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.

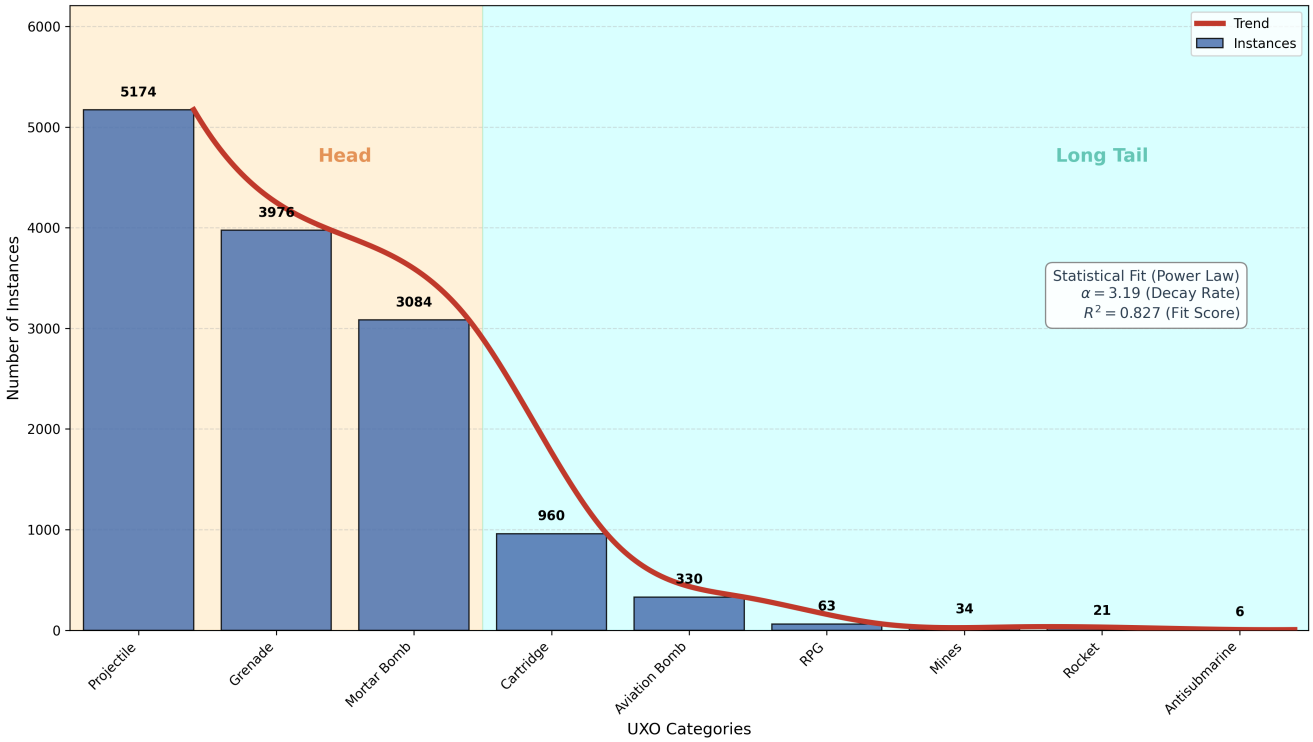
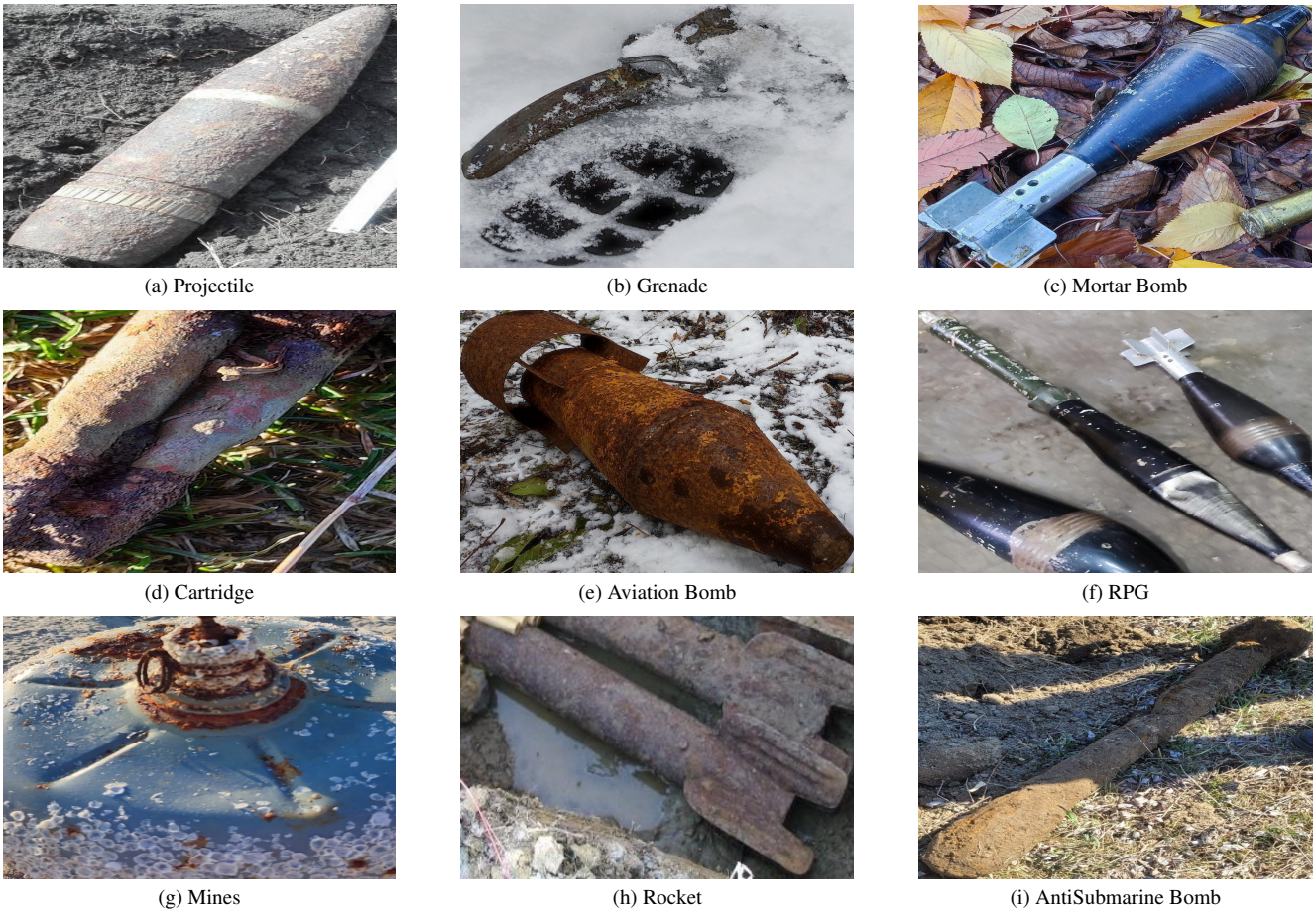


Figure 6. Unexploded Ordnances Distribution. While this distribution characterises real-world historical contamination, being the distribution from real pyrotechnic intervention, we acknowledge that in recent asymmetric conflicts, the *Landmine* class has seen a proliferation in usage and priority for 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.

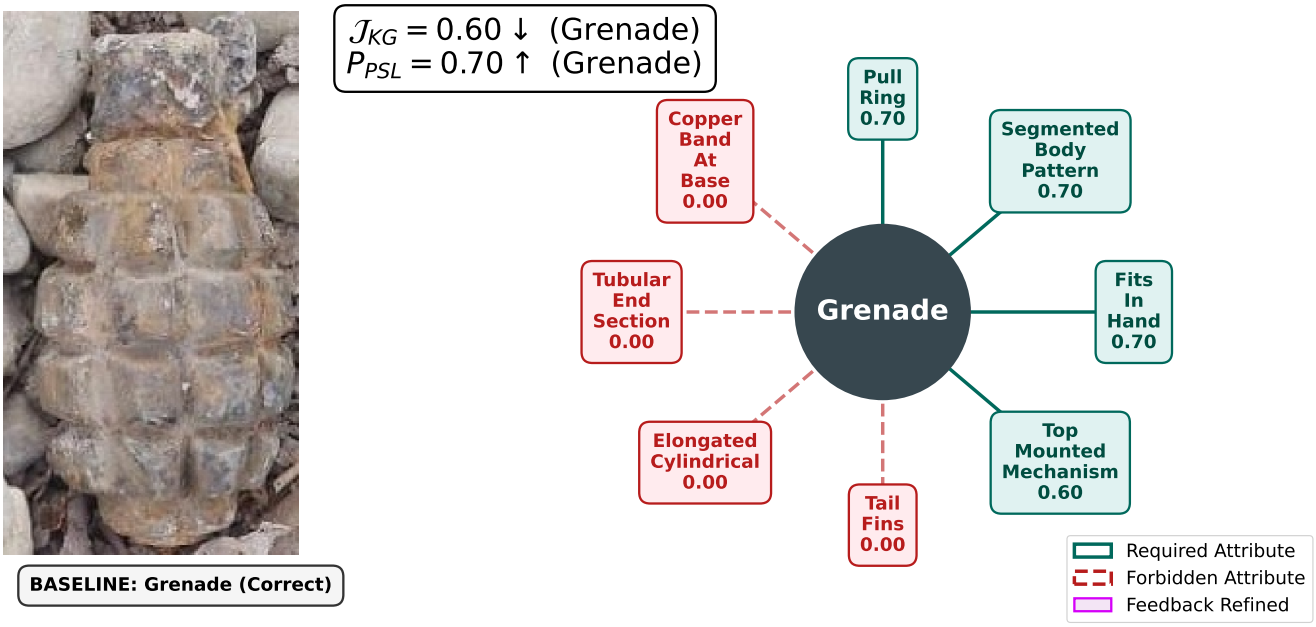


Figure 8. The **Grenade** is correctly identified by both the neuro-symbolic system (y_{PSL} = Grenade) and the baseline model. The system identifies strong supporting evidence and only minor constraint violations for **Grenade** (Energy $\mathcal{J}_{KG} = 0.60 \downarrow$), clearly outperforming the 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`.

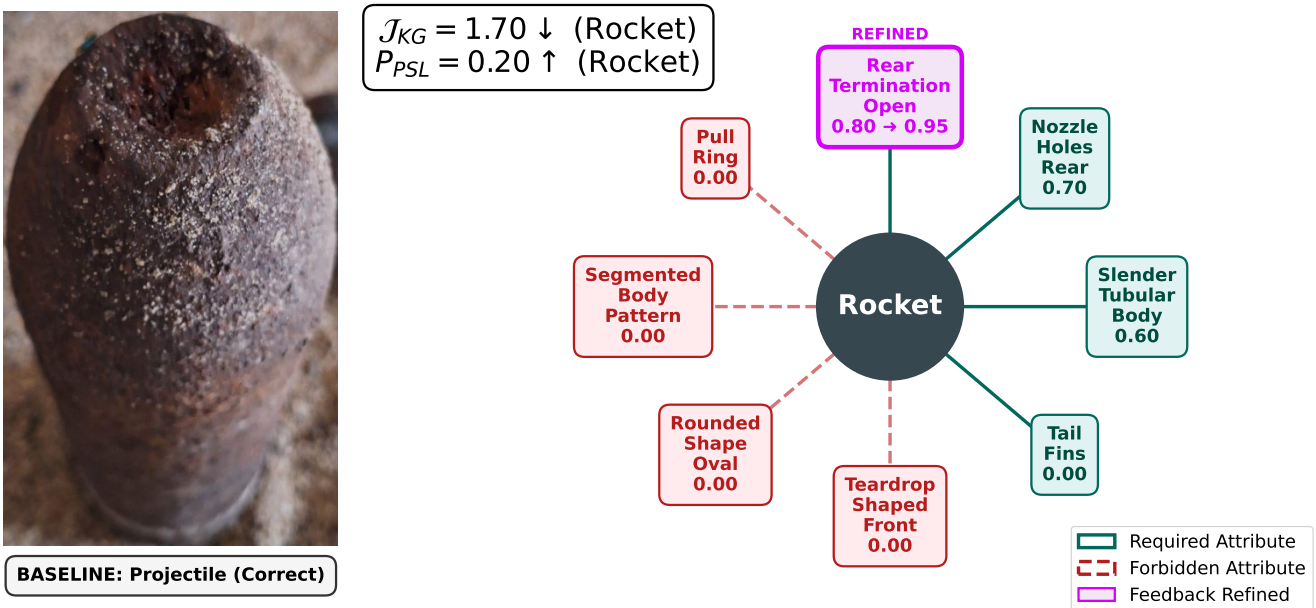


Figure 9. A case where the neuro-symbolic system chose **Rocket**, while the baseline correctly predicts **Projectile**, a common class in the head of long-tailed UXO distributions. Initial graph energies were high for all top candidates: **Rocket** (1.70), *Cartridge* (1.80), and *Projectile* (2.00). 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.

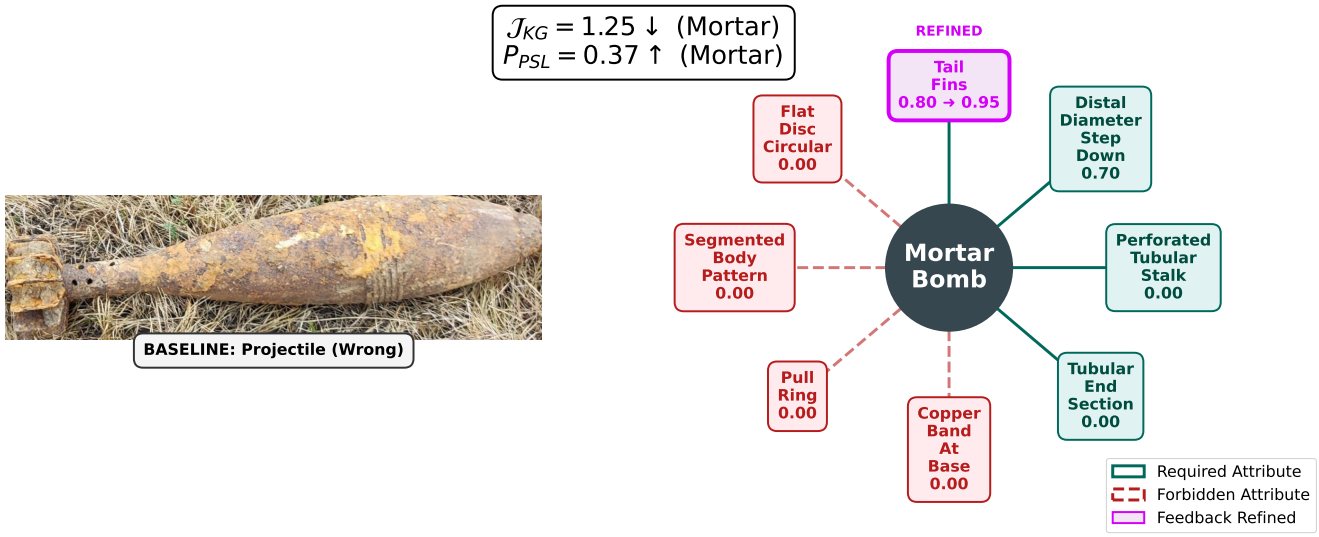


Figure 10. A successful example in which the neuro-symbolic system surpasses the baseline (which predicted *Projectile*). The initial graph-based hypothesis favored **Mortar Bomb** ($\mathcal{J}_{KG} = 1.40$) over *Mine* (1.55) and *Grenade* (2.15). The initial PSL assessment was uncertain ($\mathcal{J}^{(1)}_{PSL} = 0.50$, 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).