation (1), the first term penalises missing required attributes (the lower $I(a_i)$, the higher the penalty), while the second term penalises the presence of forbidden attributes (the higher $I(a_j)$, the greater the penalty). The initial graph hypothesis is obtained by selecting the class with minimum energy (minimum contradiction):

$$\hat{y}_{graph} = \arg\min_{y \in \mathcal{Y}} \mathcal{J}_{KG}(y, \mathbf{a}) \tag{2}$$

If the minimum energy exceeds the number of required attributes for all classes (i.e., $\min_y \mathcal{J}_{KG}(y, \mathbf{a}) > |\mathcal{N}_{req}(y^*)|$), the Uncertainty State is triggered (see Section 3.5).

We note that the claim regarding the relevance of continuous values is validated in Section 4 through comparisons with hard logical binarization.

3.3. Consistency Validator

Trust in autonomous safety-critical systems requires rigorous verification beyond a simple statistical correlation. For solving this, we use *Probabilistic Soft Logic* [1], which models the joint probability distribution of the scene attributes and class labels using a *Hinge-Loss Markov Random Field (HL-MRF)* [4]. The system ingests the confidence scores from the graph as priors (observations O) and seeks to verify the consistency of the observed attributes against specific class hypotheses by defining a log-linear probability density function:

$$P(L|O) = \frac{1}{Z} \exp\left(-\sum_{j=1}^{m} \phi_j(L, O)\right) \tag{3}$$

In Equation (3), ϕ_j is a convex hinge-loss function derived from the Lukasiewicz operators, and Z is the partition function. We focus on two critical ontological constraints: *Necessity* (implication rules, e.g., $Class \Rightarrow Attribute$) and *Mutual Exclusion* (negative constraints, e.g., $Class \Rightarrow \neg Attribute$). The energy function aggregates the weighted violations:

$$\mathcal{J}_{PSL}(y, \mathbf{a}) = \sum_{a_i \in \mathcal{N}_{req}(y)} \max(0, P(y) - I(a_i)) + \sum_{a_j \in \mathcal{N}_{forbid}(y)} \max(0, P(y) + I(a_j) - 1) \tag{4}$$

The optimal $P(y)$ values are obtained by solving the convex optimisation $\min_P \sum_y \mathcal{J}_{PSL}(y, \mathbf{a})$ subject to $P(y) \geq 0$ and $\sum_y P(y) = 1$. Computational analysis of the optimisation is detailed in Appendix J. The predicted class is $\hat{y}_{PSL} = \arg\max_y P(y)$.

3.4. Hybrid Safety Mechanism

In direct connection with the safety mechanism described above, class labels are derived from the knowledge graph using a two-step process that produces two separate inference states. The first stage is for hypothesis construction, and the second stage is for enforcing global consistency. The initial state, the *Graph Hypothesis ($\hat{y}_{graph}$)*, is obtained by propagating VLM-derived attribute scores through the graph in a forward pass that minimises energy, as exemplified in Equation (1) and Equation (2). The second state, the *logic consistent state ($\hat{y}_{psl}$)*, is a revised class assignment generated by the PSL engine, which performs global constraint optimisation by minimising the HL-MRF energy to reconcile the initial hypothesis with strict ontological rules

of necessity and mutual exclusion The safety (feedback) mechanism is activated strictly when the rigorous application of probabilistic logic forces a *semantic change* from the initial graph hypothesis, as exemplified in Figure 3. This divergence indicates that the system is operating at a fragile decision boundary. Instead of forcing a potentially erroneous decision, the feedback mechanism queries the VLM inspector, targeting the specific constraint c^* responsible for the energy shift. A *constraint-induced label shift* occurs when:

$$\hat{y}_{graph} \neq \hat{y}_{PSL}, \text{ where } \begin{cases} \hat{y}_{graph} = \arg\min_{y \in \mathcal{Y}} \mathcal{J}_{KG}(y, \mathbf{a}) \\ \hat{y}_{PSL} = \arg\min_{y \in \mathcal{Y}} \mathcal{J}_{PSL}(y, \mathbf{a}) \end{cases} \tag{5}$$

Both the graph hypothesis and the PSL inference minimise energy and a label shift occurs when they select different classes. In Equation 5, $\mathcal{J}_{KG}$ is the graph-induced energy as defined in Equation (1). When a label shift is detected, the system identifies the attribute most responsible for the inconsistency by computing the contribution of each attribute to the total energy:

$$a* = \arg\max_a 1[violation(a) > 0] \tag{6}$$

For interpretability, we assume uniform rule weights. As a result, our analysis captures the structural sources of inconsistency. The mechanism identifies the atomic attribute a^* that contributes the maximum gradient to $\mathcal{J}_{PSL}$ and generates a targeted query based on the violation of the specific restriction. If the penalty stems from the necessity term in Equation (4), the system issues a *completeness query* (*"[...] Is there any {attribute} for this object?"*), whereas a penalty arising from the exclusion term triggers a *contradiction query* (*"[...] You reported {attribute}. Verify if this feature is clutter."*). The refined VLM response updates the soft truth value $I(a^*) \leftarrow I'(a^*)$, and the PSL inference is re-executed with the corrected evidence.

3.5. Safeguards and Uncertainty Management

To ensure stability and adherence to the *Fail-Safe* principle, we enforce rigorous constraints on the refinement trajectory. We claim that a false negative outcome is not desirable, whereas a manual review request is an acceptable operational cost. We define an *Uncertainty State*, $\mathcal{U}$, which overrides the classification output. The system enters this state under the condition of persistent label shift or if the minimum energy is too high. Specifically for the energy statement, if $\min_y \mathcal{J}_{KG}(y, \mathbf{a}) > |\mathcal{N}_{req}(y^*)|$, it implies that the magnitude of violation exceeds the signal of the required attributes, indicating that the VLM did not detect sufficient evidence for any valid hypothesis (the input image is likely unclear or contains an unknown object). Additionally, if the PSL-induced class assignment $\hat{y}_{PSL}$ differs from the initial graph hypothesis $\hat{y}_{graph}$ after the first feedback iteration, the system performs a targeted re-query (Subsection 3.4). If this second inference still produces a different label ($\hat{y}_{graph} \neq \hat{y}^{(1)}_{PSL} \neq \hat{y}^{(2)}_{PSL}$), the inconsistency indicates irresolvable visual ambiguity (e.g., severe occlusion, improvised device, unknown class for the defined knowledge graph), and the sample is escalated to human review. See Appendix B and Appendix C for more details regarding the need for a feedback mechanism.

4. Experiments

As qualitative analysis, in Figure 4 we included an example for the proposed framework. Further results are provided in Appendix H. We evaluate the neuro-symbolic system on a test dataset [12] comprising 13,648 instances, employing a knowledge-enhanced zero-shot regime. Unexploded bomb identification is performed for $K = 9$ distinct ordnance types: *Mortar*, *Projectile*, *Grenade*, *Rocket*, *Aviation_Bomb*, *Mine*, *RPG*, *Cartridge*, and *AntiSubmarine_Bomb*. We benchmarked our approach against open-weight foundation models, following zero-shot inference paradigm: (i) Gemma 3 27B Instruct [46], (ii) Qwen 3 VL 32B Instruct [59], and (iii) GLM 4.6-V-Flash [47]. Multiple complementary metrics are reported. F1-Score is defined as the harmonic mean of precision and recall, computed on all test samples. In the neuro-symbolic pipeline, any sample routed to $\mathcal{U}$ (uncertainty) is counted as a misclassification, allowing for a fair comparison with baselines. In safety-critical identification, Recall is a central metric, measuring the share of true positives correctly detected among all actual positives. For UXO identification, a False Negative (failing to detect an ordnance) poses an intolerable operational risk with potentially irreversible consequences. This aspect underpins our Uncertainty State mechanism, which intentionally routes ambiguous cases to human review instead of risking an erroneous automatic decision. Consequently, we report the False Negative Rate (FNR) alongside F1-Score to explicitly characterise how effectively the system reduces missed detections in real-world EOD operations. F1-Confident is computed only on samples for which the system issues a concrete label, by excluding $\mathcal{U}$, highlighting that the neuro-symbolic method retains good metrics whenever it makes a decision. Lastly, the Human Review Rate (HRR) measures the proportion of samples assigned to the Uncertainty State $\mathcal{U}$ due to rule conflicts or missing defining primitives. All such samples are deferred for human review. A higher HRR indicates a more conservative system that prioritises safety by avoiding automatic decisions under inconsistency, whereas a lower HRR reflects increased automation at the potential cost of risk. The baseline method evaluates how accurately the pre-trained model can identify UXO types. Appendix H. To advance our analysis and allow for intermediate steps in examining

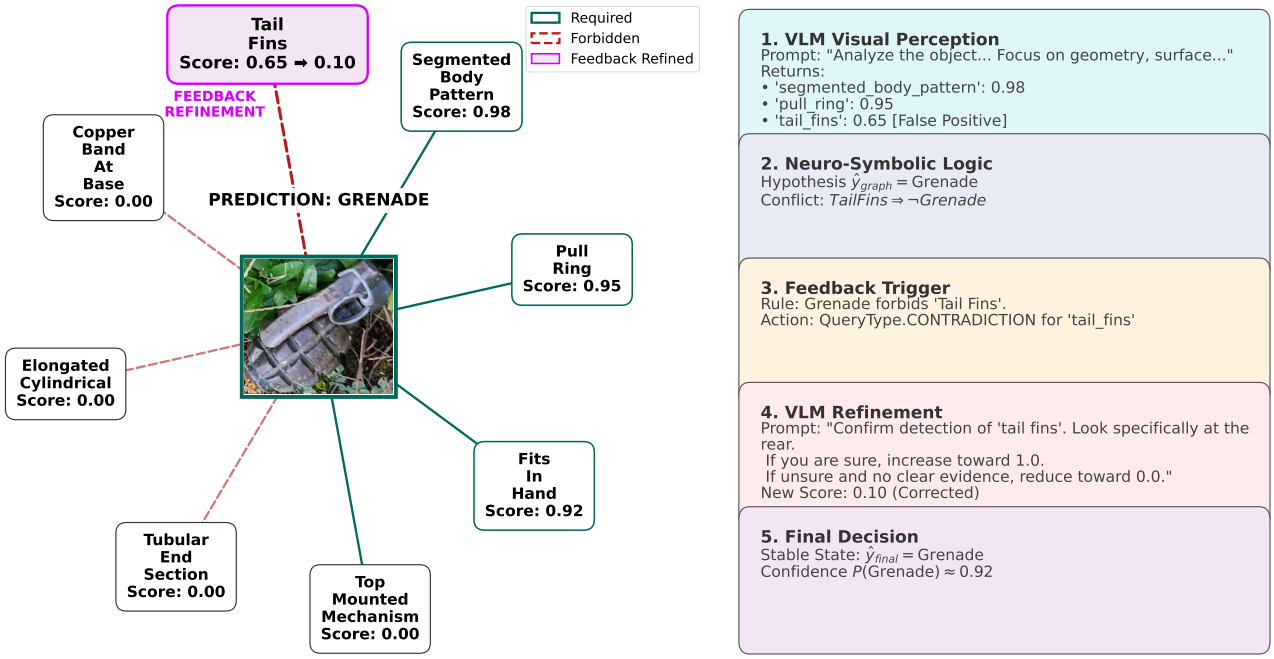


Figure 4. **Qualitative result (Qwen 3 VL 32B Instruct).** An initial visual indication of tail fins conflicts with the grenade hypothesis encoded in the knowledge graph, $\hat{y}_{graph} \neq \hat{y}_{\text{PSL}}$. A contradiction-triggered refinement step prompts a targeted visual re-evaluation, reducing the confidence of the primitive attribute from 0.65 to 0.10, thereby resolving the inconsistency, $\hat{y}^{(2)}_{PSL} = \hat{y}_{graph}$.

Table 1. Comparative results on the UXO dataset ($N = 13{,}648$) averaged over 10 experiments per configuration. Results are reported as *mean* $\pm$ *std. dev*. For the baseline models, the F1-Confident is identical to the F1-Score, since there is no Uncertainty State.

Method	F1-Score	Recall	F1-Confident	HRR
Zero-Shot (Baselines)				
GLM4.6-V-FLash	0.356 ± 0.02	0.341 ± 0.01	0.356 ± 0.02	-
Gemma3 27B Instruct	0.372 ± 0.01	0.358 ± 0.02	0.372 ± 0.01	-
Qwen3 VL 32B Instruct	0.526 ± 0.01	0.512 ± 0.01	0.526 ± 0.01	-
Zero-Shot + Attributes for Classes				
GLM4.6-V-FLash	0.508 ± 0.02	0.490 ± 0.02	0.521 ± 0.04	$2.4 \pm 0.4\%$
Gemma3 27B Instruct	0.510 ± 0.02	0.492 ± 0.02	0.524 ± 0.03	$2.7 \pm 0.3\%$
Qwen3 VL 32B Instruct	0.603 ± 0.02	0.585 ± 0.02	0.621 ± 0.03	$2.9 \pm 0.3\%$
Our Framework				
GLM4.6-V-FLash	0.543 ± 0.01	0.610 ± 0.01	0.668 ± 0.01	$18.7 \pm 0.2\%$
Gemma3 27B Instruct	0.576 ± 0.01	0.680 ± 0.01	0.610 ± 0.01	$5.50 \pm 0.2\%$
Qwen3 VL 32B Instruct	0.640 ± 0.01	0.649 ± 0.01	0.853 ± 0.03	$25.0 \pm 0.3\%$

Table 2. Comparisons with traditional type neuro-symbolic engine. Continuous logic consistently outperforms hard thresholding across all metrics.

Reasoning	F1-Score	Recall	F1-Confident	HRR	FNR
Hard Binarization	0.622	0.569	0.828	25.0%	14.3%
DEFUSAL (Ours)	**0.640**	**0.649**	**0.853**	25.0%	**13.5%**

the influence of symbolic attributes, we conducted experiments with a knowledge-enhanced prompt that incorporates class-specific attributes. For a fair comparison, we used the same attributes as those defined in the Knowledge Graph.

4.1. Results

Table 1 reports the comparative performance. For Qwen3, the neuro-symbolic pipeline attains an overall F1-Score of 0.640 (vs. a 0.526 baseline; +21.7% relative), while de-

ferring decisions on 25% of instances (HRR). On the subset of accepted predictions, performance remains high (F1-Confident = 0.853), demonstrating that the system is accurate when it decides and cautious when it does not. The knowledge-guided prompt alone already enhances identification (F1-Score 0.603), suggesting improved separation between classes. By contrast, the baseline's requirement to predict on every sample reduces F1-Score, underscoring that VLMs are miscalibrated (overly biased toward the head of the long-tail UXO distribution) in safety-critical scenarios.

To validate our choice of continuous fuzzy logic over hard binarization, we compare two reasoning configurations using the same Knowledge Graph and Qwen3 VL 32B Instruct: (i) *Hard Binarization* — VLM outputs are thresholded at 0.5 and processed through a standard boolean ruleset, and (ii) *Łukasiewicz PSL* — our proposed continuous reasoning engine. To explicitly assess long-tail robustness, Table 2 from Section 1 reports per-class F1 scores for the zero-shot baseline and our entire system. A critical design consideration for safety-critical systems is the ability to tune the balance between automation coverage and risk tolerance. The Uncertainty State can be governed by a scaling factor α, such that a sample is deferred to human review when:

$$\min_{y} \mathcal{J}_{KG}(y, \mathbf{a}) > \alpha \cdot |\mathcal{N}_{req}(y^*)| \tag{7}$$

In Equation (7), setting $\alpha = 1.0$ defines our default configuration, as presented in Table 1. We adopt uniform weights for interpretability. For a complete analysis, we also vary α to examine how different conditions affect the results. Changing α, an EOD operator can continuously trade automation coverage (lower HRR) for safety (lower FNR), without any retraining. Figure 5 illustrates this trade-off for Qwen3 VL 32B Instruct in ten configurations. This controllability is a key property for real-world EOD deployment, where mission context dictates acceptable risk levels. Also, please see the Appendix C.

Examples on difficult or rare instances can be found in Appendix H. The summary of the results per class can be found in the introductory section (Section 1), Figure 2.

4.2. Ablation Study

To validate our architectural choices, we analyse the contribution of each module. Table 3 reveals a consistent pattern: each module incrementally improves identification (F1-Score) while progressively converting hazardous false negatives into human-reviewable deferrals. Visual attributes alone reduce FNR by 8.7 percentage points with negligible coverage loss (HRR = 2.9%). The knowledge graph introduces a sharper safety–coverage trade-off, and the full system achieves the strongest safety profile (FNR 13.5%, F1-Confident 0.853), confirming that feedback-driven rejection meaningfully improves overall reliability.

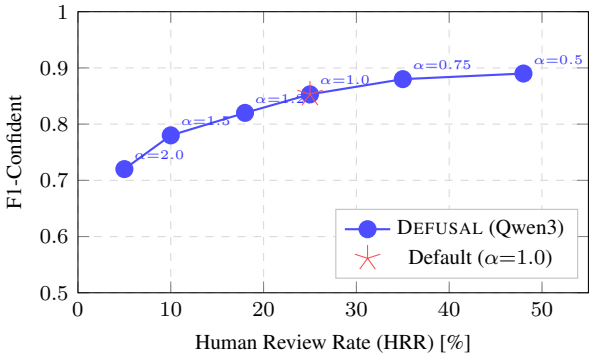


Figure 5. **Coverage–Risk trade-off for DEFUSAL (Qwen3 VL 32B).** Each point corresponds to a distinct uncertainty threshold α in Equation (7). Reducing α leads to deferring more samples (higher HRR), keeping only the most confident predictions and thereby increasing F1-Confident. The red star indicates the default setting (α=1.0, F1-Conf=0.853, HRR=25%).

Table 3. Ablation Study (Qwen3 VL 32B Instruct), impact of each component of the framework.

Component	F1-Score	F1-Confident	HRR	FNR
Baseline VLM	0.526	0.526	0.0%	43.6%
+ Visual Attributes	0.603	0.621	2.9%	34.9%
+ Knowledge Graph	0.622	0.711	12.5%	26.5%
Full System	0.640	0.853	25.0%	13.5%

5. Conclusion

In this paper, we present a neuro-symbolic framework, DEFUSAL, that bridges the semantic gap between the flexible perceptual capabilities of Vision–Language Models and the stringent safety requirements of EOD missions. Our approach decouples perception (low-level attribute extraction) from reasoning (logical constraint satisfaction), yielding three key advantages. First, it enhances safety through a logic-based rejection mechanism that reduces false negatives for long-tail UXO categories. Second, the framework generalises without additional training: new ordnance types can be incorporated simply by updating the Knowledge Graph, mitigating data scarcity for rare classes. Third, the pipeline is inherently interpretable—each identification or rejection is supported by explicit logical rules and traceable attribute evidence, reducing reliance on opaque confidence scores and limiting spurious or hallucinated reasoning. We invite readers to consult the reproducibility statement in Appendix A, the human-in-the-loop recommendations in Appendix C, and the discussion of future research directions in Appendix K.

Acknowledgements

We gratefully acknowledge the support of the University POLITEHNICA of Bucharest for financial support of this publication through the PubArt program.

References

[1] Kareem Ahmed, Stefano Teso, Kai-Wei Chang, Guy Van den Broeck, and Antonio Vergari. Semantic probabilistic layers for neuro-symbolic learning. *Advances in Neural Information Processing Systems*, 35:29944–29959, 2022. 4, 5

[2] ASSETS Plus. Identify, inspect, neutralise unexploded ordnance (UXO) at sea, 2025. Accessed: 2026-01-02. 2

[3] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. *Advances in Neural Information Processing Systems*, 33:1462–1473, 2020. 4

[4] Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research*, 18(109): 1–67, 2017. 5, 17

[5] Guangsheng Bao, Hongbo Zhang, Cunxiang Wang, Linyi Yang, and Yue Zhang. How likely do llms with cot mimic human reasoning? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7831–7850, 2025. 2

[6] Georgios Begkas, Panagiotis Giannakeris, Konstantinos Ioannidis, Georgios Kalpakis, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris. Automatic visual recognition of unexploded ordnances using supervised deep learning. In *Association for Computing Machinery*, page 286–294, New York, NY, USA, 2022. 3

[7] Dejan Blagojević, Dejan Dodić, Bojan Glamoclija, and Jelena Krstic. Application of hsv color analysis and green masking for uxo detection in dense vegetations. In *nternational Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, pages 1–4, 2025. 3

[8] Malik Boudiaf, Imtiaz Ziko, Jérôme Rony, José Dolz, Pablo Piantanida, and Ismail Ben Ayed. Information maximization for few-shot learning. *Advances in Neural Information Processing Systems*, 33:2445–2457, 2020. 15

[9] Eduardo Brito and Henri Iser. Maxsime: Explaining transformer-based semantic similarity via contextualized best matching token pairs. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2154–2158, 2023. 14

[10] CAT-UXO. Collective awareness to unexploded ordnance (cat-uxo), 2026. Accessed: 2026-01-04. 3

[11] Liang Chen, Zekun Wang, Shuhuai Ren, Lei Li, Haozhe Zhao, Yunshui Li, Zefan Cai, Hongcheng Guo, Lei Zhang, Yizhe Xiong, et al. Next token prediction towards multimodal intelligence: A comprehensive survey. *arXiv preprint arXiv:2412.18619*, 2024. 14

[12] Gheorghe Marian Craioveanu and Grigore Stamatescu. Ctx-uxo: A comprehensive dataset for detection and identification of unexploded ordnances, 2024. 6

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 14

[14] Dejan Dodić, Dejan Blagojević, Nikola Milutinović, Aleksandar Milić, and Bojan Glamočlija. Contribution of the yolo model to the uxo detection process. In *2025 24th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pages 1–6. IEEE, 2025. 3

[15] European Commission. Topic HORIZON-CL3-2023-BM-01-02: Identify, inspect, neutralise unexploded ordnance (UXO) at sea. Funding & Tenders Portal, 2023. Accessed: 2026-01-02. 2

[16] James Gallagher. Adaptive multispectral landmine identification dataset (amlid), 2025. 15

[17] Geneva International Centre for Humanitarian Demining (GICHD). *Explosive Ordnance Guide for Ukraine*. Geneva International Centre for Humanitarian Demining, Geneva, Switzerland, third edition edition, 2023. Third edition, web version. 2, 3, 13

[18] Thomas Gersbeck. *Practical military ordnance identification*. CRC Press, 2019. 2, 3, 13

[19] Oihane Gómez-Carmona, Diego Casado-Mansilla, Diego Lopez-de Ipina, and Javier García-Zubia. Human-in-the-loop machine learning: Reconceptualizing the role of the user in interactive approaches. *Internet of Things*, 25: 101048, 2024. 13

[20] International Committee of the Red Cross. Explosive remnants of war, 2025. Accessed: 2025-01-02. 1

[21] B Karthika, M Dharssinee, V Reshma, R Venkatesan, and R Sujarani. Object detection using yolo-v8. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–4, 2024. 3

[22] Markus Krötzsch, Maximilian Marx, Ana Ozaki, and Veronika Thost. Attributed description logics: Ontologies for knowledge graphs. In *International Semantic Web Conference*, pages 418–435. Springer, 2017. 3, 16

[23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 13

[24] Lao National Unexploded Ordnance Programme. Organization and background: The problem, 2026. Accessed: 2026-01-02. 1

[25] Carlos Leandro. Symbolic knowledge extraction using {\L} ukasiewicz logics. *arXiv preprint arXiv:1604.03099*, 2016. 5

[26] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6970–6979, 2022. 14

[27] Qing Li, Siyuan Huang, Yining Hong, Yixin Chen, Ying Nian Wu, and Song-Chun Zhu. Closed loop neural-symbolic learning via integrating neural perception, grammar parsing, and symbolic reasoning. In *International Conference on Machine Learning*, pages 5884–5894. PMLR, 2020. 19

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3

[29] Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer, 2024. 3

[30] Ben Malin, Tatiana Kalganova, and Nikolaos Boulgouris. A review of faithfulness metrics for hallucination assessment in large language models. *IEEE Journal of Selected Topics in Signal Processing*, 2025. 14

[31] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5222–5230, 2021. 4

[32] Vadym Mishchuk, Herman Fesenko, and Vyacheslav Kharchenko. Deep learning models for detection of explosive ordnance using autonomous robotic systems: trade-off between accuracy and real-time processing speed. *Radioelectronic and Computer Systems*, 2024:99–111, 2024. 3

[33] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023. 13

[34] North Atlantic Treaty Organization. *Glossary of Terms and Definitions*. NATO Standardization Agency, modified version 02 edition, 2000. Updated 07.08.2000. 1

[35] Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, 2023. 3, 16

[36] Wanli Ouyang, Xiaogang Wang, Cong Zhang, and Xiaokang Yang. Factors in finetuning deep model for object detection with long-tail distribution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 864–873, 2016. 14

[37] Connor Pryor, Charles Dickens, Eriq Augustine, Alon Albalak, William Wang, and Lise Getoor. Neupsl: Neural probabilistic soft logic. *arXiv preprint arXiv:2205.14268*, 2022. 3, 4, 17

[38] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3593–3602, 2019. 4

[39] Shuhan Qi, Zhengying Cao, Jun Rao, Lei Wang, Jing Xiao, and Xuan Wang. What is the limitation of multimodal llms? a deeper look into multimodal llms through prompt probing. *Information Processing & Management*, 60(6):103510, 2023. 14

[40] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015. 3

[41] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 2, 15

[42] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1):107–136, 2006. 16

[43] Janet E Simms, Robert J Larson, William Lee Murphy, Dwain K Butler, et al. *Guidelines for planning unexploded ordnance (UXO) detection surveys*. Geotechnical and Structures Laboratory (US), 2004. 2, 3, 13

[44] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 15

[45] Giuseppe Spillo, Cataldo Musto, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. Recommender systems based on neuro-symbolic knowledge graph embeddings encoding first-order logic rules. *User Modeling and User-Adapted Interaction*, 34(5):2039–2083, 2024. 3, 16

[46] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotin-

der Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. 6

[47] V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Haochen Li, Jiale Zhu, Jiali Chen, Jiaxing Xu, Jiazheng Xu, Jing Chen, Jinghao Lin, Jinhao Chen, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Ruiliang Lyu, Shangqin Tu, Sheng Yang, Shengbiao Meng, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wei Jia, Wenkai Li, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyu Zhang, Xinyue Fan, Xuancheng Huang, Yadong Xue, Yanfeng Wang, Yanling Wang, Yanzi Wang, Yifan An, Yifan Du, Yiheng Huang, Yilin Niu, Yiming Shi, Yu Wang, Yuan Wang, Yuanchang Yue, Yuchen Li, Yusen Liu, Yutao Zhang, Yuting Wang, Yuxuan Zhang, Zhao Xue, Zhengxiao Du, Zhenyu Hou, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026. 6

[48] Spyros Theodoropoulos, Georgios Makridis, Dimosthenis Kyriazis, and Panayiotis Tsanakas. Robust novel defect detection with neurosymbolic ai. In *IFIP International Conference on Advances in Production Management Systems*, pages 381–396. Springer, 2024. 3, 15

[49] Efthymia Tsamoura, Timothy Hospedales, and Loizos Michael. Neural-symbolic integration: A compositional perspective. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5051–5060, 2021. 3

[50] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023. 14

[51] United Nations. Protocol on explosive remnants of war to the convention on prohibitions or restrictions on the use of certain conventional weapons which may be deemed to be excessively injurious or to have indiscriminate effects (protocol v). United Nations Document CCW/P.V/CONF/2, 2003. Adopted at Geneva, 28 November 2003. 1

[52] United Nations Development Programme. Uxo project document 2022-2026: Supporting effectiveness and efficiency in the uxo sector to contribute to the achievement of sdg 18 and safe path forward iii. Project document, United Nations Development Programme, Vientiane, Lao PDR, 2022. Accessed: 2026-01-02. 1

[53] Olivier Veilleux, Malik Boudiaf, Pablo Piantanida, and Ismail Ben Ayed. Realistic evaluation of transductive few-shot learning. *Advances in Neural Information Processing Systems*, 34:9290–9302, 2021. 15

[54] Shuihua Wang and Yudong Zhang. Grad-cam: understanding ai models. *Comput. Mater. Contin*, 76(2):1321–1324, 2023. 15

[55] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens Van Der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019. 15

[56] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, 2014. 3, 16

[57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2, 14

[58] Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. Adversarial robustness under long-tailed distribution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8659–8668, 2021. 14

[59] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. 6

[60] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3208–3216, 2021. 3, 16

[61] Hongda Zhang, Carlos S Lima, Rafael Samorinha, Adriano Tavares, Fausto Giunchiglia, Donglei Song, Hao Xu, and Deming Guo. A memory-assisted neuro-symbolic approach for chestradiography classification. *Available at SSRN 5388368*, 2025. 15

[62] Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023. 3, 16

[63] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection, 2023. 3

[64] Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36:44502–44523, 2023. 2

Look, Reason, Defuse: Bridging Perception and Domain Knowledge for Real-World Unexploded Ordnance Identification

Supplementary Material

Appendix

The following sections provide additional information and complement the main paper:

A. Reproducibility Statement

The proposed system is designed to be architecture-agnostic, enabling the integration of any model provided it is accessible through an OpenAI-compatible API or the HuggingFace Transformers framework. For the experiments reported in Section 4, all models were deployed locally using PagedAttention vLLM [23]. The framework does not impose specific hardware requirements. Knowledge nodes are defined through a configurable JSON structure that is parsed by the system, allowing users to easily adapt or extend the knowledge base. To ensure cross-platform compatibility with respect to package dependencies and versioning, the framework was validated across multiple environments, including Windows 11, Ubuntu 25.10, and Fedora 40, with and without GPU support (via public cloud providers). Details on code availability and dataset access are provided in Appendix D. The dataset is described in Appendix E. Notation is introduced in Section 3.1, and additional qualitative results are presented in Appendix H.

B. Limitations

The external reasoning capability depends on the coverage of the defined Knowledge Graph. If a rare munition or a novel improvised explosive device (IED) is absent from the symbolic taxonomy, the Consistency Checker cannot evaluate it and therefore assigns high uncertainty. However, the Knowledge Graph can be easily extended through the JSON-based representation, allowing new ordnance types to be incorporated without requiring additional training data. In this sense, the system can generalise without retraining. Second, although the implemented safeguards reduce the risk of reasoning loops, the underlying VLM remains stochastic. As a result, the model may occasionally oscillate between two similarly plausible yet incorrect interpretations that remain logically consistent.

C. Human-In-The-Loop Recommendation

We emphasise that this work is designed as a Decision Support System (DSS) and must never autonomously execute the final decision in a defusal operation. The neuro-symbolic framework is intended to augment the cognitive capabilities of human EOD operators, not to replace them. Given the severe consequences of error in the disposal of unexploded ordnance, the final verification and neutralisation decision must remain the responsibility of a qualified specialist. The system is designed to highlight and explain potential risks that human operators might overlook due to fatigue or distraction, effectively acting as an always-on automated "second pair of eyes" for UXO assessment. A human-in-the-loop methodology [19, 33] is therefore strongly recommended for all final UXO-related decisions.

D. Code and Datasets

To foster transparency and accelerate safety research in the humanitarian sector, we make all assets available:

- **Code**: The complete implementation of the Neuro-Symbolic Loop, including the Knowledge Graph, energy functions, and feedback mechanism, is released under **Creative Commons Attribution 4.0 International** at Anonymous GitHub. We explicitly encourage community audits of the safety logic and submit any request for improvements.
- **Dataset**: The entire dataset, including additional classes, can be found at IEEE Dataport, Zenodo or Hugging Face. We used a custom repository of images collected from real-world pyrotechnic interventions. The dataset was annotated by a certified EOD specialist. Specialised technical manuals [17, 18, 43] were also used to validate information on ammunition that the labelling expert was unfamiliar or uncertain about. This dataset explicitly addresses the real distribution of munitions found in post-conflict zones. The complete dataset and its description were validated in the past. We provide a detailed description and a representative subset in Appendix E.

E. Long-Tailed Distributions in UXO Context

Classical supervised models for localisation, recognition and specific identification methods often require a large, representative dataset that is balanced between classes. However, in practice, this is difficult to achieve for UXO, as the prevalence of munitions varies. In statistics, this distribution phenomenon is known as the long-tail effect [36]. Long-tailed distributions are frequently encountered in environments that mirror real-world situations (primarily in finance and industry) and represent a significant challenge [26], which is itself the subject of research. Although primary metrics may appear promising, systems often fail during real-world deployment due to the lack of a representative test set [58]. Addressing this issue in the context of object detection, researchers [36] attempted a hierarchical feature learning approach, achieving a 4.7% improvement in mAP over baseline for the ImageNet dataset [13].

The current study uses a real-world dataset that has previously been validated. The long-tailed distribution of this dataset is illustrated in Figure 6. The real-world UXO prevalence is characterised by a multi-level long-tailed distribution. First, the domain of unexploded weapons/ordnance is sparsely represented in the general training corpora of foundation models, making it a rare and specialised knowledge area. Second, the intra-domain class distribution is highly imbalanced, with a long tail of rare UXO types/categories. The dataset is representative and consists of actual UXO items collected over a four-year period. The dataset $\mathcal{D} = \{(x_i, y_i, \mathbf{a}_i)\}_{i=1}^{N}$ comprises $N = 13,648$ annotated ordnance instances. Data were collected over a four-year period from active demining sectors, reflecting the operational reality of field interventions targeting remnants of the World Wars. Consequently, the class distribution follows a heavy-tailed Zipfian power law (validated with $R^2 \approx 0.827$):

$$P(y) \propto \text{rank}(y)^{-\alpha}, \quad \alpha \approx 3.19 \tag{8}$$

As shown in Figure 6, the dataset is heavily biassed toward the "head" categories, with classes such as *Projectiles*, *Mortar Bombs*, and *Grenades* accounting for 89% of all samples. In contrast, certain other categories, for instance *Mines (Land Mines and Naval Mines* ($n = 34$), fall into the far end of the distribution tail and are represented only by a very small number of examples. This distribution validates the necessity for a logic-driven, zero-shot capable framework, as standard data-driven learning fails to generalise to these rare instances. We observe that in more recent conflicts, a broader variety of UXO types has emerged, including adapted or modified versions. The strength of the current system is that it is straightforward to update (via KG defined in JSON format) to generalise to these new types.

In Figure 7, we added a sample for each class from the used dataset.

F. Use of Foundation Models

Within the landscape of multimodal models, the fundamental distinction between classical Internal Reasoning, for example Chain-of-Thought [57] and our proposed closed loop feedback lies in the validation of the facts while forcing the logic-guidance (no unfaithful reasoning). Although internal reasoning unfolds as a linear probabilistic path through latent space, producing a smooth, seemingly coherent chain of logic that can suffer from hallucinatory consistency [30, 50] and gradual semantic drift when visual attributes are missing, our approach instead functions as an external stepwise reasoning procedure that breaks this process into discrete, verifiable stages both for the framework itself and for the human operator, stages specially designed for UXO identification. By fragmenting the decision, the process is made into a series of iterative queries (e.g., '[...] Is there a [attribute]? Look for the following criteria: guidance_for_attributes'), the feedback loop forces the model to break free from linguistic inertia and execute a mandatory visual re-grounding within the image input tensors at every step. Each step can be inspected by machine and by human.

We hypothesise that, although general-purpose Foundation Models (FMs) exhibit strong zero-shot performance on everyday objects, they are fundamentally unsuitable for direct use in Unexploded Ordnance (UXO) mitigation unless heavily constrained or adapted to the domain, being difficult due to missing representative and balanced datasets. This hypothesis is based on three fundamental inconsistencies between the training objectives of the foundation model and the constraints of high-stakes safety environments, originating from distributional shift and long-tailed scarcity. Specifically, the pretraining distribution $\mathcal{D}_{train}$ assigns a statistically negligible probability mass to UXO imagery ($P(x_{UXO}) \approx \epsilon$), causing models to default to high-probability priors of everyday objects or most common UXO (see Appendix E), due to feature space overlaps in surface texture. This representational failure is severely exacerbated by the intrinsic incompatibility of probabilistic generation with safety guarantees. Current Vision-Language Models (VLMs) are trained by maximising token-sequence likelihood [9, 11, 39], $\max_\theta \sum \log P(w_t | w_{<t}, \mathbf{I})$, a stochastic procedure that favours semantic plausibility over factual accuracy. thus lacking the specific framework/methods required to minimise the False Negative Rate essential for defusal operations. Furthermore, base models (object classification/detectors) without internal reasoning rely on superficial manifold mapping rather than causal physical reasoning, effectively mapping visual inputs to semantic labels $f : \mathcal{X} \to \mathcal{Y}$, a key shortcoming that our logic-guided framework overcomes by embedding structured domain knowledge directly into at test-time.

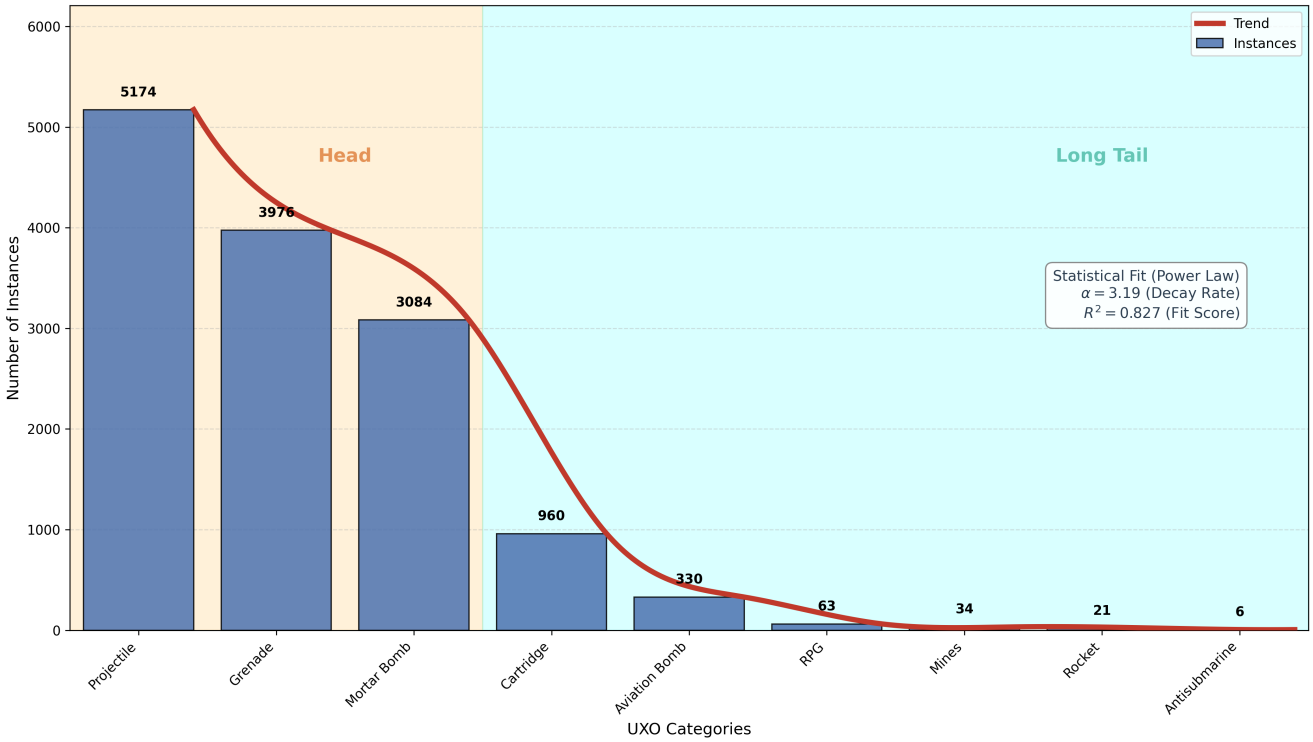
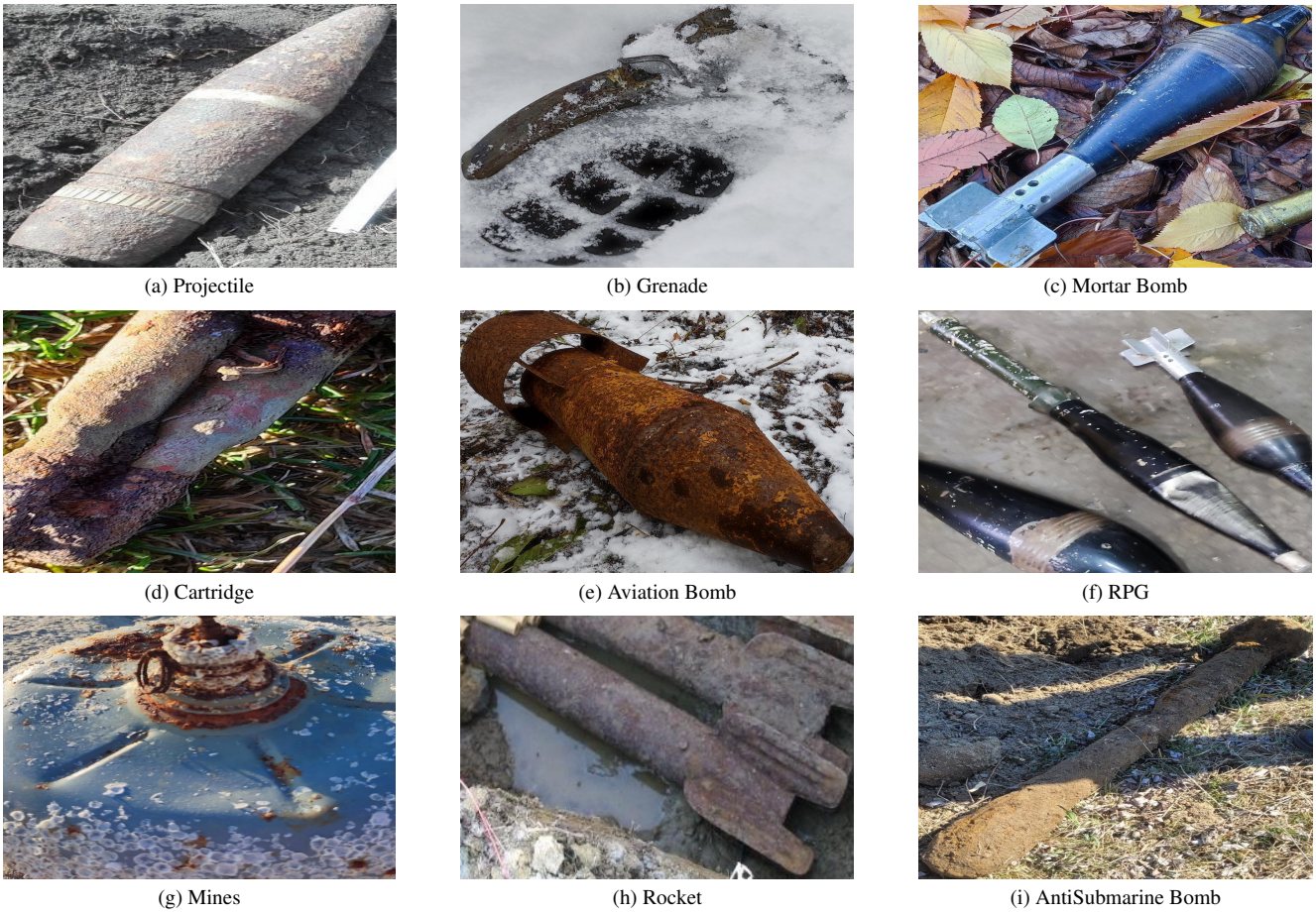


Figure 6. Unexploded Ordnances Distribution. While this distribution characterises real-world historical contamination, being the distribution from real pyrotechnic intervention, we acknowledge that in recent asymmetric conflicts, the *Landmine* class has seen a proliferation in usage and priority for intelligent analysis [16], effectively shifting it towards the 'head' of the distribution in modern operational theatres. The tail can be extended with UXOs from more recent conflicts (eg. submunitions, bombs modified to be used by UAVs). The system is capable of generalising by making updates to the knowledge graph, including new types of UXOs. New fields are automatically interpreted and incorporated into described logical mechanisms

G. Additional Related Works

The necessity for interpretability and robustness has driven the adoption of neuro-symbolic AI in safety-critical fields, such as medical imaging and industrial inspection. For example, knowledge graphs have been used to guide chest radiograph diagnosis, ensuring that neural predictions align with medical ontologies [61]. They achieved a 4% gain in F1-Score along with improved explainability. Similarly, in industrial settings, researchers [48] introduced a neurosymbolic framework using Logic Tensor Networks (LTN) that effectively bridges the performance gap in open-set defect detection. Although supervised baselines suffer from performance drops in new/unseen defects, their method maintains recall rates comparable to the state-of-the-art unsupervised (e.g., 85% in unseen anomalies) without sacrificing precision in known classes. They validate that symbolic constraints can successfully adapt to unseen variations. This hybrid approach is relevant in safety-critical applications, as it ensures that data-driven predictions remain physically grounded and logically consistent. Supervised methods are feasible for UXO localisation and binary detection, but they cannot reliably identify specific UXO types because representative, balanced datasets are lacking, representing real-world distribution. Inductive few-shot learning [44, 55] can handle situations with imbalanced datasets, but they often struggle with cross-domain adaptation. Transductive few-shot methods [53], such as maximisation of test information [8], require optimised centroids or adaptive layers, the main dependency being the choice of a representative support subset. However, such systems lack explainability capabilities, making post-hoc analyses based on Local Interpretable Model-Agnostic Explanations (LIME) [41] or gradient-based methods [54] necessary to understand final output/predictions in safety-critical applications such as UXO identification. With respect to safety-critical domains, the system operator needs feedback along with the prediction itself. Otherwise, the decision-support system may lead to the generation of false negative predictions that can have serious repercussions. Knowledge-Enhanced Computer Vision systems integrate external knowledge, often in semantic form, to improve decision-support systems and/or to explainability capabilities. Such dual approaches based on semantic knowledge added on top of multimodal models



(a) Projectile (b) Grenade (c) Mortar Bomb

(d) Cartridge (e) Aviation Bomb (f) RPG

(g) Mines (h) Rocket (i) AntiSubmarine Bomb

Figure 7. **Visual examples of the dataset classes.** The UXOs in the dataset are exactly in the same condition, position, configuration, and appearance as they were found in the field during real interventions.

have been introduced in the past [62], where the authors implemented a multimodal decision system that outperforms three out of five human experts and achieves superior results compared to classical transfer learning methods in the zero-shot paradigm in four different datasets. Researchers [60] introduced the ERNIE-ViL framework, which consists of parsing multimodal output using Scene Graph Knowledge. By establishing connections between objects, attributes, and relationships, ERNIE-ViL [60] achieves a 3.7% improvement on the VCR Leaderboard. To incorporate neuro-symbolic knowledge effectively, it is requisite to define a logical reasoning mechanism that complements the pattern analysis/recognition capabilities of neural networks. According to researchers [35, 45], a foundational approach is *first-order logic (FOL)*, which introduces expressivity through predicates, quantifiers, and variables to enforce "hard constraints" on model output. For example, in UXO detection contexts, FOL can impose axiomatic rules such as $\forall x(Landmine(x) \implies Bakelite(x))$, ensuring that the properties of specific materials inherently imply the presence of a target. However, the primary challenge in integrating FOL with deep learning lies in its discrete, non-differentiable nature. Beyond rule-based constraints, reasoning can be structured through Ontological Logic, often formalised as Description Logic (DL) [22]. This paradigm focusses on the hierarchical organisation of semantic relationships, effectively distinguishing between terminological schemas (TBox) and specific assertions (ABox). When integrated into neuro-symbolic architectures, these ontologies are frequently mapped into vector spaces via Knowledge Graph Embeddings [56], allowing the neural network to learn and respect the underlying topology of the domain while maintaining semantic fidelity. Addressing the inherent noise and uncertainty of real-world signal acquisition requires moving beyond binary truth values to probabilistic frameworks. Probabilistic logic based on Markov Chains, specifically Markov Logic Networks (MLNs) [42], extends first-order logic by attaching weights to formulas. For this model, a violated rule does not render a world impossible, but rather less probable. To resolve this efficiency bottle-

neck, Probabilistic Soft Logic (PSL) [4, 37] has emerged as a compelling alternative that uses continuous truth values within the interval $[0, 1]$. In the current study, PSL relaxes logical operators using the Lukasiewicz t-norms, such that a logical conjunction $A \wedge B$ transforms into the arithmetic operation $\max(0, A + B - 1)$, and negation $\neg A$ becomes $1 - A$. We elaborate on specific aspects related to KG and logical operators in Section 3.

H. Qualitative Results

To gain deeper insight into the behavior of the neuro-symbolic framework, we perform a qualitative study of three representative cases: (1) robust identification, (2) confirmation bias, and (3) error correction. We characterize the system's dynamics by tracking the Knowledge Graph Energy ($\mathcal{J}KG$), which quantifies constraint violations (Equation 1), together with the PSL consistency score ($PPSL$), computed from the energy configuration (Equation 3).

Listing 1. Prompt used for the baseline model

```
You are an expert EOD (Explosive Ordnance
    Disposal) technician.

Analyze the object in this image and classify it
    into one of the following categories.

CLASSES:
{classes_str}

Respond with a JSON object containing the
    predicted class.

Format:
{
  "class": "ClassName"
}

If you are unsure, choose the most likely class
    based on visual features.
```

Each prompt functions as a *soft program* that constrains VLM behaviour while preserving flexibility. Listing 1 is employed for baseline inference (VLM), Listing 2 is utilized for knowledge-augmented inference, and Listing 3 is used for our proposed framework.

Listing 2. Class-specific constraints used in the prompt

```
You are an expert EOD (Explosive Ordnance
    Disposal) technician.

Analyze the object in this image and classify it
    into one of the following categories,
    adhering strictly to the visual constraints
    defined below.

CLASSES AND CONSTRAINTS:
{constraints}

Unknown:
Select this class ONLY if the visible features do
     NOT match the REQUIRED attributes of any
    other class.

Respond only with a JSON object containing the
    predicted class:

{
  "class": "ClassName"
}

If the object fits multiple classes, choose the
    one with the most matching REQUIRED
    attributes.
```

Listing 3. Prompt used in the DEFUSAL Framework

```
Analyze the object in this image.
Score each attribute from 0.0 to 1.0 based ONLY
    on visual evidence.

SCORING GUIDELINES:
- 0.0: not visible
- 0.3--0.5: partially visible / uncertain
- 0.6--0.8: clearly visible
- 0.9--1.0: very certain

{pairs_text}

VISUAL ATTRIBUTE DEFINITIONS (use JSON keys with
    underscores):

{formatted_descriptions}

OUTPUT FORMAT:
- Return a JSON object with ALL keys listed below
    .
- Start from this template and only change values
     you see evidence for:
{template_json}

Missing keys will be treated as 0.0, but do not
    omit keys.

Example:

{
  "tail_fins": 0.7,
  "teardrop_shaped_front": 0.6
}

{guidelines}
```

I. Use of Large Language Models

We used Gemini 3 Pro Academics for sentence-level grammar refinement and code readability edits. All conceptual contributions, proofs, experiments, implementations, and analyses are our own.

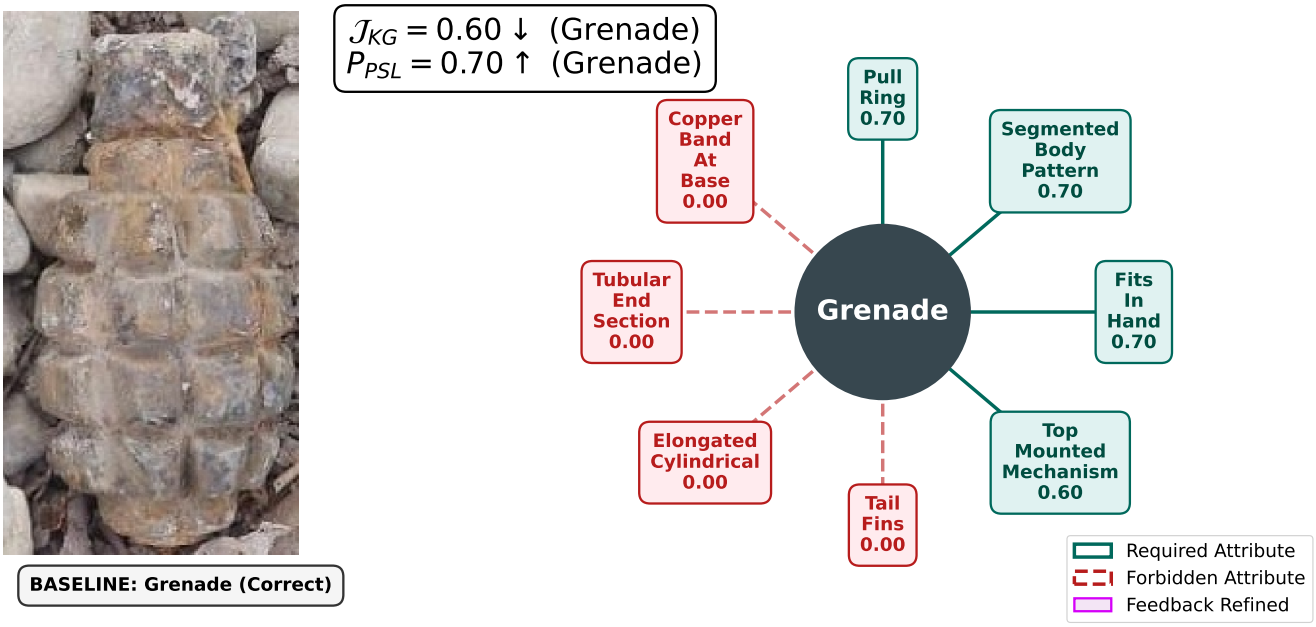


Figure 8. The **Grenade** is correctly identified by both the neuro-symbolic system (y_{PSL} = Grenade) and the baseline model. The system identifies strong supporting evidence and only minor constraint violations for **Grenade** (Energy $\mathcal{J}_{KG} = 0.60 \downarrow$), clearly outperforming the next best hypotheses: *Mine* ($\mathcal{J}_{KG} = 1.50$) and *Aviation Bomb* ($\mathcal{J}_{KG} = 2.10$). The probabilistic logic corroborates this with high confidence: $P_{PSL} = 0.70 \uparrow$, compared to $P(\text{Mine}) \approx 0.30$. The identification is supported by strong evidence for required attributes such as `pull_ring` and `segmented_body_pattern`.

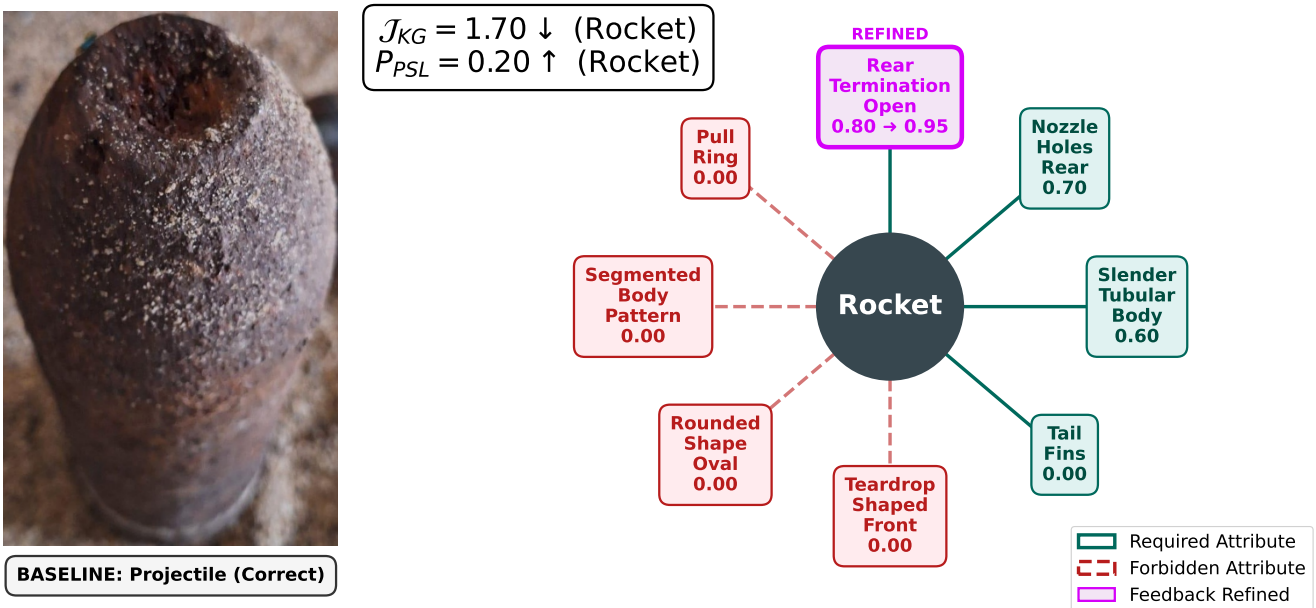


Figure 9. A case where the neuro-symbolic system chose **Rocket**, while the baseline correctly predicts **Projectile**, a common class in the head of long-tailed UXO distributions. Initial graph energies were high for all top candidates: **Rocket** (1.70), *Cartridge* (1.80), and *Projectile* (2.00). The PSL engine ($\mathcal{J}^{(1)}_{PSL} = 0.80$) triggered the Feedback Loop. The mechanism queried the VLM again ("Is the rear open?"), and the VLM reinforced it, with confidence increasing $0.80 \rightarrow 0.95$. This confirmation bias lowered the energy for Rocket ($\mathcal{J}^{(2)}_{PSL} = 0.20$), misleading the system into a confident but wrong final decision. This is a difficult case, as many distinctive projectile features are missing, the image angle is unfavorable, and the munition is oxidized; however, $y =$ *Projectile* is ranked among the top three candidates. In such cases, the tracing (including extracted attributes) can be monitored by the specialist to make informed decisions.

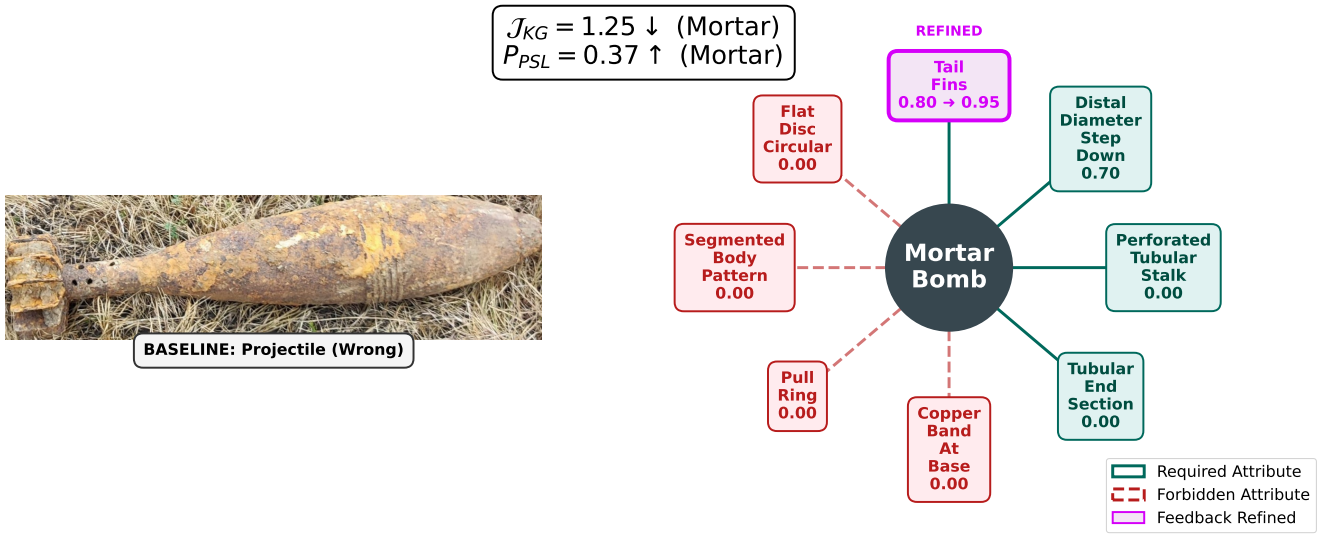


Figure 10. A successful example in which the neuro-symbolic system surpasses the baseline (which predicted *Projectile*). The initial graph-based hypothesis favored **Mortar Bomb** ($\mathcal{J}_{KG} = 1.40$) over *Mine* (1.55) and *Grenade* (2.15). The initial PSL assessment was uncertain ($\mathcal{J}^{(1)}_{PSL} = 0.50$, Prob ≈ 0.108) because of conflicting information. The feedback loop was activated on `tail_fins`. The VLM's follow-up evaluation confirmed the presence of tail fins with higher confidence ($0.80 \rightarrow 0.95$). This increased the probability to $P_{PSL} = 0.37$, effectively resolving the earlier ambiguity.

J. Computational Analysis

The MAP objective is addressed using a constrained SLSQP (Sequential Least Squares Programming) optimiser on the probability simplex, enforcing bounds $P(y) \in [0, 1]$ and the equality constraint $\sum_y P(y) = 1$. We initialise $P(y)$ from the knowledge-graph scores by applying a softmax to the negative energies, and we handle disjunctive requirements/exclusions via the max t-conorm (for example, $I(a_1 \vee a_2) = \max(I(a_1), I(a_2))$). If the optimiser either fails or yields non-finite outputs, we revert to the initial posterior. The final prediction is assigned to *UNCERTAIN* whenever the PSL energy of the top-ranked class exceeds the number of required attributes. Optimisation methods for neuro-symbolic frameworks were validated in other closed-loop neural symbolic approaches [27]. The knowledge graph comprises $|V| = 47$ vertices and $|E| = 137$ edges that encode the domain constraints. For each image, PSL inference instantiates $O(|Y| \times |A|) = 9 \times 38 = 342$ candidate facts. In the convex formulation of HL-MRF, the inference of MAP has a complexity $O(n^2)$ in the number n of grounded atoms. The PSL optimization step itself requires little computational effort, taking just 3.58 ms per image, while the feedback loop adds , on average, a further 367 ms when this cost is amortised across all samples (including those that never invoke feedback). The capability to deploy the framework on real-time systems also depends on the selected VLM backbone's size and the extent of the optimizations applied to it.

K. Future Research Directions

Future work will extend the current framework beyond class-level labels toward dense attribute-level annotations (e.g., specific fuze types, corrosion levels, and paint markings) in order to more accurately assess the risk associated with both the ordnance class and its observed attributes. We also aim to close the learning loop by using the successful reasoning dialogues generated by the system as training data. By fine-tuning a smaller, specialised VLM on these logically verified reasoning traces, we intend to distil the knowledge of the neuro-symbolic framework into a faster end-to-end model suitable for edge deployment. This distilled model will additionally serve as a baseline for analysing the trade-off between primary performance metrics and secondary constraints, particularly inference time and memory usage. Another research direction concerns deeper identification capabilities. Beyond recognising the UXO category, we aim to infer characteristics related to the calibre and origin of the ammunition in order to estimate explosion parameters and appropriate safety measures. While prior studies primarily focused on UXO/NON-UXO localisation and classification, the present work advances the task by identifying the UXO type (e.g., mortar bomb) together with a reasoning trace. Future work will extend this capability to subtype identification and risk estimation (e.g., a mortar bomb of calibre 82 mm associated with elevated risk due to the presence of an oxidised fuze).