Bridging Explanations and Logics: Opportunities for Multimodal Language Models

Nicolas Sebastian Schuler[1][0009−0009−6688−9416], Vincenzo Scotti[1][0000−0002−8765−604X], Matteo Camilli[2][0000−0003−2491−5267], and Raffaela Mirandola[1][0000−0003−3154−2438]

[1] KASTEL,
Karlsruhe Institute of Technology, Am Fasanengarten 5, 76131, Karlsruhe, Germany
`nicolas.schuler@kit.edu`,
`vincenzo.scotti@kit.edu`, `raffaela.mirandola@kit.edu`
[2] DEIB, Politecnico di Milano, Via Golgi 42, 20133, Milano (MI), Italy
`matteo.camilli@polimi.it`

Abstract. As subsymbolic Artificial Intelligence (AI) systems have become increasingly integrated into decision support tools, there is a consequent need for transparency and interpretability. While eXplainable AI (XAI) techniques offer valuable insights into model behavior, they often lack the formal rigor required for causal interpretation and verification – qualities inherent to symbolic AI. This paper presents a framework designed to bridge the gap between subsymbolic explanations and symbolic reasoning through the application of Multimodal Language Models (MLMs). Our approach combines the output of XAI methods with symbolic knowledge bases encoded in a logic programming language, enabling abductive reasoning and yielding causal interpretations of explanations produced over predictions. In our framework, MLMs serve as intersymbolic translators, converting visual or textual explanations into structured logical assertions that can be processed by reasoning engines for verification. Through this integration, we aim to enhance the interpretability of AI systems and promote the use of sound reasoning to increase the trustworthiness of the observed AI-based system. We outline our proposed methodology, conduct a preliminary experiment, and discuss both future directions and key challenges. Our work contributes to the emerging field of intersymbolic AI, which calls for the integration of symbolic and subsymbolic paradigms in the pursuit of trustworthy AI.

Keywords: Intersymbolic AI · XAI · MLM · Abduction.

1 Introduction

In the last few years, the proliferation of data and advances in computational resources have led to an impressive acceleration in the development and adoption of Artificial Intelligence (AI). At the core of this progress lie *subsymbolic* (i.e., data-driven) techniques, such as deep learning, which have achieved remarkable performance across domains ranging from computer vision and natural language processing to scientific discovery [9]. While these advances have expanded the scope and capability of intelligent systems, they have also brought to light concerns regarding the *opacity* of their internal operations and the difficulty of explaining their decisions to human stakeholders. This growing tension has given rise to the research field centered on *explainability* and eXplainable Artificial Intelligence (XAI) [24].

XAI seeks to provide human-comprehensible justifications for model predictions, addressing one of the central challenges of contemporary AI. Methods such as LIME [20], SHAP [11], and GradCAM [22] have become standard tools for attributing a model's output to specific input features, offering insights into which aspects of the input most strongly influenced a given prediction and, in many cases, the resulting decision-making process.

However, subsymbolic XAI methods, rooted in statistical association, often lack the representational structure necessary to express causal relationships explicitly. They can highlight influential factors but struggle to justify whether these associations correspond to true causal links or spurious correlations. Moreover, validating such explanations formally, for instance by testing their logical consistency or confirming their invariance under intervention, remains a major challenge.

Addressing these limitations calls for more rigorous, logic-based approaches, where symbolic AI frameworks provide principled mechanisms to encode, interpret, and verify causal claims within a formally grounded setting. Among classical forms of inference, namely (i) *deduction*, (ii) *induction*, and (iii) *abduction*, *abduction* holds particular promise for XAI applications. Abductive reasoning seeks the most plausible causes for observed effects, thereby offering a mechanism to bridge the gap between a model's output with structured, causally grounded explanations [7,14,6].

Yet, abductive reasoning presents two critical obstacles. First, it demands a priori explicit encoding of domain knowledge within a formal logic system. Second, this structured knowledge is typically absent from the raw outputs that subsymbolic models produce. Consequently, practitioners face fundamental challenges: how to translate XAI outputs into logic-compatible representations while simultaneously incorporating relevant world knowledge to support abductive inference, all while accommodating the inherent uncertainty characteristic of data-driven models.

We propose leveraging Multimodal Language Models (MLMs) as a potential avenue for addressing these challenges. MLMs can combine several input modalities, have a vast – albeit potentially incomplete or erroneous – knowledge base across diverse domains, and they can generate structured outputs. Therefore, MLMs can in principle function as intersymbolic translators – transforming subsymbolic explanations into formal symbolic representations compatible with downstream reasoning engines. By exploratively examining how MLMs bridge the representational gap between explanation techniques and logical reasoning systems, we demonstrate the feasibility of a pathway toward more rigorous causal analysis while maintaining practical applicability to real-world model outputs. This work serves as an initial foundation and conceptual basis for future studies in that area.

The remainder of the paper is structured as follows. Section 2 presents background and related work. Section 4 presents our proposed methodology. Section 5 provides a demonstrative analysis. Section 6 discusses ongoing and future research directions. Finally, Section 7 concludes the paper.

2 Background

To contextualize our proposed framework, we review the fundamental concepts of post-hoc explainability, abductive inference, and probabilistic logical programming.

2.1 Post-Hoc Explainability

In this section, we provide an overview of post-hoc methods for XAI and we introduce the concept of *semantic gap* in the context of visual explanations.

Overview Subsymbolic AI systems, particularly deep learning models like Convolutional Neural Networks (CNNs), are inherently opaque, creating a critical need for transparency and interpretability. XAI encompasses a diverse set of methods designed to enhance the transparency and interpretability of machine learning-based systems by providing human-understandable rationales for model behavior. [24]. In this context we distiguish between interpretability and explainability. *Interpretability* refers to the degree to which a human can comprehend a model's internal mechanics or decisions through direct inspection—essentially treating it as a passive property of the model. *Explainability*, by contrast, is an active process: it focuses on the mechanisms through which a machine learning system conveys the reasons for its outputs, enabling users to form or update beliefs based on these explanations [**?**]. This distinction is central to clarifying whether a system is inherently interpretable or requires post-hoc explanation, shaping the goals and evaluation criteria of XAI research.

Current XAI research distinguishes between two primary methods: (i) *attributive methods*, which quantify the influence of input features on model outputs, typically producing importance scores or relevance maps, and (ii) *counterfactual methods*, which explore how inputs must change to alter a prediction, supporting causal reasoning about alternative outcomes [2]. Emerging research directions extend these approaches to encompass *causal and mechanistic interpretations*, which aim to extract the internal computational pathways and dependencies underlying a model's predictions [15].

These methods coexist with other taxonomies, such as model-agnostic versus model-specific techniques. *Model-agnostic methods*, like LIME [**?**] and SHAP [**?**], treat the target model as a black box. *Model-specific methods*, such as Grad-CAM [**?**], Integrated Gradients [**?**], and DeepLift [**?**], exploit architectural properties of neural networks to deliver more faithful attributions.

Semantic Gap For the scope of this work, we focus on attributive methods for explainability and we extend them with causal inference to ground the prediction of a model in an interpretable and sound process. To this end, we need to encode the information coming from subsymbolic modlels in a symbolic format In fact, despite the progress in post-hoc explainability, a persistent challenge is the *semantic gap* between low-level attribution outputs and high-level human concepts.

If we consider the visual domain, for example, gradient-based attribution methods, such as Grad-CAM and its generalization, Grad-CAM++ [3] are standard for visualizing the spatial regions of an input image that maximize the class activation. However, while these methods offer valuable insights into where a model focuses, they often lack the formal rigor required for verification.

We refer to the absence of this explicit mapping as semantic gap: a heat map provides attributive localization (e.g., highlighting an area) but fails to explain what concepts were detected or why they matter [4]. This lack of formal verification necessitates a transition from purely visual attributions to structured, symbolic explanations.

2.2 Symbolic Reasoning and Logical Inference

In this section, we provide an overview of logical inference. We then delve into the details of abductive reasoning (the inference method we exploit in our explanation pipeline), and finally introduce the concept of (probabilistic) logical programming.

Overview Symbolic reasoning refers to the use of formal representations—such as logic, rules, and structured knowledge—to perform inference and decision-making. [**?**] Unlike sub-symbolic approaches (e.g., neural networks), symbolic systems operate on discrete symbols and well-defined semantics, enabling explicit reasoning about concepts and relationships. This paradigm is often implemented in many classical AI techniques, including*knowledge bases*, *ontologies*, and *expert systems*, and remains essential for tasks requiring transparency, verifiability, and alignment with human-understandable logic. In modern AI, symbolic reasoning is increasingly integrated with statistical and neural methods to bridge the gap between structured knowledge and data-driven learning, forming the foundation for *inter-symbolic* AI [**?**] and *neuro-symbolic* AI [**?**].

In this context, *logical inference*, the process of deriving conclusions from premises according to formal rules, is central to symbolic reasoning. We distinguish among three types of inference (or reasoning): deduction, induction, and abduction. *Deductive inference* derives logically certain conclusions from general premises (e.g., from “All humans are mortal” and “Socrates is human,” deducing “Socrates is mortal”). *Inductive inference* generalizes from specific observations to broader rules, often introducing uncertainty (e.g., observing many white swans and inferring “All swans are white”). *Abductive inference* seeks the most plausible explanation for observed facts, often used in diagnostic or hypothesis-generation contexts (e.g., inferring a disease as the likely cause of symptoms). These inference modes collectively enable reasoning under certainty, uncertainty, and incomplete information, making them indispensable for explainable and trustworthy AI.

Abductive Reasoning Abductive reasoning, formally defined as *inference to the best (i.e., most plausible) explanation* [7]. In our work, we propose to use abductive reasoning to help bridge the semantic gap.

Unlike deduction (which derives conclusions from rules) or induction (which generalizes rules from observations), abduction operates backwards: given a knowledge base KB and a specific observation O (in our case, the model prediction), it seeks a hypothesis H (the explanation) that explains O [8]. Formally, an abductive problem can be characterized as finding a set of features Δ such that:

$$KB \cup \Delta \models O \quad \text{and} \quad KB \cup \Delta \text{ is consistent} \tag{1}$$

This formalism provides the missing verification layer for XAI: it mathematically ensures that the highlighted features (Δ) are sufficient to logically entail the observation (O) given the domain rules encoded in (KB) [6,14].

Logic Programming and Probabilistic Logic Programming ...

However, the strict logical notion of abduction is usually not practical in the presence of uncertainty [18]. Therefore, we utilize the probabilistic version of this concept with Probabilistic Programming – called Probabilistic Logic Programming – specifically ProbLog[19]. ProbLog extends standard Prolog by introducing probabilistic facts, denoted as $p :: f$, stating that an atom f is true with a probability p. This formalism allows for *soft* confidence scores, creating a natural harmonization between *hard* logical constraints and *soft* statistical confidence.

2.3 Intersymbolic AI

Intersymbolic AI represents the integration of symbolic and subsymbolic paradigms in the pursuit of trustworthy AI. It closely aligns with the field of neuro-symbolic AI, seeking to combine the robust learning capabilities of statistical models (subsymbolic) with the transparency and formal rigor of logical systems (symbolic) [13,23]. While subsymbolic models excel at processing high-dimensional, noisy data, they often lack the ability to provide causal justifications. Intersymbolic AI aims to bridge this divide by grounding these statistical predictions in verifiable logic.

3 Related Work

Effective explanations for AI systems must bridge the gap between statistics and logical justification. This section reviews recent advancements in intersymbolic AI, specifically focusing on the challenges of symbolic grounding, architectural approaches to rule extraction, and the emerging role of Large Language Models (LLMs) in reasoning.

A fundamental challenge in trustworthy AI is the *grounding problem* and ensuring that the internal representation of a model truly corresponds to real-world concepts – often expressed as symbols. As suggested by Marconato et al. [12], modern deep Machine Learning (ML) models are prone to learning *reasoning* *shortcuts*, where predictions are based on spurious correlations in the data rather than causal mechanisms. For example, a model might classify an object based on a specific background texture, rather than its shape. While standard XAI methods can visually hint at these shortcuts, they cannot express whether the model's internal *reasoning* is logically sound. In this work, we utilize abductive reasoning to explicitly check grounding results, to distinguish between mere correlation and valid logical inference.

To achieve logically grounded interpretability, one prominent research stream involves modifying the neural architecture itself to support symbolic rule extraction. For instance, Padalkar and Gupta [16] recently introduced a framework for interpretable image classification using Vision Transformer. By introducing a sparse concept layer and subsequent application of specific algorithms, they map attention-guided representations directly to Answer Set Programming rules. While powerful, such approaches are inherently model-specific and require architectural constrains – necessitating careful design of concept layers and retraining. In contrast, our framework is model-agnostic; it operates solely on the outputs and the explanation that is generated, effectively treating the underlying model as a black box.

Concurrent with architectural approaches, the concept of LLMs as a judge has been explored to facilitate symbolic reasoning. He et al. [5] introduced *ProtoReasoning*, a framework that translates reasoning tasks into Prolog prototypes, demonstrating that mapping unstructured input to a logical program can significantly enhance generalization. Similarly, Allen et al. [1] propose integrating LLMs directly into the interpretation function of formal semantics for paraconsistent logic, leveraging the LLM's vast knowledge base while maintaining logical soundness and completeness. While these works validate the use of LLMs as *translators* to formal logic, they primarily operate within the textual domain. Our work directly addresses the challenge of visual grounding – specifically, how to translate the abstract visual explanation provided by common XAI methods into logical assertions without human intervention, while preserving the connection to pixel-level evidence. These grounding results are then handed off to a probabilistic reasoning engine for abductive inference.

Current research literature presents a dichotomy, where architectural methods offer visual rigor but lack generality, while translation methods offer flexibility but have been limited to textual domains. This paper addresses this gap by employing MLMs as intersymbolic translators, mapping visual explanations to probabilistic Prolog facts, thereby bridging visual perception, logical grounding, and formal reasoning.

4 Methodology

Our framework establishes an intersymbolic pipeline to ground a model's prediction on its own visual explanation. The architectural concept is illustrated in Figure 1, and consists of three phases: (i) *subsymbolic perception*, (ii) *intersymbolic translation*, and (iii) *abductive reasoning.* Before running data through our pipeline, we consider the preceding phase: *ontology and knowledge base creation.*

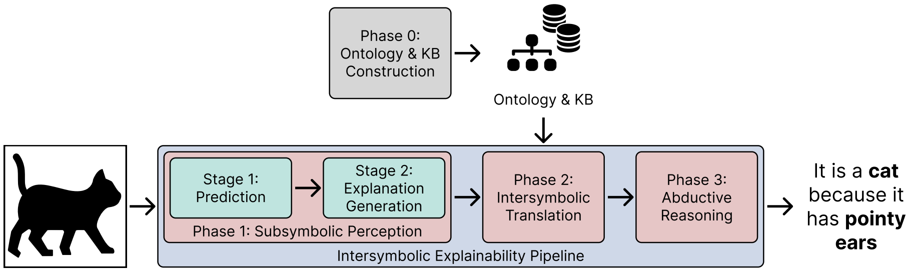
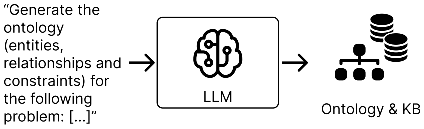


Fig. 1: Probabilistic Abduction Framework for Intersymbolic Explainability.


"Generate the ontology (entities, relationships and constraints) for the following problem: [...]"
LLM
Ontology & KB

(a) Ontology & *KB* Construction

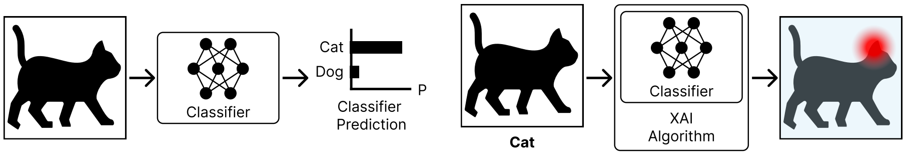
```
0.9::has(cat, pointy_ears).
0.8::has(cat, whiskers).
0.2::has(cat, round_ears).

0.8::has(dog, round_ears).
0.2::has(dog, pointy_ears).
0.3::has(dog, whiskers).

explains(Feature, Animal) :- (
    highlighted(Feature),
    has(Animal, Feature)
).
```

(b) Example Encoding

Fig. 2: Phase 0: Ontology & *KB* Construction

To bridge the gap between visual pixels and logical semantics, we introduce an ontology $\mathcal{O}$ to encode our knowledge into a symbolic representation. We use a *knowledge base KB* grounded in the ontology we defined, to store the domain knowledge describing the causal relationships between features and classes. $\mathcal{O}$ and KB can come from several sources: crafted by hand, fetched from an external resource, or automatically constructed. In our approach, we propose to use LLMs to build it automatically (or semi-automatically) starting from the problem description (see Figure 2). The idea, represented in Figure 2a, is to elicit the LLM to output its *internal world knowledge* concerning the problem in a structured format as if it were a *domain expert* [17]. The output of phase 0 resembles the example in Figure 2b.

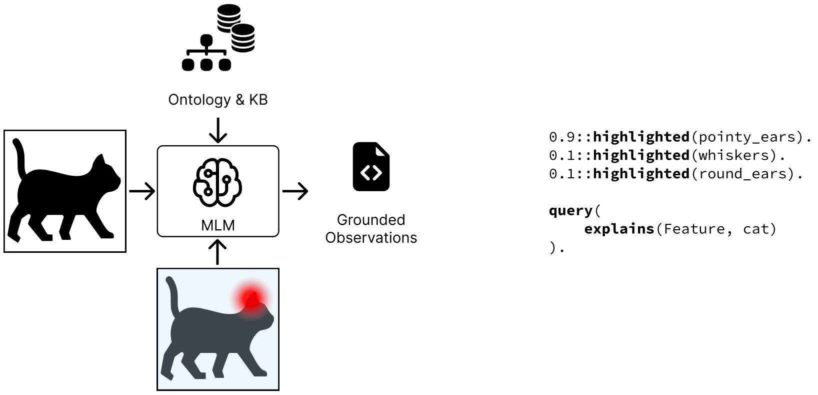
Phase 1 of the pipeline is divided into two stages (see Figure 3): *Prediction* and *explanation generation*. The pipeline begins with a standard deep learning classification model M, mapping an input image I to a probability distribution over classes C (see Figure 3a). Upon generating a prediction $\hat{y} \in C$, where C denotes the classes, we employ a post-hoc XAI technique, such as *Grad-CAM++* [3] to generate a heat map H (see Figure 3b). This map H identifies the spatial regions of I that maximized the activation for class $\hat{y}$, indicating the spatial locus of attention without conveying the underlying semantic content; this limitation represents the challenge of semantic grounding. In this case, H could prominently highlight the area of the image that includes the animal's ears.



(a) Stage 1: prediction

(b) Stage 2: explanation generation

Fig. 3: Phase 1: subsymbolic perception



```
0.9::highlighted(pointy_ears).
0.1::highlighted(whiskers).
0.1::highlighted(round_ears).

query(
    explains(Feature, cat)
).
```

(a) Translation

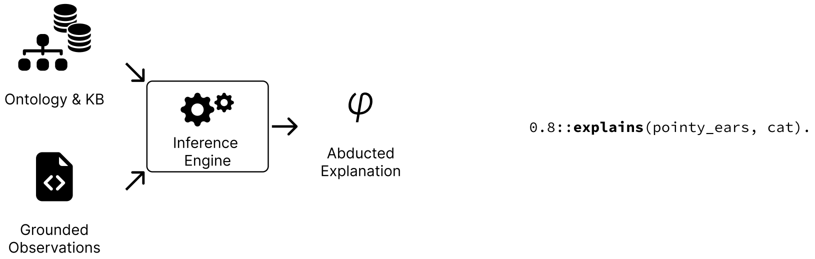
(b) Example Grounding

Fig. 4: Phase 2: Intersymbolic Translation

Phase 2 is where MLMs can serve as a bridge between low-level model outputs and high-level symbolic reasoning, as depicted in Figure 4. MLMs are capable of interpreting visual inputs (e.g., heat maps), extract relevant features (e.g., “ears are highlighted”), and generate structured logical assertions (e.g., `0.9::has(Cat, pointy_ears)`) that are compatible with the symbolic knowledge base KB. We refer to the result of this translation process as *grounding* [10]. MLMs can serve as mediators between the subsymbolic and symbolic layers by performing the following tasks: (i) mapping visual features to semantic concepts using contextual knowledge, (ii) generating logic-compatible hypotheses for abductive reasoning, or (iii) handling uncertainty by interfacing with probabilistic logic systems like ProbLog.

In this instance, the MLM (f_{MLM}) receives the ontology $\mathcal{O}$, and the heat map H as input, and is prompted to ground the explanation in a structured format (see Figure 4a). The MLM objective is to ground the highlighted visual features into symbols, compatible with $\mathcal{O}$. The output is a set of logical facts F associated with the highlighted features identified by the MLM. These facts should be associated with a probability to handle uncertainty and allow probabilistic reasoning. However, extracting the correct probability estimates is beyond the scope of this work; thus, we resort to the LLM self-reported probabilities.

Following the example of Figure 4b, if H highlights the animal's ears and part of the muzzle, the MLM generates the ProbLog fact `0.9::has(Cat, pointy_ears)` and `0.5::has(Cat, whiskers)`. This step is critical, since it converts the visual explanation into discrete symbolic concepts $F = f_{\mathrm{MLM}}(\mathcal{O}, KB, H, I)$.



```
0.8::explains(pointy_ears, cat).
```

(a) Abduction (b) Example Explanation

Fig. 5: Phase 3: Abductive Reasoning

In the final phase of our pipeline, we employ probabilistic abductive reasoning to determine if the extracted observations F constitute a sufficient explanation for the predicted label $\hat{y}$. Unlike classical binary entailment, we operate within a probabilistic logic framework where grounded features possess confidence scores. Formally, we define the verification condition via probabilistic entailment, requiring that the posterior probability of the label – conditioned on the knowledge base KB and the feature set F – exceed a validity threshold τ:

$$P(\hat{y} \mid KB, F) \geq \tau \tag{2}$$

If the probabilistic entailment holds, the prediction is considered symbolically grounded with probability p. Interestingly, by inverting this logic, we can propose candidates for reasoning shortcuts. Specifically, if the classifier predicts $\hat{y}$ with high confidence, yet the symbolic verification yields a low probability $P(\hat{y} \mid KB, F) < \tau$ – given a sufficiently expressive KB – it implies that the classifier's reasoning was insufficient to logically justify the label, likely relying on spurious correlations.

5 Analysis

To demonstrate the feasibility of our pipeline, we conducted a preliminary evaluation on a custom image dataset. The goal of this analysis is to assess whether the intersymbolic translation and abductive verification can successfully validate correct predictions and detect potential grounding errors in Out-of-Distribution (OOD) samples. We structured our experiment into a multi-stage execution pipeline, corresponding to the workflow illustrated in Figures 6-8.

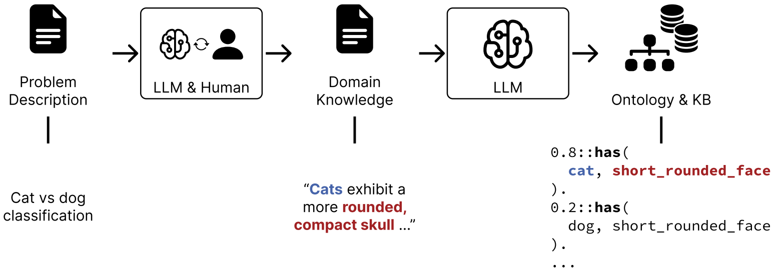
5.1 Experimental Setup

We curated a small, controlled dataset containing five distinct classes of animals: Cat ($n=15$) and Dog ($n=15$) as in-distribution classes, and Fox ($n=12$), Tiger ($n=12$), and Wolf ($n=12$) as OOD classes. These OOD classes were selected to test the system's robustness when detecting features on visually similar but semantically distinct animals. We employed EfficientNet-B0 [26] as the classifier, a CNN pre-trained on ImageNet [21]. To adapt it for binary Cat-vs-Dog classification without retraining, we implemented a weight aggregation protocol: we synthesized a two-class output head by averaging the weight vectors and biases of ImageNet's different Cats and Dogs (breeds) classes. For the Cat class, the weights with indices $[281;285] \in N_{cat}$, and for the Dog class $[151;268] \in N_{Dog}$ have been aggregated. Therefore, the following new weights w_c and biases b_c have been synthesized for $n \in \{Cat, Dog\}$:

$$w_c = \frac{1}{|N_n|} \sum_{i \in N_n} w_i \qquad b_c = \frac{1}{|N_n|} \sum_{i \in N_n} b_i \tag{3}$$

We denote the prediction of this CNN as $\hat{y}_{CNN}$ in the remainder of this paper.

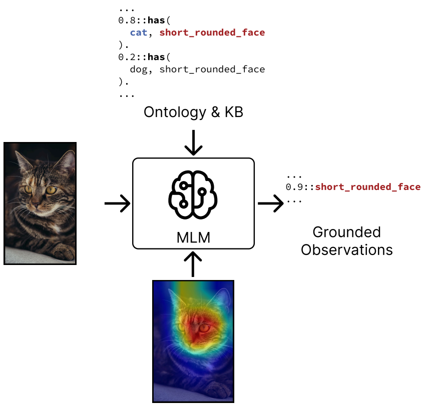
As post-hoc XAI method, we employed Grad-CAM++ [3] to generate heat maps for every prediction. Finally, we utilized the MLM Qwen3-vl:235b-a22b-instruct to interpret the heat maps and ground the features [27].


Problem Description
LLM & Human
Domain Knowledge
LLM
Ontology & KB
Cat vs dog classification
"Cats exhibit a more rounded, compact skull ..."

```
0.8::has(
  cat, short_rounded_face
).
0.2::has(
  dog, short_rounded_face
).
...
```

Fig. 6: Implementation of Phase 0: Human-in-the-Loop Knowledge Elicitation and Ontology Synthesis in Logical Programming Format.

Stage 1 & 2: Knowledge Base Construction Before processing individual images, our pipeline needs to construct the symbolic knowledge base KB that encodes the domain logic for the classification task (see Figure 6). Unlike black-box approaches, our KB requires explicit formulation. We employed a Human-in-the-Loop generation strategy, leveraging two LLMs to assist in creating task-specific natural language reasoning descriptions (Qwen:Qwen3-30b-a3b-2507), where we first prompted the LLM for a comparative analysis between Cats and Dogs, their inherent traits and biological characteristics visible in images[27]. For instance, the LLM responded that Cats have a *more rounded and compact skull compared to Dogs.* This natural language description was then used as an input argument for another, more capable LLM (Gemini 2.5 Pro), to synthesize a logical program that aligns with the intent of the classification task, encoding the features mentioned in the natural language description as facts that can later be added as evidence before the abductive inference stage in the pipeline. The prior probabilities associated with each feature are also derived by the LLM. As previously noted, utilizing LLMs and MLMs offers the advantage of leveraging vast world knowledge, which approximates reasonable prior probabilities for each feature based on learned data distributions. To ensure logical consistency, the authors manually reviewed the generated rules and probabilities.



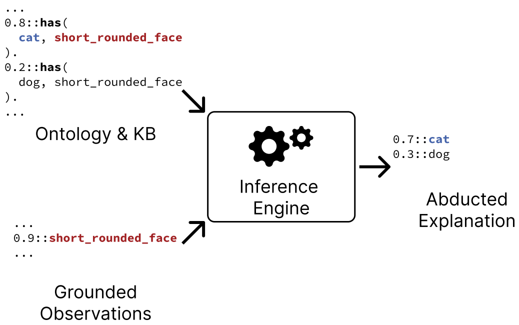
```
...
0.8::has(
  cat, short_rounded_face
).
0.2::has(
  dog, short_rounded_face
).
...
```

```
...
0.9::short_rounded_face
...
```

Fig. 7: Stage 3: Preparation and Extraction of Input Data for the Intersymbolic Translation.

Stage 3: Feature extraction In this stage (see Figure 7), we extract the features encoded in the logical program from Stage 2. To streamline this process, we utilized additional LLM (Qwen:Qwen3-coder-30b)[27]. The extracted features (e.g., `short_rounded_face`), along with the image overlaid with the heat map, are then passed to Stage 4.

Stage 4 & 5: Intersymbolic Translation and Abductive Inference Given the input from the previous stage, the MLM analyzes the image overlaid with the heat map and correlates the extracted features with the highlighted regions (see Figure 8). Further, the MLM assigns a probability score on how certain it is that it sees the given feature highlighted. The resulting grounded features are asserted as evidence



```
...
0.8::has(
  cat, short_rounded_face
).
0.2::has(
  dog, short_rounded_face
).
...
```

```
...
0.9::short_rounded_face
...
```

```
0.7::cat
0.3::dog
```

Fig. 8: Stage 4 and 5: Intersymbolic Translation Results Provide Evidence for Knowledge Base *KB* and Abductive Reasoning Engine.

into the logical program created in Stage 2. The probabilistic reasoning engine – in this case ProbLog – then performs abductive inference to compute the final probability score $\hat{y}_{Sym}$. In this instance, based on the features grounded by the MLM and the logical program, the most probable explanation was correctly identified as a cat.

5.2 Results

Ground truth	**Support**	$\hat{\mathbf{y}}_{\text{CNN}}$		$\hat{\mathbf{y}}_{\text{Sym}}$		$\Delta = \hat{\mathbf{y}}_{\text{Sym}} - \hat{\mathbf{y}}_{\text{CNN}}$	
		$P(\texttt{Cat} \mid \mathbf{X})$	$P(\texttt{Dog} \mid \mathbf{X})$	$P(\texttt{Cat} \mid \mathbf{X})$	$P(\texttt{Dog} \mid \mathbf{X})$	$\Delta P(\texttt{Cat} \mid \mathbf{X})$	$\Delta P(\texttt{Dog} \mid \mathbf{X})$
`Cat`	15	0.994	0.006	0.874	0.126	−0.120	0.120
`Dog`	15	0.357	0.643	0.403	0.597	0.046	−0.046
`Fox`	12	0.717	0.283	0.795	0.205	0.078	−0.078
`Tiger`	12	0.893	0.107	0.665	0.335	−0.228	0.228
`Wolf`	12	0.461	0.539	0.063	0.937	−0.398	0.398

Table 1: Average classification results on the custom dataset with **X** denoting the ground truth.

We compared the raw probability scores of our CNN ($\hat{y}_{CNN}$) against the abductively inferred probabilities ($\hat{y}_{Sym}$) derived from our pipeline. The results are summarized in Table 1.

The last column highlights the impact of the symbolic grounding: for genuine Cats (0.874), the symbolic confidence remains high, though slightly lower than the CNN's raw confidence (0.994). This expected drop occurs because the symbolic engine requires explicit feature evidence, whereas the CNN may rely on non-causal background textures. The most significant result is observed in the Wolf class. The CNN struggled to distinguish Wolves from Cats, assigning a probability of 0.461 to

Cat. However, the abductive verification phase dropped this probability notably to 0.063. The underlying reasoning is now transparent: the MLM detected features, that did not align with Cat features present in the knowledge base. Consequently, the abductive solver could not logically entail the label Cat, therefore, correctly identifying that the visual evidence did not support the CNN's confusion.

In Figure 9, we visualize the feature set and overlaps produced by the grounding stage – given a probability threshold of 0.5. The MLM successfully mapped visual regions to shared concepts, while correctly identifying unique discriminators, allowing for the distinguishing of classes based on the ontology coded in the knowledge base KB. For example, the Cat and Tiger classes, across the whole dataset, share a combined total of 17 features detected by the MLM after the Intersymbolic Translation stage, with a probability threshold greater than 0.5.

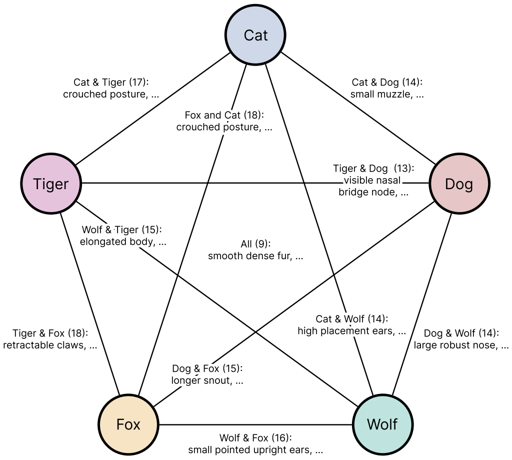


Fig. 9: Excerpt of the Cross-Paired Features from the Grounding Stage. Number of Shared Features in Parentheses.

5.3 Discussion

In this brief preliminary study, we acknowledge several limitations.

MLM Hallucinations: The translation phase relies on the MLM faithfully reporting what is in the heat map and – more crucially – following the instructions without relying on a predisposition of knowledge regarding the features (e.g., first recognizing the image as a Cat, and then deducing the necessary features of the Cat). If the MLM hallucinates a feature, the subsequent logic will be sound but based on false premises. However, at least this process generates a reasoning trace that can be followed and analyzed compared to the direct classification approach of the CNN.

Prompt Sensitivity: We observed that adherence to prompt instructions varies significantly across different MLM. Therefore, our analysis is limited to the MLM and prompts used for our experiment.

Ontology Completeness: Our *KB* was limited to morphological features. It currently lacks spatial reasoning (e.g., ears must be above the eyes), which could lead to false positives.

6 Further developments

The results of our proposed framework and exploratory study are encouraging, and open up several future research directions, in the usage of MLMs to ground subsymbolic predictions in symbolic logic, particularly regarding the reliability of the translation and the methodology of knowledge base construction.

Automated Ontology Alignment Currently, our knowledge base *KB* relies on a semi-automated generation process with human verification. A significant barrier arises in the scaling of this approach with a vast number of classes. Future work will need to investigate the ontology alignment problem, where the MLM extracts features from images and proposes updates to the *KB* upon encountering consistent visual evidence that contradicts the current encoded knowledge. This would enable the system to move from a pure static verifier into a dynamic learner, refining its own symbolic world model over time.

Spatial and Relational Logic The current implementation treats features as unstructured data. This approach is susceptible to simple adversarial attacks where features are present but spatially incoherent. We plan to extend the intersymbolic translation to extract spatial predicates, as has been done in previous work [25]. By encoding these spatial constraints, we can enforce a stricter form of structural verification, ensuring that the model is not merely relying on patches but coherent physical objects.

Explicit Feedback Loops Perhaps the most promising direction is moving from passive verification to active correction and learning. Currently, our pipeline acts as a passive verifier, identifying discrepancies without introducing correction mechanisms to the underlying classifier. Future research will explore explicit feedback loops, where the output of the abductive solver serves as a supervising signal. This *semantic loss* will enable the model to unlearn spurious correlations and align its internal feature representation with causal domain logic.

Formalization of the Problem Domain Beyond the technical verification, our approach introduces a methodological shift, because of the explicit construction *KB*, the system designer is forced to explicitly articulate the causal definitions of a class, rather than relying on implicit – and error prone – correlations. This process inherently facilitates the discovery of edge cases. Future work will explore tool assistance in systematic identification of logical gaps where ontology fails to cover OOD samples like the Wolf or Fox.

Computational Scalability and Efficiency The computational cost of this intersymbolic pipeline, especially when involving large-scale MLMs and probabilistic reasoning engines raises practical concerns. Efficient implementations, caching strategies, and selective reasoning mechanisms will be essential to make this approach scalable and applicable to real-world scenarios. Furthermore, our initial results indicate that the model size and internal architecture highly influence the reasoning capability, as already observed in contemporary research. Therefore, in future research, it would be exciting to investigate whether the mechanisms and distinct stages we have shown could be more directly integrated into the MLM architecture to support native, yet logically sound, abductive reasoning while maintaining scalability.

7 Conclusion

While the combination of explainability, abduction, and MLMs offers interesting perspectives for intersymbolic AI, it also presents several open challenges. Addressing these will require interdisciplinary collaboration across machine learning, formal methods, and knowledge representation communities. We posit that this line of research extends beyond fundamental acxai, benefiting high-stakes disciplines like medicine or law, which require AI models to be explainable for decision-making.

Data availability Our full experiment code and dataset are made available at https://github.com/NicolasSchuler/abduction-demo for reference.

Acknowledgments. This work was supported by funding from KASTEL Security Research Labs, Karlsruhe and the pilot program Core Informatics at KIT (KiKIT) of the Helmholtz Association (HGF).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allen, B.P., Chhikara, P., Ferguson, T.M., Ilievski, F., Groth, P.: Sound and complete neurosymbolic reasoning with llm-grounded interpretations. CoRR **abs/2507.09751** (2025). https://doi.org/10.48550/arxiv.2507.09751, https://doi.org/10.48550/arXiv.2507.09751
2. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion **58**, 82–115 (2020). https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012, https://www.sciencedirect.com/science/article/pii/S1566253519308103
3. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 839–847 (March 2018). https://doi.org/10.1109/WACV.2018.00097

4. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: Deep learning for interpretable image recognition. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper_files/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf
5. He, F., Chen, Z., Liang, X., Ma, T., Qiu, Y., Wu, S., Yan, J.: Protoreasoning: Prototypes as the foundation for generalizable reasoning in llms. CoRR **abs/2506.15211** (2025). https://doi.org/10.48550/arxiv.2506.15211, https://doi.org/10.48550/arXiv.2506.15211
6. Hoffman, R.R., Miller, T., Clancey, W.J.: Psychology and AI at a Crossroads: How Might Complex Systems Explain Themselves? The American Journal of Psychology **135**(4), 365–378 (Dec 2022). https://doi.org/10.5406/19398298.135.4.01
7. Josephson, J.R., Josephson, S.G.: Abductive Inference: Computation, Philosophy, Technology. Cambridge University Press, Cambridge (Oct 2009)
8. Kowalski, R.A.: Logic for problem solving, The computer science library : Artificial intelligence series, vol. 7. North-Holland (1979), https://www.worldcat.org/oclc/05564433
9. LeCun, Y., Bengio, Y., Hinton, G.E.: Deep learning. Nat. **521**(7553), 436–444 (2015). https://doi.org/10.1038/NATURE14539, https://doi.org/10.1038/nature14539
10. Li, R., Li, L., Ren, S., Tian, H., Gu, S., Li, S., Yue, Z., Wang, Y., Ma, W., Yang, Z., Ma, J., Sui, Z., Luo, F.: Groundingme: Exposing the visual grounding gap in mllms through multi-dimensional evaluation (2025), https://arxiv.org/abs/2512.17495
11. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 4765–4774 (2017), https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
12. Marconato, E., Bortolotti, S., van Krieken, E., Morettin, P., Umili, E., Vergari, A., Tsamoura, E., Passerini, A., Teso, S.: Symbol grounding in neuro-symbolic AI: A gentle introduction to reasoning shortcuts. CoRR **abs/2510.14538** (2025). https://doi.org/10.48550/arxiv.2510.14538, https://doi.org/10.48550/arXiv.2510.14538
13. Mileo, A.: Towards a neuro-symbolic cycle for human-centered explainability. Neurosymbolic Artificial Intelligence **1**, NAI–240740 (2025). https://doi.org/10.3233/NAI-240740, https://doi.org/10.3233/NAI-240740
14. Miller, T.: Explainable AI is dead, long live explainable ai!: Hypothesis-driven decision support using evaluative AI. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023. pp. 333–342. Acm (2023). https://doi.org/10.1145/3593013.3594001, https://doi.org/10.1145/3593013.3594001
15. Moraffah, R., Karami, M., Guo, R., Raglin, A., Liu, H.: Causal interpretability for machine learning - problems, methods and evaluation. SIGKDD Explor. **22**(1), 18–33 (2020). https://doi.org/10.1145/3400051.3400058, https://doi.org/10.1145/3400051.3400058
16. Padalkar, P., Gupta, G.: Symbolic rule extraction from attention-guided sparse representations in vision transformers. Theory and Practice of Logic Programming **25**(4), 722–738 (2025). https://doi.org/10.1017/s1471068425100318
17. Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.: Language models as knowledge bases? In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.)

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 2463–2473. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1250, https://aclanthology.org/D19-1250/

18. Poole, D.: Probabilistic horn abduction and bayesian networks. Artif. Intell. **64**(1), 81–129 (1993). https://doi.org/10.1016/0004-3702(93)90061-F, https://doi.org/10.1016/0004-3702(93)90061-F
19. Raedt, L.D., Kimmig, A., Toivonen, H.: Problog: A probabilistic prolog and its application in link discovery. In: Veloso, M.M. (ed.) IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007. pp. 2462–2467 (2007), http://ijcai.org/Proceedings/07/Papers/396.pdf
20. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144. Acm (2016). https://doi.org/10.1145/2939672.2939778, https://doi.org/10.1145/2939672.2939778
21. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
22. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. **128**(2), 336–359 (2020). https://doi.org/10.1007/s11263-019-01228-7, https://doi.org/10.1007/s11263-019-01228-7
23. Sheth, A.P., Roy, K., Gaur, M.: Neurosymbolic AI - why, what, and how. CoRR **abs/2305.00813** (2023). https://doi.org/10.48550/arxiv.2305.00813, https://doi.org/10.48550/arXiv.2305.00813
24. Speith, T.: A review of taxonomies of explainable artificial intelligence (XAI) methods. In: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022. pp. 2239–2250. Acm (2022). https://doi.org/10.1145/3531146.3534639, https://doi.org/10.1145/3531146.3534639
25. Suchan, J., Bhatt, M., Wałega, P., Schultz, C.: Visual explanation by high-level abduction: On answer-set programming driven reasoning about moving objects. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Apr 2018). https://doi.org/10.1609/aaai.v32i1.11569, https://ojs.aaai.org/index.php/AAAI/article/view/11569
26. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. Pmlr (2019), http://proceedings.mlr.press/v97/tan19a.html
27. Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388