Concept-RuleNet: Grounded Multi-Agent Neurosymbolic Reasoning in Vision Language Models

Sanchit Sinha, Guangzhi Xiong, Zhenghao He, Aidong Zhang
University of Virginia, USA
{sanchit,guangzhi,zhenghao,aidong}@virginia.edu

Abstract

Modern vision-language models (VLMs) deliver impressive predictive accuracy yet offer little insight into 'why' a decision is reached, frequently hallucinating facts, particularly when encountering out-of-distribution data. Neurosymbolic frameworks address this by pairing black-box perception with interpretable symbolic reasoning, but current methods extract their symbols solely from task labels, leaving them weakly grounded in the underlying visual data. In this paper, we introduce a multi-agent system - **Concept-RuleNet** that reinstates visual grounding while retaining transparent reasoning. Specifically, a multimodal concept generator first mines discriminative visual concepts directly from a representative subset of training images. Next, these visual concepts are utilized to condition symbol discovery, anchoring the generations in real image statistics and mitigating label bias. Subsequently, symbols are composed into executable first-order rules by a large language model reasoner agent - yielding interpretable neurosymbolic rules. Finally, during inference, a vision verifier agent quantifies the degree of presence of each symbol and triggers rule execution in tandem with outputs of black-box neural models, predictions with explicit reasoning pathways. Experiments on five benchmarks, including two challenging medical-imaging tasks and three underrepresented natural-image datasets, show that our system augments state-of-the-art neurosymbolic baselines by an average of 5% while also reducing the occurrence of hallucinated symbols in rules by up to 50%. An extended version of the paper can be found at https://arxiv.org/abs/2511.11751.

Introduction

With the increasing size and complexity of pre-trained large-scale Vision Language Models (VLMs) achieving widespread success in diverse vision tasks, it is tempting to utilize them for diverse use cases. Usually, VLMs are pre-trained on vast amounts of paired image-text data, which makes their learned decision-making process increasingly *misaligned* with the human thought process. Researchers refer to such systems as **System-1** due to their speed, scalability, and unintuitive reasoning pathways behind a prediction. On the other hand, the human thought process is slower, more deliberate, and *logical* - often composing multiple semantics *neurosymbolically* to reach a more accurate and trustworthy decision - classified as **System-2** (Nye et al. 2021). For example, consider the image in Figure 1, where both VLMs and logical rules output the same prediction, but System-2 reasoning is much more explainable and intuitive. Most of the current research on *alignment problem* focuses on inducing System-2 reasoning in System-1 reasoners during pre-training, chain of thought (Wei et al. 2022), etc., with limited success due to the fundamental assumptions geared towards scalability and efficiency. However, human cognition naturally weaves together System-1 (fast, associative) and System-2 (slow, deliberative) reasoning. Recognizing this, recent research has begun combining both systems to harness System-1's efficiency while using System-2 mechanisms to refine predictions and provide transparent, step-by-step rationales - thus leveraging both System-1 and System-2 *in tandem*.

Popular works that combine System-1 and System-2 reasoning often utilize meticulously curated symbols to form neurosymbolic rules, and the prediction score of a rule is computed using *First-order Logic*. For example, in (Yi et al. 2018), the authors utilize several visual classifiers to select relevant functional tools that output the likelihood of a particular concept present in the image, which is then composed using a curated logical rule. Although such approaches can be an effective solution in a closed setting with limited rules and a closed set of concepts (e.g., in (Yi et al. 2018), only four types of objects are considered), they are often not generalizable to large-scale complex datasets. As a consequence, a relatively new line of research proposes utilizing external knowledge from Large Language Models (LLMs) to automatically generate symbols (Generation) and subsequently construct logical rules using the generated symbols through neurosymbolic composition (Reasoning). For example, Symbol-LLM (Wu et al. 2024) utilizes an LLM (GPT-3.5) first to extract all relevant symbols relating to human activity labels and then uses the same LLM as a reasoner for rules construction, overcoming expensive and slow manual logical rule generation.

Even though approaches like (Wu et al. 2024) utilize LLMs as agents, they implicitly make a strong assumption that the parametric knowledge encoded in LLMs (during pre-training) is sufficient for effective symbol and rule generation. Note that the symbol discovery process in such approaches is conditioned *only on a single task label*, with **no**

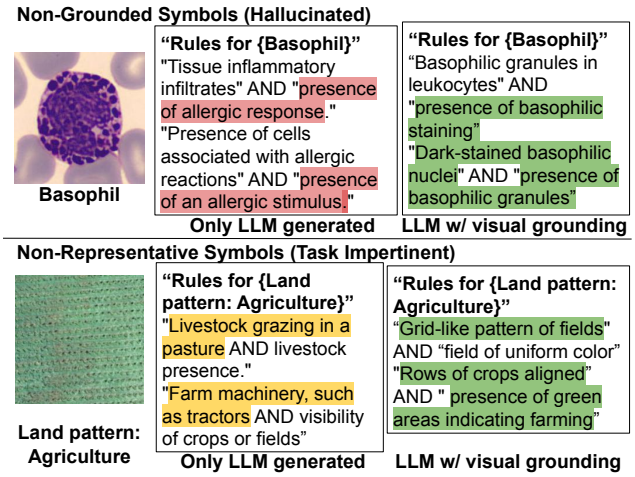

Figure 1: Examples sampled from the BloodMNIST (Yang et al. 2023a) and UC-Merced Land Use (Yang and Newsam 2010) test datasets demonstrating a sample from the 'Basophil' and 'Agriculture Land Pattern' classes, respectively. We list the top rules influencing the decision-making process. (TOP) We observe that utilizing no images during the symbolic rule generation process (Symbol-LLM) generates rules with non-grounded symbols, i.e., symbols NOT present in test images (hallucinations). (BOTTOM) We observe that the generated rules are often somewhat semantically related to the task label but not representative of the task. The highlighted symbols are most relevant for prediction, with red symbols being hallucinated, green being appropriate, and yellow being non-representative.

information from the actual training images in the dataset. This presents a significant problem with datasets out-of-distribution to LLM pre-training, where the LLM parametric knowledge is lacking (Li et al. 2024, 2023b). Utilizing training images during the symbol generation and rule formation process provides multiple benefits, namely, grounding and representativeness as discussed below.

Grounding: Firstly, as LLMs are susceptible to hallucinations on contexts with limited knowledge (Simhi et al. 2025; Zhang et al. 2024a), the symbols and rules generated can be adulterated with *outright incorrect symbols* that are never encountered in the dataset. Consider the top example in Figure 1. Here, the symbols 'presence of allergic response' and 'presence of allergic stimulus' are encountered in neither the training set nor the test set - implying the symbol is hallucinated in context to this setting. However, conditioning symbol generation on training images adds semantic context during the generation process and is essential to ensure grounding.

Representativeness: Secondly, one of the main reasons why LLMs output irrelevant symbols for underrepresented tasks and surprisingly accurate symbols for a select few tasks is not due to their extensive parametric encoded knowledge, but due to the well-known phenomenon of *dataset leakage* during pretraining (Carlini et al. 2021) wherein LLMs are overfitted on descriptions of commonly utilized benchmark datasets. This phenomenon is subtle, but it is clear to observe in the results presented in (Wu et al. 2024) and the similar work (Zhang et al. 2024b), where performance on in-domain datasets is benchmarked, but on out-of-domain data is lacking. Similarly, consider the symbols generated for the 'Land Pattern: Agriculture' class in Figure 1 using (Wu et al. 2024), where the symbols are relevant to the description of the class but are irrelevant for the task at hand, i.e., recognizing land use patterns - making the symbols non-representative of the task.

As a consequence, in this paper, we propose **Concept-RuleNet** - a collaborative multi-agent framework which enforces **grounding** and **representativeness** in the neurosymbolic rule generation process. Concept-RuleNet - (i) effectively leverages a subset of training images to first extract grounded concepts using a **Visual Concept Extraction agent**, (ii) creates neurosymbolic symbols and composes them into logical rules using a strong **Symbol Exploration and Neurosymbolic agent**, and (iii) further augments standard System-1 prediction with symbolic predictions using a **Verifier agent**. With this three-agent system, we achieve the desired symbol properties by conditioning automatic symbol generation not only on the target labels but also grounded descriptions extracted as visual concepts from images in the training set. Subsequently, we demonstrate the drawbacks of current label-conditioned symbols and rule formation methods by empirically evaluating the improved prediction performance and degree of grounding compared to Concept-RuleNet. Lastly, we propose an extension, Concept-RuleNet++, which not only utilizes relevant symbols but also utilizes counterfactual symbols as a combination of conjunctive and disjunctive rule formation. More objectively, our contributions are as follows:

- We propose Concept-RuleNet, a neuro-symbolic multi-agent system that utilizes three distinct LLM and VLM agents collaborating to extract grounded visual concepts and create representative and grounded symbols.
- We benchmark Concept-RuleNet across 5 challenging datasets on 4 modern VLMs and empirically demonstrate Concept-RuleNet's superior performance as compared to the state of the art (SOTA) approaches.
- We show that Concept-RuleNet produces **grounded** and **representative** symbols for more accurate rule generation and reduced hallucinations of LLM.
- We propose an extension - Concept-RuleNet++, which augments propositional rule generation in System-2 reasoning systems by leveraging **counterfactual** symbols for even higher prediction performance.

Related Work

Neurosymbolic Reasoning. Neurosymbolic reasoning seeks to bridge the gap between the high-dimensional, often opaque representations learned by deep neural networks and the discrete, interpretable symbols fundamental to human reasoning (Besold et al. 2021). Early works in this area utilized specialized architectures and regularization techniques to explain deep neural networks (DNNs) via propositional logic (Riegel et al. 2020; Dong et al. 2019;

Garcez and Lamb 2023; Sinha, Xiong, and Zhang 2025a). Building on these foundations and informed by taxonomical frameworks such as that proposed by Nye et al. (2021), recent research has increasingly aimed to integrate fast, intuitive System-1 processes with slower, deliberative System-2 reasoning (Saha et al. 2024; Wu et al. 2024; Mao et al. 2019). Moreover, studies combining concept-based explanations with neurosymbolic approaches have enhanced both interpretability and robustness (Barbiero et al. 2023; Sinha et al. 2023).

Utilizing Agents to Augment Black-box Models. Multiple recent approaches leverage the extensive semantic knowledge encoded in large language models (LLMs) to directly generate meaningful symbolic representations and enhance the reasoning capabilities of VLMs and supplement their limited linguistic understanding (Chen et al. 2023). For example, methods such as those in (Oikarinen et al. 2023) leverage the inherent language understanding of LLMs to extract meaningful symbols, while other works (Moayeri et al. 2023; Yang et al. 2023b) not only generate these concepts but also align them with visual data (VLMs). By harnessing both pre-training knowledge and the in-context learning capabilities of LLMs, these approaches are able to generate semantically rich symbols that serve as a bridge between raw visual inputs and higher-level reasoning tasks (Wu et al. 2024; Zhang et al. 2024b). Approaches such as (Cho et al. 2023; Hu et al. 2023; Sinha, Xiong, and Zhang 2025b) incorporate rich language cues into VLM inference through mechanisms like scene graphs or language priors, while other work (Zhou et al. 2023) directly feeds LLM outputs into the visual understanding process.

Comparisons to Related Work. Our approach can be directly compared against agentic neurosymbolic systems, which leverage LLMs as symbol extractors and logical rule generators. We compare against Symbol-LLM (Wu et al. 2024), which relies solely on task labels as conditioning for generating symbols and rules, whereas our framework leverages visual concepts to induce grounding and representativeness. Even though Symbol-LLM achieves benchmark performance on HICO and Stanford datasets, the methodology is not generalizable to underrepresented or out-of-domain datasets.

Methodology

In this section, we discuss the proposed Concept-RuleNet approach. We first begin by detailing the problem setting and formalizing notations. Next, we discuss the three primary stages of Concept-RuleNet, namely, Image-conditioned Visual Concept Extraction, Context-dependent Conditional Symbol Exploration and Rule Formation, and Neurosymbolic Rule-based Predictions. Finally, we discuss an extension to Concept-RuleNet - Concept-RuleNet++, which leverages counterfactual symbols to augment the neurosymbolic reasoning process. A schematic diagram of Concept-RuleNet is depicted in Figure 2. The subsequent inference process is depicted in Figure 3.

Preliminaries and Problem Setting

System-1: Let $\mathcal{X}$ be the space of input images and $\mathcal{Y}$ be the set of corresponding class labels. A typical System-1 model for image classification can be mathematically defined as a function mapping $F_{\text{sys1}} : \mathcal{X} \rightarrow \mathcal{Y}$ such that,

$$y = F_{\text{sys1}}(x) \;\; \forall \, (x, y) \in \{(\mathcal{X}, \mathcal{Y})\}$$

The function F_{sys1} can be modeled as a neural network. Note that in this work, we primarily consider a *zero-shot* setting where the System-1 model is used off the shelf, as fine-tuning large System-1 models is extremely expensive.

System-2. As opposed to learning a single function F_{sys1}, a System-2 model can be thought of as a composition of three separate functions $F_{concept}$, $F_{neurosymbolic}$, and F_{verify} such that $F_{\text{sys2}} = F_{verify} \circ F_{neurosymbolic} \circ F_{concept}$. Note that $\circ$ represents function composition, i.e., $F_{neurosymbolic} \circ F_{concept}$ represents the output of $F_{concept}$ is input to $F_{neurosymbolic}$. More precisely, the function $F_{concept}$ maps $\mathcal{X} \rightarrow \mathcal{P}(\mathcal{C})$, where $\mathcal{P}(\mathcal{C})$ represents the power set of $\mathcal{C}$, a set consisting of relevant, human-understandable descriptive *concepts*. Mathematically,

$$c = F_{concept}(x), \text{ s.t., } c \subseteq \mathcal{C} \tag{1}$$

Note that concepts $\mathcal{C}$ represent human-understandable descriptions of the images. Next, the function $F_{neurosymbolic}$ utilizes these visual concepts to explore task-relevant symbols, which are further composed into logical rules. Formally, $F_{neurosymbolic} : \mathcal{P}(\mathcal{C}) \rightarrow \mathcal{P}(\mathcal{L})$. Let $\mathcal{S}$ be the vocabulary of binary atomic symbols. A logical rule $l \in \mathcal{L}$ is formed by composing multiple symbols $s_i \in \mathcal{S}$. Mathematically,

$$l = F_{neurosymbolic}(c), \text{ s.t., } l \subseteq \mathcal{L} \tag{2}$$

where each $l_i \in l$ is of the form $l_i = \bigwedge_{s_i \in s} s_i$. Finally, the function F_{verify} maps $\mathcal{P}(\mathcal{L}) \rightarrow \mathcal{Y}$ by implicitly scoring each symbol in a rule and then aggregating the scores into an entailment confidence, and returning the rule's prediction. Mathematically,

$$F_{sys2} = F_{verify}(l), \text{where } l = F_{neurosymbolic} \circ F_{concept}(x) \tag{3}$$

To leverage both System-1 and System-2 reasoning together in prediction, the final composite prediction is expressed by a weighted sum of System-1 and System-2 reasoning models:

$$\hat{y} = (1 - \lambda) F_{sys1}(x) + \lambda F_{sys2}(x) \tag{4}$$

where λ controls the influence of System-2 model's prediction to the final prediction. In practice, all the individual functions of F_{sys2}, i.e., $F_{concept}$, $F_{neurosymbolic}$ and F_{verify} are implemented using agents as discussed below.

Image-conditioned Visual Concept Extraction

As discussed, a System-2 reasoning model requires the generation of logical rules to emulate the human reasoning process. As discussed in Sec , the first stage is represented as the function $F_{concept}$ and generates a set of grounded and representative visual concepts. We utilize the VLM agent ($\mathcal{A}_{\mathcal{V}}$) to first extract visual concepts present in each training image.

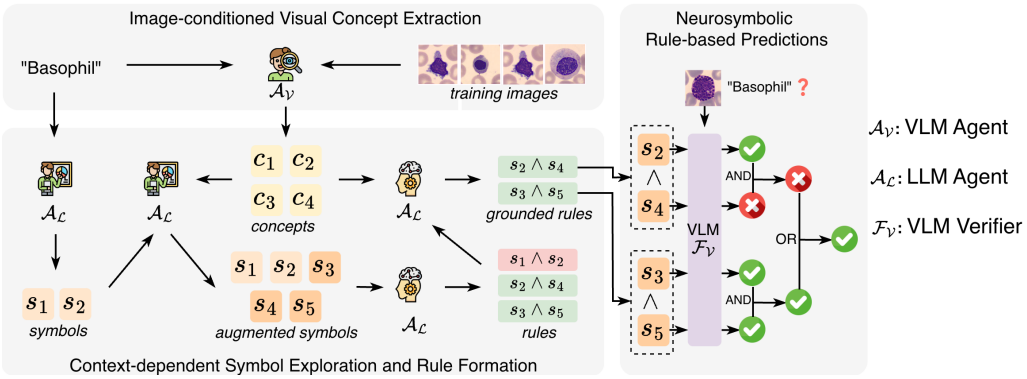

Figure 2: Schematic figure of Concept-RuleNet approach. Concept-RuleNet operates in three sequential stages - (i) Grounded Visual Concept Extraction: outputs visual concepts grounded in representative training images, (ii) Conditional Symbol Generation and Neurosymbolic Rule-based Predictions: explores relevant symbols and composes them into logical rules, and (iii) Neurosymbolic Rule-based Predictions: verifies the presence of each symbol in the rule to provide final predictions.

Recent research has found that VLMs are extremely effective in identifying attributes in the images but less effective in identifying complex relationships between the discovered attributes, and hence act better as 'Bag-of-visual attributes' than extracting complex relationships (Doveh et al. 2023; Herzig et al. 2023). Hence, we extract low-level visual concepts $c_y \in \mathcal{P}(\mathcal{C})$ for each training image $x \in \mathcal{X}_{train}$ in the dataset, conditioned also on the task label y. Mathematically,

$$c_y = \bigcup \mathcal{A}_\mathcal{V}(x_i, y, M),\ \forall\ x_i\ \in \mathcal{X}_y^{\text{train}} \tag{5}$$

where $\mathcal{A}_\mathcal{V}$ is a function characterized by the task label y and a number of concepts M, and $\mathcal{X}_y^{\text{train}}$ is the training subset of $\mathcal{X}$ with labels y. The visual concepts for randomly selected images belonging to a particular label y in the training set are extracted and appended to a set, discarding duplicates. Finally, we get a set c_y of observed visual concepts for each task label $y \in \mathcal{Y}$.

Context-dependent Symbol Exploration and Rule Formation

In the next stage, the function $F_{neurosymbolic}$ generates symbols and logical rules conditioned on the visual concepts. This stage has two distinct components - exploration and rule-formation, discussed below.

Exploration. To form grounded and representative symbols, we utilize a strong linguistic agent depicted by $\mathcal{A}_\mathcal{L}$. The process of symbol discovery is conditioned on both the task labels y and the generated concepts c_y in the last stage. To begin, we initialize the symbol set with dataset-specific symbols generated using an Initialization Symbol function ($\mathcal{IS}$), which is characterized by the task label y and the number of initial symbols K. Mathematically, the symbol set is initialized as:

$$S = \mathcal{A}_\mathcal{L}(\mathcal{IS}(y, K)) \tag{6}$$

Subsequently, we begin iterative rule exploration using both the initial symbols and the concepts associated with each label, as collected in the last stage. The visual concepts provide grounded context to $\mathcal{A}_\mathcal{L}$ during exploration to minimize hallucinations. We utilize an Explore-Symbol function ($\mathcal{ES}$) characterized by the task label y with context c_y. Mathematically,

$$S = S \bigcup \mathcal{A}_\mathcal{L}(\mathcal{ES}(c_y, y)) \tag{7}$$

where the set S collects all explored symbols following certain constraints as discussed in the next section.

Rule Formation. The next stage composes the symbols generated during the Exploration stage into logical neurosymbolic rules through the linguistic agent $\mathcal{A}_\mathcal{L}$ and an entailment function $\mathcal{EN}$. To construct a rule, we utilize the initialized symbol set S as input, and each new symbol discovered during the explore stage is evaluated in the form of a rule in the Disjunctive Normal Form (DNF). Mathematically, a set of rules l^* can be formed as:

$$l^* = \{\bigwedge s_i \rightarrow y\}, \text{where } s_i \in S \text{ and } y \in \mathcal{Y} \tag{8}$$

To ascertain the soundness of each rule in l^*, we utilize $\mathcal{A}_\mathcal{L}$ to calculate entailment. Mathematically, it can be written as:

$$l = \{\ l_i \in l^*\ \mid\ \mathcal{A}_\mathcal{L}(\mathcal{EN}(c_y, l_i)) > \epsilon\} \tag{9}$$

where $\mathcal{A}_\mathcal{L}$ calculates the entailment for a rule l_i and ϵ is the entailment threshold signifying if the rule is plausible. Note that the rule scoring is dependent not just on the labels y, but also on the visual concepts extracted in the previous stage.

Based on the scoring mechanism, only the rules above a pre-defined threshold (ϵ) are considered for System-2 reasoning. We limit the length of each rule to N symbols, preventing overfitting a particular rule to the task. We point out that the functions $\mathcal{IS}$, $\mathcal{ES}$, and $\mathcal{EN}$ are modeled as prompt templates, which are detailed in the extended version.

Neurosymbolic Rule-based Predictions

For the final System-2 prediction, the process is decomposed into two steps: (i) computing scores for individual symbols using the *verifier agent* ($\mathcal{F}_\mathcal{V}$), and (ii) aggregating symbol scores to evaluate confidence of a rule.

For a neuro symbolic rule l_i of the form $l_i = s_{i_1} \wedge s_{i_2} \wedge \cdots \wedge s_{i_k}$, where $\{s_{i_1}, s_{i_2}, \ldots, s_{i_k}\} \in \mathcal{P}(\mathcal{S})$ and a test image $x \in \mathcal{X}^{test}$, the overall score is then computed by taking the minimum of the scores of the individual symbols:

$$F_{verify}(l) = \max_{i \in \{1,..,|l|\}} \{\min\{\mathcal{F}_{\mathcal{V}}(x, s_{i_1})..,\mathcal{F}_{\mathcal{V}}(x, s_{i_k})\}\}. \tag{10}$$

where $\mathcal{F}_{\mathcal{V}}$ is a VLM, and the final System-2 prediction over all rules in $\mathcal{L}$.

Sample Inference Process. An example of an inference procedure is demonstrated in Figure 3. During inference, the image is passed through the System-1 model to infer the probability of each class label. (Basophil=0.48 and Eosinophil=0.52). In parallel, the Verifier Agent predicts the likelihood of each symbol for all neurosymbolic rules. For each class, the most likely rule is calculated using Equation 10. The final class prediction is performed through a weighted sum of System-1 (Basophil=0.48) and neurosymbolic output (Basophil=0.95). As can be seen, neurosymbolic output 'corrects' System-1 output to instill higher trust.

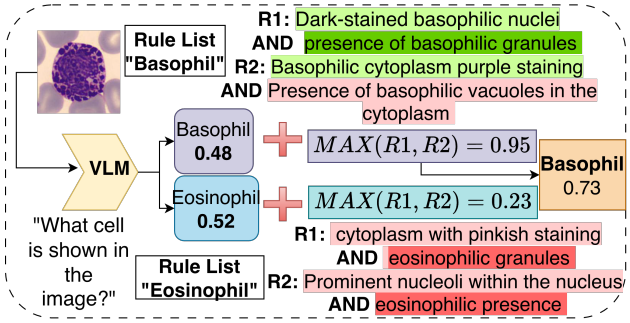

Figure 3: Inference process for a sample from the 'Basophil' class from BloodMNIST dataset. VLM inference output assigns a probability score of 0.48 to correct class.

Concept-RuleNet++

Recent work has underscored the value of counterfactual reasoning in enhancing the interpretability of neurosymbolic systems. Counterfactual symbols serve as a complement to relevant symbols, for clearer decision boundaries and to reduce the impact of spurious correlations. Approaches like (Wachter, Mittelstadt, and Russell 2017; Dandekar et al. 2023) empirically demonstrate that incorporating counterfactual symbols leads to improved generalization. We propose an extension to the standard Concept-RuleNet setup - Concept-RuleNet++ which augments reasoning with *counterfactual* symbols. We utilize symbols present in rules from other classes with inverse verification, i.e., the probability of not being present. Mathematically, the new logical rules are formed as,

$$l = \{\bigwedge\{\bigvee\{\tilde{s}_i, s_i\}\} \rightarrow y\}, \tag{11}$$

where $s_i, \tilde{s}_i \in S$ and $y \in \mathcal{Y}$. $\tilde{s}_i$ represents counterfactual symbols discovered. Note that Concept-RuleNet++ expands the rule structure to be a combination of both DNF and Conjunctive Normal Forms (CNF), forming a Mixed Normal Form structure of System-2 reasoning.

Experiments

Dataset and Model Description

Dataset Description. We utilize 5 medical and real-world datasets. **MedMNIST** (Yang et al. 2023a) - BloodMNIST and DermaMNIST are designed for blood cell and skin abnormality classification, respectively. **UC-Merced Satellite Land Use** (Yang and Newsam 2010) and **WHU** (Xia et al. 2010) are large-scale satellite image-based remote sensing datasets with high-resolution images for categorizing the land-use pattern. **iNaturalist-21** (Van Horn et al. 2018) consists of 13 classes referring to various biological species.

System-1 Models. We benchmark our approach on 3 open-source VLMs - InstructBLIP-XXL (Dai et al. 2023), LLaVA-1.5 (Liu et al. 2023), and LLaVA-1.6 (Team 2024). For MedMNIST, we also utilize a medical VLM - LLaVA-Med (Li et al. 2023a).

Visual Concept Extraction and Symbol-Generation Models. For the MedMNIST family of datasets, we utilize LLaVA-Med as a visual concept extractor agent ($\mathcal{A}_{\mathcal{V}}$) while for real-world datasets, we utilize LLaVA-1.6. The symbol exploration and rule formation agents are chosen to be strong LLMs with large-scale pre-training. We utilize **GPT-4o-mini** by OpenAI (OpenAI 2024) - a SOTA LLM with advanced reasoning capabilities.

Hyperparameter Settings

Visual Concept Extraction. We utilize LLaVA-Med for extracting visual concepts in the MedMNIST dataset and LLaVA-1.6 for the other datasets with temperature 0.2.

Symbol Generation. We utilize 5 initial premise symbols (N) followed by a maximum rule length of 3. To reduce overfitting, the rules on the visual context, we ensure maximum entailment (ϵ) scores greater than 0.7 for a rule to be relevant. Increasing rule sizes beyond 3 provides diminishing returns. We run recursive symbol exploration and rule composition for 10 and 7 iterations, respectively, for MedMNIST and other datasets. The temperature is set at 0.7 for exploration and 0 for entailment.

Dataset Specific. We utilize $\lambda = 0.5$ for the Blood and Derma datasets while $\lambda = 0.7, 0.5, 0.7$ for Satellite, WHU, and iNaturalist datasets respectively, based on tuning on the validation set.

Implementation Details

Baseline Replication. We recreate the Symbol-LLM baseline (Wu et al. 2024) by adapting it to different datasets. Note that Symbol-LLM is exclusively tested on Human Activity Recognition (HOI) datasets (which, as discussed before, is in-domain to LLM pre-training), hence its efficacy on the selected datasets is unknown. We utilize the implementation with minor changes in prompts by changing task labels corresponding to the datasets used in this paper. We limit the rule lengths to 3 symbols each, with a lowered minimum entailment value of 0.7 (to achieve speed up). As demonstrated in (Wu et al. 2024), rules with at most 3 symbols are functionally equivalent to longer rules.

Verification. Although VLMs are adept at generating natural language descriptions of input images, the actual token

	BloodMNIST			DermaMNIST			UCMerced-Satellite			WHU			iNaturalist		
	S1	**S-LLM**	**CRN**	**S1**	**S-LLM**	**CRN**	**S1**	**S-LLM**	**CRN**	**S1**	**S-LLM**	**CRN**	**S1**	**S-LLM**	**CRN**
Verifier: Same as System-1															
InstructBLIP	11.55	13.56	**18.09**	5.05	5.05	**8.54**	41.33	48.0	**57.33**	14.28	14.28	**20.40**	52.13	52.65	**53.21**
LLaVA-1.5	11.55	10.05	**14.57**	9.54	30.15	**47.73**	65.33	64.00	**69.33**	18.36	18.36	**19.22**	58.12	**60.24**	**60.24**
LLaVA-1.6	10.05	9.67	**19.35**	36.68	38.19	**48.74**	38.66	49.33	**50.66**	27.27	**28.43**	**28.43**	61.30	61.30	**63.45**

Table 1: Comparison between prediction accuracy on BloodMNIST, DermaMNIST, UCMerced-Satellite, WHU, and iNaturalist datasets across multiple VLMs. The columns under each dataset indicate System-1 (S1), S1 augmented with Symbol-LLM (S-LLM) and Concept-RuleNet (CRN), respectively.

	BloodMNIST			DermaMNIST		
Model	**S1**	**S-LLM**	**CRN**	**S1**	**S-LLM**	**CRN**
InstructBLIP	11.55	11.55	**13.56**	5.05	5.05	**7.86**
LLaVA-1.5	11.55	**13.56**	**13.56**	9.54	30.15	**34.21**
LLaVA-1.6	10.05	9.54	**13.21**	36.68	48.25	**66.33**
LLaVA-Med	11.05	11.05	**12.06**	4.81	2.77	**12.56**

Table 2: Prediction accuracy for BloodMNIST and DermaMNIST datasets with a medical verifier agent - LLaVA-Med. We observe that LLaVA-Med boosts performance significantly on the DermaMNIST dataset.

predictions are underexplored. We augment VQAScore (Lin et al. 2025) scoring strategy by reformatting the visual description into a binary 'Yes/No' question. Mathematically, for a given $x \in \mathcal{X}^{test}$, the prediction probability of a symbol s for a VLM (F_V) with logit outputs ($\hat{F}_V$),

$$P(\text{“yes”}|x, s) = e^{\hat{F}_V[\text{“yes”}]}/(e^{\hat{F}_V[\text{“yes”}]} + e^{\hat{F}_V[\text{“no”}]}) \quad (12)$$

The probability of the token 'Yes' is taken as the proxy for the *confidence* of prediction.

Experiment-1: Prediction Performance

Choice of Verifier. As the verifier ($\mathcal{A}_V$) is one of the most important aspects of System-2 models, we consider 2 real-world settings - one where all the symbols and rules are pre-computed *ante-hoc* and only the rules and test-images are available. In this case, the System-1 model can act as a verifier with minor modifications. The next setting is where we are provided both test images and an *inference* only verifier. In this case, to provide more confident symbol probabilities, we can utilize a domain-specific verifier. We chose LLaVA-Med, a strong medical VLM, as a verifier.

Medical Datasets. We report the prediction performance of Concept-RuleNet as compared to baseline Symbol-LLM and only without any System-2 integrations (S1) in Table 1. We observe that Concept-RuleNet consistently outperforms System-1 only and Symbol-LLM baselines on most datasets. For the BloodMNIST dataset and DermaMNIST datasets, Concept-RuleNet beats Symbol-LLM by an average of about 5% across all model settings.

Real-world Datasets. Here, we observe that System-1 models demonstrate good performance out of the box on the Satellite datasets - UCMerced and WHU as they share resemble with types of images encountered in the pre-training setup of VLMs. We observe that Concept-RuleNet outperforms Symbol-LLM by a considerable margin of about 5% on most models, with the highest improvement being on the Instruct-BLIP-XXL model of 9.33% on the UCMerced dataset, while an average of 2-4% on the WHU dataset, possibly due to the more challenging nature of the images. Finally, the prediction performance of Concept-RuleNet is superior to Symbol-LLM on the iNaturalist dataset on all models. The results are a testament to our approach, as it improves prediction performance on underrepresented, out-of-distribution datasets.

Utilizing domain-specific verifiers. Next, we observe that using a medical verifier improves performance in both medical datasets in Table 2, making a strong case to consider designing even more powerful medical VLMs as verifiers in the future. Lastly, in Row-4, we conduct an interesting experiment where we utilize LLaVA-Med as both a System-1 model and a verifier, which interestingly does not give good results. This is an important insight - implying that models like LLaVA-Med do not have a deep understanding of the images (weak reasoning) but can be good verifiers.

Improvements using Concept-RuleNet++. Next in Table 3, we report the performance improvement using Concept-RuleNet++. Note as Concept-RuleNet++ utilizes counterfactual symbols, it is not directly comparable to Symbol-LLM but rather an approach similar to combining Symbol-LLM with counterfactual symbols which is out of scope for this work. We observe Concept-RuleNet++ outperforms Concept-RuleNet by an average of 1-2%, making it ideal for use cases that are performance-sensitive.

Dataset	**Concept-RuleNet**	**Concept-RuleNet++**
BloodMNIST	18.09	**21.43**
DermaMNIST	8.54	**14.23**
Satellite	57.33	**58.12**
WHU	20.40	**21.52**
iNaturalist	53.21	**54.15**

Table 3: Prediction performance improvements for Concept-RuleNet++ over Concept-RuleNet. We observe a consistent improvement in prediction performance over all datasets except iNaturalist, possibly due to the extreme diversity in the training samples, as the occurrence of a counterfactual symbol is still pretty high. (All datasets tested on InstructBLIP)

Experiment-2: Symbol Quality

Quantitative Grounding Measures. To validate if the symbols generated by Concept-RuleNet are better grounded, we compute the average likelihood of each symbol in the generated rules being present in both the train and test images using a VLM as shown in Figure 4. We observe that symbols

Initialization	Exploration	Entailment	Prediction
✗	✗	✗	48.00
✔	✗	✗	49.50
✔	✔	✗	55.10
✔	✔	✔	57.33

Table 4: Ablation study for presence of visual context in each stage. First row corresponds to Symbol-LLM setting.

generated by Concept-RuleNet are more likely to be present in both the Train and Test sets than those by Symbol-LLM. For Satellite and WHU datasets, the difference is even more stark, where Symbol-LLM's symbol occurrence rate is less than 0.5 - highlighting the need for Concept-RuleNet in underrepresented domains.

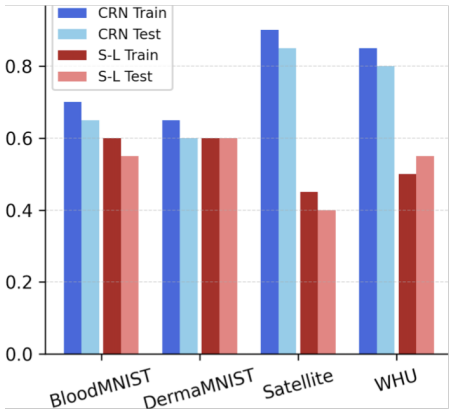

Figure 4: Degree of grounding on symbols generated by Concept-RuleNet (CRN) and Symbol-LLM (S-LLM) using the InstructBLIP model.

Representativeness of Symbols. Similarly, to evaluate the representativeness of generated symbols we format them as a question, 'How likely are {symbol1, symbol2,..} in predicting {class} for a {task}?' and pass them through an advanced reasoning model. We find that the average likelihood outputted for symbols generated by Concept-RuleNet is **0.54** as compared to 0.49 for Symbol-LLM (Refer to the extended version for experiment design). Figure 5 shows concepts and symbols generated using Concet-RuleNet and Symbol-LLM on the UCMerced Satellite dataset.

Ablation Study

Impact of Visual Concepts. We conduct an extensive ablation study by providing visual concepts as a context in each stage of the symbol generation and entailment process. Visual context improves each stage of Concept-RuleNet for the UC-Merced dataset, highlighting its usefulness.

Impact of hyperparameters. We further report the impact of λ, which controls the impact of System-2 reasoning on the final output and the initial number of images used for visual concept extraction in Table 5. We observe that too high or too low λ degrades performance. Similarly, considering too many images is detrimental due to overfitting on obscure, irrelevant concepts.

Method	λ			# of images	
	0.3	0.7	0.9	50	90
Symbol-LLM	44.00	48.00	38.66	48.00*	48.00*
Concept-RuleNet	49.33	57.33	48.00	57.33	56.28

Table 5: Ablation Study on the impact of hyperparameter λ and number of images used for visual concept extraction. (*) implies no training images are utilized.

Complexity analysis. The time complexity of visual concept extraction depends on the number of images selected. Assuming the number of images is n and the number of classes is c, the time complexity is given as $\mathcal{O}(nc)$. No other change in complexity between CRN and S-LLM.

Concept-RuleNet

Visual Concepts:
1. Grid-like pattern of buildings: satellite view of a residential area, and one of the most prominent features is grid-like pattern.
2. Shadows cast by the buildings.

Symbols:
"Multiple cars parked on street": 0,
"Rows of closely spaced houses":1,
"Shared sidewalks/pathways": 2,
"houses are built on small lots.": 3,

Rules:
1. "Rows of close space houses", "small lots.":**0.95**
2. "Small front yards with minimal space", "urban environ.":**0.93**
3. "Shared sidewalk paths","urban environment.": **0.93**

Symbol-LLM

Symbols:
"Close-packed apartments": 0,
"Multiple parked cars on the street": 1,
"Nearby playground or community area": 2,
"Sidewalks lined with pedestrians": 3,
"Small yards with limited space": 4,
"presence of multiple houses apartment buildings": 5,
"high demand for housing.": 6,
"high urban population.": 7,

Rules:
1. {"Close-packed apartment buildings", "high urban population} : **0.95**
2. { "Multiple parked cars on the street", "streets are lined with homes."} : **0.95**
3. {"high demand for housing.", "high urban population."}: **0.95**

Figure 5: (LEFT) Concept-RuleNet generated visual concepts, symbols, and rules. (RIGHT) Symbol-LLM generated symbols and rules for the 'denseresidential' class in the satellite dataset. We observe that Symbol-LLM outputs multiple symbols with high probability of being presented (represented as bold numbers in the rules), but which are non-representative of the task - classifying 'denseresidential'. E.g. 'high-demand for housing' is irrelevant.

Conclusion

In this paper, we propose Concept-RuleNet, a novel image-conditioned neurosymbolic reasoning framework designed to improve image classification performance. By leveraging both target labels and training images to generate grounded and representative symbols, our approach effectively mitigates the issues of hallucination and dataset leakage that have limited prior methods relying solely on label-conditioned symbol generation. The empirical evaluations across a diverse set of benchmark datasets demonstrate Concept-RuleNet's superior performance. Furthermore, we introduced Concept-RuleNet++, an extension that incorporates counterfactual symbols into the logical rule formation process. Overall, our work underscores the importance of integrating visual context into the neurosymbolic reasoning process and opens up promising avenues for future research.

Acknowledgments

This work is supported in part by the US National Science Foundation (NSF) and the National Institute of Health (NIH) under grants IIS-2106913, IIS-2538206, IIS-2529378, CCF-2217071, CNS-2213700, and R01LM014012-01A1. Any recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIH or NSF.

References

Barbiero, P.; Ciravegna, G.; Giannini, F.; Zarlenga, M. E.; Magister, L. C.; Tonda, A.; Lió, P.; Precioso, F.; Jamnik, M.; and Marra, G. 2023. Interpretable neural-symbolic concept reasoning. In *International Conference on Machine Learning*, 1801–1825. PMLR.

Besold, T. R.; d'Avila Garcez, A.; Bader, S.; Bowman, H.; Domingos, P.; Hitzler, P.; Kühnberger, K.-U.; Lamb, L. C.; Lima, P. M. V.; de Penning, L.; et al. 2021. Neural-symbolic learning and reasoning: A survey and interpretation 1. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 1–51. IOS press.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, 2633–2650.

Chen, L.; Li, B.; Shen, S.; Yang, J.; Li, C.; Keutzer, K.; Darrell, T.; and Liu, Z. 2023. Large language models are visual reasoning coordinators. *Advances in Neural Information Processing Systems*, 36: 70115–70140.

Cho, J.; Hu, Y.; Garg, R.; Anderson, P.; Krishna, R.; Baldridge, J.; Bansal, M.; Pont-Tuset, J.; and Wang, S. 2023. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*.

Dandekar, R.; et al. 2023. Counterfactual Symbols in Neurosymbolic Architectures for Robust Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1234–1243.

Dong, H.; Mao, J.; Lin, T.; Wang, C.; Li, L.; and Zhou, D. 2019. Neural logic machines. *arXiv preprint arXiv:1904.11694*.

Doveh, S.; Arbelle, A.; Harary, S.; Schwartz, E.; Herzig, R.; Giryes, R.; Feris, R.; Panda, R.; Ullman, S.; and Karlinsky, L. 2023. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2657–2668.

Garcez, A. d.; and Lamb, L. C. 2023. Neurosymbolic AI: The 3 rd wave. *Artificial Intelligence Review*, 56(11): 12387–12406.

Herzig, R.; Mendelson, A.; Karlinsky, L.; Arbelle, A.; Feris, R.; Darrell, T.; and Globerson, A. 2023. Incorporating structured representations into pretrained vision & language models using scene graphs. *arXiv preprint arXiv:2305.06343*.

Hu, Y.; Liu, B.; Kasai, J.; Wang, Y.; Ostendorf, M.; Krishna, R.; and Smith, N. A. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20406–20417.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Li, W.; Fan, H.; Wong, Y.; Yang, Y.; and Kankanhalli, M. 2024. Improving context understanding in multimodal large language models via multimodal composition learning. In *Forty-first International Conference on Machine Learning*.

Li, Y.; Hu, B.; Chen, X.; Ding, Y.; Ma, L.; and Zhang, M. 2023b. A multi-modal context reasoning approach for conditional inference on joint textual and visual clues. *arXiv preprint arXiv:2305.04530*.

Lin, Z.; Pathak, D.; Li, B.; Li, J.; Xia, X.; Neubig, G.; Zhang, P.; and Ramanan, D. 2025. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 366–384. Springer.

Liu, H.; Li, C.; Wu, Q.; and Li, Y. J. 2023. LLaVA: Large Language and Vision Assistant. *arXiv preprint arXiv:2304.08485*.

Mao, J.; Gan, C.; Kohli, P.; Tenenbaum, J. B.; and Wu, J. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*.

Moayeri, M.; Rezaei, K.; Sanjabi, M.; and Feizi, S. 2023. Text-to-concept (and back) via cross-model alignment. In *International Conference on Machine Learning*, 25037–25060. PMLR.

Nye, M.; Tessler, M.; Tenenbaum, J.; and Lake, B. M. 2021. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. *Advances in Neural Information Processing Systems*, 34: 25192–25204.

Oikarinen, T.; Das, S.; Nguyen, L. M.; and Weng, T.-W. 2023. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*.

OpenAI. 2024. GPT-4o Technical Report.

Riegel, R.; Gray, A.; Luus, F.; Khan, N.; Makondo, N.; Akhalwaya, I. Y.; Qian, H.; Fagin, R.; Barahona, F.; Sharma, U.; et al. 2020. Logical neural networks. *arXiv preprint arXiv:2006.13155*.

Saha, S.; Prasad, A.; Chen, J. C.-Y.; Hase, P.; Stengel-Eskin, E.; and Bansal, M. 2024. System-1. x: Learning to Balance Fast and Slow Planning with Language Models. *arXiv preprint arXiv:2407.14414*.

Simhi, A.; Itzhak, I.; Barez, F.; Stanovsky, G.; and Belinkov, Y. 2025. Trust Me, I'm Wrong: High-Certainty Hallucinations in LLMs. *arXiv preprint arXiv:2502.12964*.

Sinha, S.; Huai, M.; Sun, J.; and Zhang, A. 2023. Understanding and enhancing robustness of concept-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 15127–15135.

Sinha, S.; Xiong, G.; and Zhang, A. 2025a. ASCENT-ViT: Attention-based Scale-aware Concept Learning Framework for Enhanced Alignment in Vision Transformers. *arXiv preprint arXiv:2501.09221*.

Sinha, S.; Xiong, G.; and Zhang, A. 2025b. COCO-Tree: Compositional Hierarchical Concept Trees for Enhanced Reasoning in Vision-Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2695–2711.

Team, L. 2024. LLaVA-NeXT: Advancing Multimodal Large Language Models. Accessed: Jan 30, 2024.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2): 841–887.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, X.; Li, Y.-L.; Sun, J.; and Lu, C. 2024. Symbol-LLM: leverage language models for symbolic system in visual human activity reasoning. *Advances in Neural Information Processing Systems*, 36.

Xia, G.-S.; Yang, W.; Delon, J.; Gousseau, Y.; Sun, H.; and MaÎtre, H. 2010. Structural high-resolution satellite image indexing. Vienna, Austria.

Yang, J.; Shi, R.; Wei, D.; Liu, Z.; Zhao, L.; Ke, B.; Pfister, H.; and Ni, B. 2023a. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1): 41.

Yang, Y.; and Newsam, S. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 270–279.

Yang, Y.; Panagopoulou, A.; Zhou, S.; Jin, D.; Callison-Burch, C.; and Yatskar, M. 2023b. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19187–19197.

Yi, K.; Wu, J.; Gan, C.; Torralba, A.; Kohli, P.; and Tenenbaum, J. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *NeurIPS*, 31.

Zhang, Y.; Li, S.; Liu, J.; Yu, P.; Fung, Y. R.; Li, J.; Li, M.; and Ji, H. 2024a. Knowledge overshadowing causes amalgamated hallucination in large language models. *arXiv preprint arXiv:2407.08039*.

Zhang, Y.; Wang, X.; Chen, H.; Fan, J.; Wen, W.; Xue, H.; Mei, H.; and Zhu, W. 2024b. Large Language Model with Curriculum Reasoning for Visual Concept Recognition. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6269–6280.

Zhou, K.; Lee, K.; Misu, T.; and Wang, X. E. 2023. ViCor: Bridging Visual Understanding and Commonsense Reasoning with Large Language Models. *arXiv preprint arXiv:2310.05872*.