

# DOLPHIN: A Programmable Framework for Scalable Neurosymbolic Learning

Aaditya Naik<sup>1</sup> Jason Liu<sup>1</sup> Claire Wang<sup>1</sup> Amish Sethi<sup>1</sup> Saikat Dutta<sup>2</sup> Mayur Naik<sup>1</sup> Eric Wong<sup>1</sup>

## Abstract

Neurosymbolic learning enables the integration of symbolic reasoning with deep learning but faces significant challenges in scaling to complex symbolic programs, large datasets, or both. We introduce DOLPHIN, a framework that tackles these challenges by supporting neurosymbolic programs in Python, executing complex symbolic reasoning on the CPU while vectorizing probabilistic computations and gradient propagation on the GPU. Across 13 benchmarks spanning tasks over text, image, and video data, with symbolic reasoning features like recursion and black-box functions, DOLPHIN converges to state-of-the-art accuracies on the more complex benchmarks while existing frameworks such as Scallop, ISED, and IndeCateR+ fail to converge within the time limit. On simpler benchmarks, DOLPHIN matches their performance, while achieving these results 1.71x to 62x faster than the baselines. Overall, DOLPHIN advances the scalability of neurosymbolic frameworks, achieving state-of-the-art efficiency and convergence on difficult benchmarks where existing frameworks struggle. The code is published at <https://github.com/Dolphin-NeSy/Dolphin>.

## 1. Introduction

Deep learning has made great strides in tasks such as image classification, speech recognition, and natural language processing. With the emergence of foundation models like GPT-4 and SAM, deep learning is increasingly applied to more complex tasks. Despite significant strides, these models remain limited in their ability to reliably perform reasoning required for tasks involving structure, logic, and planning, where symbolic approaches traditionally excel (Kamb-

<sup>1</sup>Department of Computer and Information Science, University of Pennsylvania <sup>2</sup>Department of Computer Science, Cornell University. Correspondence to: Aaditya Naik <as-naik@seas.upenn.edu>.

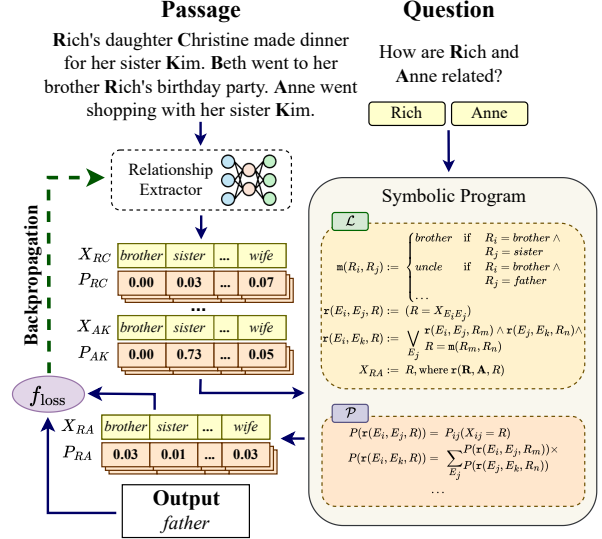


Figure 1: Illustration of the execution of a neurosymbolic program for the kinship reasoning task CLUTRR. While existing neurosymbolic frameworks run the neural models on the GPU, they run the symbolic program entirely on either the CPU (e.g. Scallop) or the GPU (e.g. Logic Tensor Networks), rendering them inefficient in terms of compute and memory, respectively. In DOLPHIN, both neural models and symbolic programs are specified as PyTorch modules, but only probabilistic computations ( $\mathcal{P}$ ) are vectorized on GPU whereas symbolic computations ( $\mathcal{L}$ ) execute on CPU.

hampati et al.). Neurosymbolic programming (Chaudhuri et al., 2021) has emerged as a promising paradigm to incorporate symbolic reasoning into deep learning models, providing the best of both worlds.

Various frameworks have been developed to improve the programmability and accessibility of neurosymbolic applications (Manhaeve et al., 2018; Li et al., 2023; Solko-Breslin et al., 2024). These frameworks support complex symbolic reasoning features like recursion and black-box functions, implement efficient differentiable reasoning algorithms, and provide bindings for deep learning frameworks like PyTorch. However, they incur significant overhead during training.

Figure 1 shows an example of a kinship reasoning task called CLUTRR whose goal is to infer the relationship between two people based on a passage describing interactions

and relationships within a family. A natural neurosymbolic formulation splits this task’s computation into a neural component, which extracts relationships from the input passage, and a symbolic component which infers new relationships to obtain the final answer. The latter involves specifying complex manipulations over symbols ( $\mathcal{L}$ ), e.g., multi-hop kinship reasoning between pairs of family members, and performing probabilistic computations ( $\mathcal{P}$ ) to track the probabilities of the symbols derived using  $\mathcal{L}$ . In general, as the complexity of the symbolic program increases, the number of possible results and their associated weights also grows exponentially, leading to a combinatorial explosion in the number of required computations. This issue is exacerbated by larger datasets usually found in deep learning tasks. Deep learning frameworks typically address this challenge by batching computations across multiple data samples.

Neurosymbolic frameworks like LYRICS (Marra et al., 2019) and Logic Tensor Networks (LTN) (Badreddine et al., 2022) also batch the computations of both  $\mathcal{L}$  and  $\mathcal{P}$  on the GPU. LTN grounds all logits and discrete symbols as tensors, and the aforementioned computations ( $\mathcal{L}$  and  $\mathcal{P}$ ) are specified in differentiable first-order logic as operations over those tensors. These programs output a value quantifying the satisfiability of model outputs with respect to logic constraints. This approach is highly performant for smaller tasks, such as MNIST SumN, where the goal is to predict the sum of  $N$  MNIST images. As we see in Table 3 in § 4.3, LTN takes around 90 seconds to converge for  $N = 5$ . However, when the complexity increases to  $N = 10$ , LTN runs out of memory on consumer-grade GPUs (here, with a capacity of 11GBs), due to the combinatorial explosion of required symbols (from  $10^5$  to  $10^{10}$ ) and their probabilities that require to be grounded on the GPU.

On the other hand, neurosymbolic frameworks like DeepProbLog (Manhaeve et al., 2018) and Scallop (Li et al., 2023) run neural models on the GPU but use a separate CPU-based backend for executing both  $\mathcal{L}$  and  $\mathcal{P}$ . This avoids issues of memory consumption on the GPU, but the lack of batched computations on CPU results in slowdowns as the problem complexity increases. We see this in MNIST SumN (Table 2). Scallop requires around 15 minutes to converge for  $N = 5$ , but needs around 1 and 2 hours to converge for  $N = 10$  and  $N = 15$ , respectively.

In this paper, we propose DOLPHIN as a solution for scaling neurosymbolic learning. In DOLPHIN, we build three key components that effectively tackle scalability challenges with existing neurosymbolic frameworks. First, we develop a unified representation that efficiently captures the relationships between neural network outputs as PyTorch tensors on GPU and associated discrete symbols as Python objects on CPU. Second, we introduce a set of primitives to enable writing symbolic manipulations that can be mapped to com-

putations over these representations, while allowing support for black-box Python functions that simplify the writing of complex symbolic programs. Third, we develop a set of vectorized *provenance semirings* (Green et al., 2007) that are easily pluggable into DOLPHIN and enable to efficiently compute symbolic gradients.

Together, these components enable DOLPHIN to construct a computation graph that integrates both neural and probabilistic computations ( $\mathcal{P}$ ), ensuring high parallelism and end-to-end differentiability on GPU. At the same time, it runs  $\mathcal{L}$  over discrete symbols on CPU, allowing flexible manipulation over arbitrary Python objects. This allows DOLPHIN to scale effectively to complex problems such as CLUTRR-N (Table 2) where  $N$  denotes the max length of the reasoning chain in the training dataset. In the case of Scallop, as the length of the reasoning chain increases, the gap between convergence times dramatically widens. In contrast, for  $N = 3$ , DOLPHIN takes around 13 minutes to converge, about 5x faster than Scallop, while for  $N = 4$ , DOLPHIN takes around 15 minutes, about 8.5x faster than Scallop. Finally, DOLPHIN is implemented as a library integrated with PyTorch, allowing users to easily incorporate it into their existing deep learning pipelines.

We evaluate DOLPHIN on a diverse set of neurosymbolic tasks involving text, image, and video, using rich reasoning features like recursion and black-box Python functions. On simpler problems, neurosymbolic programs written using DOLPHIN match the accuracy of state-of-the-art methods, while achieving these results 47x, 62x, 8x, and 1.7x faster than baselines like Scallop, sampling-based frameworks like ISED and IndeCater+, and solely GPU based methods like LTN respectively. We also observe that DOLPHIN efficiently scales to more complex benchmarks and larger datasets, achieving state-of-the-art accuracies. While baselines fail to converge on 5 out of 8 such benchmarks within 10 hours, DOLPHIN requires 5.5 hours in the worst case.

We make the following contributions in this work:

- We propose DOLPHIN, a novel neurosymbolic programming framework for end-to-end differentiable symbolic reasoning in a scalable manner (§3).
- We develop novel Pythonic abstractions and primitives to enable writing complex symbolic manipulations for neurosymbolic programs (§3.1).
- We design DOLPHIN to be extendable to new provenances and develop vectorized provenances that can be plugged into DOLPHIN for efficient computation of symbolic gradients on parallelizable hardware such as GPU (§3.2).
- We evaluate DOLPHIN on a diverse range of 13 challenging neurosymbolic tasks across different domains and show that it effectively scales with increasing problem complexity and dataset size (§4).

```

1 class SumNNet(torch.nn.Module):
2     def __init__(self):
3         super(SumNNet, self).__init__()
4         self.CNN = MNISTNet()
5
6     def forward(self, imgs):
7         d = range(10)
8         D_res = Distribution(self.CNN(imgs[0]), d)
9         for i in range(1, len(imgs)):
10             D_i = Distribution(self.CNN(imgs[i]), d)
11             D_res = apply(D_res, D_i, lambda x, y: x + y)
12         return get_logits(D_res)
    
```

Figure 2: DOLPHIN code for the MNIST SumN task.

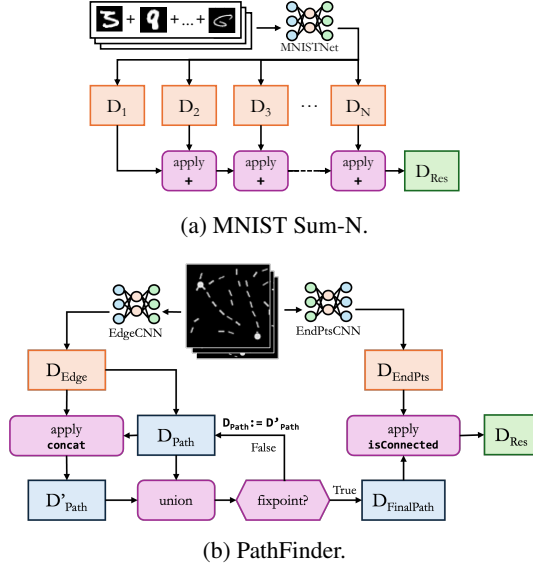


Figure 3: Computation graphs for two neurosymbolic programs written using DOLPHIN.

## 2. Overview

We illustrate DOLPHIN using the MNIST SumN task (De Smet et al., 2024), where the goal is to add  $N$  MNIST digit images. The task grows exponentially difficult, with  $10^N$  possible states and only  $9N + 1$  labels, making supervision sparse. Figure 2 shows the code for this task using DOLPHIN with PyTorch. The neural module `MNISTNet` is a PyTorch model classifying a batch of images into one of 10 classes representing the digits 0-9. This is done for each of the  $N$  batches of images in the tuple `imgs`. The logits produced by `MNISTNet`, representing probability distributions over the digits, are then passed as inputs to the symbolic program. Lines 8-11 depict a symbolic program written in Python using DOLPHIN primitives.

To support training, the symbolic program must track digit probabilities, compute probability distributions over all possibilities (0 to 9N), and propagate gradients for backpropagation. Batched computations further complicate this, making native PyTorch implementations cumbersome.

DOLPHIN abstracts symbolic computation, letting program-

mers express logic without handling underlying complexities. Lines 8 and 10 of Figure 2 show how `MNISTNet`’s output can be captured within `Distribution` objects. Each `Distribution` associates a single collection of digits with the corresponding batch of logits produced by `MNISTNet`, along with any gradients and associated metadata.

The programmer can now express the symbolic program in terms of operations over `Distributions`. For instance, in line 11, the `apply` function is used, taking two `Distributions` as arguments, along with a lambda function that specifies the addition operation. Under the hood, `apply` combinatorially explores all possible sums of the symbols from `D_res` and `D_i` and calculates their associated probabilities. The result of `apply` is a new `Distribution` over the calculated sums, and is stored back into `D_res`. This is repeated iteratively until all the outputs of the CNN are summed appropriately.

DOLPHIN provides additional primitives to support more complex symbolic programs. Figure 3b shows the computation graph for the PathFinder task (Tay et al., 2021), which involves recursively building paths to identify if two points in a maze are connected. The `union` primitive is used to support the recursive nature of this program. Since `Distribution` objects associate symbols with the batched logits themselves, probabilistic computations are vectorized and directly operate over PyTorch tensors.

This deep integration of DOLPHIN into PyTorch allows programmers to write symbolic programs as *symbolic layers* that interact with PyTorch neural layers within a neurosymbolic model. DOLPHIN can thus leverage the hardware acceleration supported by PyTorch. This contrasts with systems like Scallop, where tensors are converted into Scallop-friendly tags transferred to a process outside the Python environment with CPU-bound probability computations, restricting scalability.

## 3. The DOLPHIN Framework

We based DOLPHIN’s design on four core principles. First is *flexible programmability*, to allow writing complex symbolic manipulations ( $\mathcal{L}$ ) with Python’s rich and expressive language features. Second, probabilistic computations ( $\mathcal{P}$ ) must allow *end-to-end differentiability* on the GPU. Third, DOLPHIN must be *scalable* to tasks with large data and problem complexity. Finally, it must be *tunable*, allowing developers to define and choose provenances, treating them like deep learning hyperparameters.

Together, these principles help address the challenges of scaling neurosymbolic frameworks. Flexible programmability and tunability allow us to write complex neurosymbolic programs, while GPU differentiability and scalability work towards tackling problem and data complexity. We now describe DOLPHIN and show how we realize these principles.

Symbol	::	$s \in S$ (objects)
Tag	::	$t \in T$ (tensors)
Distribution	::	$D \in \mathbb{D} = S \rightarrow T$
APPLY	:	$\mathbb{D}^K \times (S^K \rightarrow S) \rightarrow \mathbb{D}$
FILTER	:	$\mathbb{D} \times (S \rightarrow \mathbb{B}) \rightarrow \mathbb{D}$
APPLYIF	:	$\mathbb{D}^K \times (S^K \rightarrow S) \times (S^K \rightarrow \mathbb{B}) \rightarrow \mathbb{D}$
UNION	:	$\mathbb{D} \times \mathbb{D} \rightarrow \mathbb{D}$
GETPROBS	:	$\mathbb{D} \rightarrow [0, 1]^N$

Figure 4: Formal definition of DOLPHIN’s programming abstractions (top) and primitives (bottom).

### 3.1. The DOLPHIN Syntax

To allow flexible programmability, DOLPHIN provides an interface that developers can use to express symbolic programs in a Pythonic manner.

#### 3.1.1. ABSTRACTIONS

DOLPHIN provides three main abstractions for expressing symbolic programs, shown in Figure 4. *Symbols*  $S$  represent symbolic entities relevant to the program. These entities can be any Pythonic object, such as hand-written digits in MNIST-SumN or coordinates of points in PathFinder. *Tags*  $T$  are tensors that represent their likelihoods. Typically, tags for symbols are derived from the outputs of machine learning models, such as the logits produced by the CNN classifier in MNIST-SumN. Finally, *Distribution*  $D$  maps a collection of symbols to their corresponding tags.

Distributions serve as the fundamental datatype of a DOLPHIN program and act as its main interface with a PyTorch model. As seen in the following code snippet from Figure 2:

```
D_res = Distribution(self.CNN(imgs[0]), d)
```

the logits output by the model are directly passed to the Distribution object, effectively acting as an input to the symbolic program. These logits form the batched *tags* within a Distribution object which also maintains the set of corresponding symbols  $d$ . The symbolic manipulations in a DOLPHIN program occur over the discrete symbols, while the probabilistic computations occur over tags stored as PyTorch tensors. This enables a seamless integration between the PyTorch model and the symbolic program.

This has several advantages. First, it preserves the gradients of the model output throughout the symbolic program, enabling end-to-end differentiability via PyTorch’s autograd, addressing the second core principle of DOLPHIN. Second, it allows DOLPHIN to perform operations over an entire batch of tags as per the principle of scalability, leveraging the vectorized operations provided by PyTorch. DOLPHIN can thus operate efficiently on specialized hardware like GPUs, allowing the symbolic program to scale effectively. Third, since symbol and tag computations are effectively

decoupled, operations over symbols can be run on the CPU, allowing the support for arbitrary Python objects and functions in the symbolic program, even while the tag computations are performed on GPU.

**Example 3.1.** Consider MNIST images  $I_1, I_2$  for the SumN task discussed in Section 2. Let  $f_{CNN}$  be the neural model which classifies image  $I_j$  into one of 10 classes representing digits 0 to 9. Let  $f_{CNN}(I_1) = \{0.00, 0.90, \dots, 0.01\}$  and  $f_{CNN}(I_2) = \{0.78, 0.09, \dots, 0.00\}$ . We thus define the following Distributions:

1.  $D_1 = \{0 \rightarrow 0.00, 1 \rightarrow 0.90, \dots, 9 \rightarrow 0.01\}$
2.  $D_2 = \{0 \rightarrow 0.78, 1 \rightarrow 0.09, \dots, 9 \rightarrow 0.01\}$

#### 3.1.2. OPERATIONS

DOLPHIN provides five operations to allow the expression of complex neurosymbolic programs in conjunction with user-defined functions, shown in Figure 4.

**APPLY.** This is the primary operation that can be used to manipulate Distributions. It takes as inputs  $K \geq 1$  Distributions, along with a function  $f$  of the same arity. This function defines operations over the symbols of  $K$  distributions. APPLY then computes the results of  $f$  over all possible combinations of arguments sourced from the symbols of the Distributions as well as their associated tags, and returns a new Distribution with these results and tags. This occurs in two stages akin to the popular map-reduce pattern. In the *map* stage, APPLY computes the results of  $f$  over the symbols of the input Distributions and conjuncts their tags:

$$R = \{ (f(s_1, s_2, \dots, s_k), (t_1 \otimes t_2 \otimes \dots \otimes t_k)) \mid D_i(s_i) = t_i, i = 1, \dots, k \} \quad (1)$$

Here, the tag of each result symbol  $f(s_1, s_2, \dots, s_k)$  is the conjunction  $\otimes$  of the tags  $(t_1, t_2, \dots, t_k)$  of the input symbols it was derived from. The function  $f$  is executed sequentially on the CPU for each combination of symbols as function  $f$  can be any user-defined Python function, including complex control flows and operations like regex parsing, image processing, or Python’s `eval()`. It may also be a many-to-one function and the tags shared by a resulting symbol must be aggregated to form the final tags of the output Distribution. We, therefore, *shuffle* the results from the map stage to compute a function  $M$  from each symbol to tags from  $R$  associated with it:

$$M = \lambda s . \{ t \mid (s, t) \in R \} \quad (2)$$

We then proceed to the *reduce* stage, where we aggregate the tags of each symbol in  $M$  using disjunction  $\oplus$  to produce the final Distribution  $D_{res}$ :

$$D_{res} = \lambda s . \bigoplus \{ t \mid t \in M(s) \} \quad (3)$$

Since the tags here are PyTorch tensors representing probabilities, the implementations of the conjunction and disjunction operations are dictated by the underlying provenance specified by the program, detailed in Section 3.2.



**Example 3.2.** Continuing from Example 3.1, let function  $f(x, y) = x + y$  be applied to  $D_1$  and  $D_2$  to produce a new Distribution  $D' = \text{apply}(D_1, D_2, f)$ .  $D'$  thus represents a Distribution over the sum of the symbols from  $D_1$  and  $D_2$ :

$$D' = \{0 \rightarrow 0.00, 1 \rightarrow 0.70, \dots, 18 \rightarrow 0.00\}$$

Consider the tag of  $D'(1)$ . 1 can be a result of  $D_1$  being 0 and  $D_2$  being 1, or of  $D_1$  being 1 and  $D_2$  being 0:

$$\begin{aligned} D'(1) &= (D_1(0) \otimes D_2(1)) \oplus (D_1(1) \otimes D_2(0)) \\ &= (0.00 \otimes 0.09) \oplus (0.90 \otimes 0.78) \end{aligned}$$

This expression’s value (e.g. 0.70) depends on the provenance specified (e.g. DAMP), discussed in Section 3.2.

**FILTER.** The FILTER operation is used to filter out symbols from a Distribution. It takes in a single Distribution, along with a user-defined function that returns a boolean value. This operation then returns a new Distribution that contains only symbols that satisfy the condition with their tags.

**Example 3.3.** Continuing from Example 3.1, assume we want a Distribution over just the even symbols of  $D_1$ . We can consider a filtering function  $f(x) = (x \bmod 2 == 0)$ . The resulting Distribution will have all the odd-numbered symbols completely removed:

$$\begin{aligned} D' &= \text{filter}(D_1, f) \\ &= \{0 \rightarrow 0.00, 2 \rightarrow 0.02, \dots, 8 \rightarrow 0.01\} \end{aligned}$$

**APPLYIF.** This operation is a conditional version of APPLY. It takes in  $K$  Distributions and functions  $f_{\text{apply}}$  and  $f_{\text{cond}}$  of the same arity. For each combination of symbols from the  $K$  Distributions, APPLYIF computes  $f_{\text{apply}}$  and its associated tags only if the condition  $f_{\text{cond}}$  is satisfied over that combination of symbols. The operation then returns a new Distribution with these results and tags.

**UNION.** The UNION operation takes in two Distributions and returns a new Distribution containing the union of the input symbols, along with their tags. Any symbols common to both input Distributions have their tags disjuncted.

**Example 3.4.** Consider Distributions  $D_1 = \{0 \rightarrow 0.01, 1 \rightarrow 0.24\}$  and  $D_2 = \{0 \rightarrow 0.63, 4 \rightarrow 0.37\}$ . The union will be:

$$\text{union}(D_1, D_2) = \{0 \rightarrow 0.64, 1 \rightarrow 0.24, 4 \rightarrow 0.37\}$$

**GETPROBS.** The GETPROBS operation extracts the probabilities from the tags of a Distribution. This is used mainly once the symbolic program has been executed to extract the final probabilities of the symbols in the output Distribution. These probabilities can then be used to compute the loss function for training the neural model.

### 3.1.3. WRITING COMPLEX DOLPHIN PROGRAMS

Some neurosymbolic tasks require the writing of programs containing complex control flows and recursion (e.g. the PathFinder task, § 4.1). This can be done in one of two ways. The simplest way is to specify any control flow operations within the user-defined functions supplied to the DOLPHIN operations. Alternatively, one can specify branches of control flow separately and merge their results via UNION, as shown in Figure 6 (Appendix B for more details).

In some cases, even though DOLPHIN limits the effect of combinatorial explosion in terms of efficiency and memory usage, the number of combinations may still be excessive. In such cases, DOLPHIN allows developers to sample subsets of symbols from Distribution objects specified within the symbolic program, effectively limiting the number of symbols processed in each operation.

## 3.2. DOLPHIN Provenances

The DOLPHIN primitives discussed above define how to conjunct or disjunct tags corresponding to the symbol manipulations, e.g. Equations (1) and (3). These tag operations are achieved by using a mathematical framework called *provenance semirings* (Green et al., 2007). Provenance semirings provide generalized algebraic structure to propagate probabilities over tagged data.

Designing and implementing provenances can be challenging since they must be accurate enough to capture the semantics of the symbolic program, while at the same time being coarse enough to maintain computational feasibility. Furthermore, the provenances must be differentiable.

While neurosymbolic frameworks like Scallop (Li et al., 2023) implement differentiable provenances, they are not designed to leverage hardware accelerations or batched optimizations due to the CPU-bound nature of their implementations. Frameworks like LTN use t-norms that are more amenable to vectorization, but lack support for more complex provenances such as Differentiable Top- $k$  Proofs (DTKP) (Huang et al., 2021). We thus design differentiable, vectorized provenances in DOLPHIN to enable GPU support.

We simplify the definition of provenances as a 5-tuple:  $(T, \mathbf{0}, \mathbf{1}, \otimes, \oplus)$ . Here,  $T$  is the tag space,  $\otimes : T \times T \rightarrow T$  is the conjunction operator with identity  $\mathbf{0}$ , and  $\oplus : T \times T \rightarrow T$  is the disjunction operator with identity  $\mathbf{1}$ . We then implement two differentiable provenances in DOLPHIN: Differentiable Add-Mult Probabilities (DAMP) and Differentiable Top-K Proofs (DTKP). Table 1 summarizes the operations of these provenances. While building the neurosymbolic program, the developer may specify which provenance to use, satisfying the core principle of tunability.

**Differentiable Add-Mult Probabilities.** Differentiable

Table 1: DOLPHIN provenances implemented in PyTorch.

Provenance	Domain	0	1	$t \oplus t'$	$t \otimes t'$
DAMP	$[0, 1]$	0	1	$\text{clamp}_0^1(t + t')$	$t \cdot t'$
DTKP-AM	$[0, 1] \cup \{\infty, -\infty\}$	$\hat{\mathbf{0}}_{ij} = -\infty$	$\hat{\mathbf{1}}_{ij} = \begin{cases} \infty & i = 1 \\ -\infty & i > 1 \end{cases}$	$\text{top}_k(\text{cat}(t, t'))$	$\text{top}_k([\min( t_i ,  t'_j ) \mid (t_i, t'_j) \in t \times t'])$

Add-Mult Probabilities (DAMP) is a popular technique that uses the probability space as its tag space:  $T = [0, 1]$ . Its conjunction operation  $\otimes$  is defined as the product of probabilities, clamped at 1, and its disjunction operation  $\oplus$  is defined as the sum of probabilities. The main assumption underlying the DAMP operations is that the input Distributions are mutually exclusive and independent. This assumption allows DAMP to compute probabilities extremely efficiently, as the operations are simple and can be easily vectorized.

**Differentiable Top- $k$  Proofs.** Differentiable Top- $k$  Proofs (DTKP) (Huang et al., 2021) was proposed to overcome the shortcomings of DAMP. This provenance tracks a set of up to  $k$  proofs for each symbol. Each proof denotes the set of input symbols necessary to derive the output symbol. These proofs are then used to compute the probabilities of the output symbols. In Scallop, DTKP tags are converted into probabilities via differentiable weighted model counting (WMC). This form of DTKP, which we call DTKP-WMC, is computationally hard and is by nature difficult to vectorize due to the varying sizes of proof sets and the WMC procedure. We hence design a novel vectorized approximation of DTKP-WMC; we term DTKP-AM (DTKP with Add-Mult), that can be efficiently computed on the GPU.

We first define the structure of tags in DTKP-AM to conform to the constraints of PyTorch tensors. Each tag  $t$  for a symbol  $s$  is a 2-dimensional tensor of shape  $(k, |I|)$ , where  $k$  is the maximum number of proofs to be retained and  $I$  is an ordered list of all *input symbols* (symbols that are present in the input Distributions). Each row  $t_i$  of  $t$  corresponds to one of the tag’s  $k$  proofs. Each element  $t_{ij}$  thus represents the probability of the  $j$ th input symbol in the  $i$ th proof:

$$t_{ij} = \begin{cases} p_j & \text{if the } j\text{th symbol is present in the } i\text{th proof} \\ \hat{\mathbf{0}}_{ij} & \text{otherwise} \end{cases}$$

where  $p_j$  is the probability of the  $j$ th input symbol. The probability of each proof is then computed by taking the product of the normal:

$$\Pr(t_i) = \prod_j \text{norm}(t_{ij}), \text{ where } \text{norm}(t_{ij}) = \begin{cases} 1 & t_{ij} = +\infty \\ 0 & t_{ij} = -\infty \\ t_{ij} & \text{otherwise} \end{cases}$$

We next define the operations of DTKP-AM in Table 1. The  $\oplus$  operation is defined as the union of two tag tensors  $t$  and  $t'$  while  $\otimes$  is defined as the element-wise minimum of the normalized elements of all possible combinations of proofs

in  $t$  and  $t'$ . In each case, the  $\text{top}_k$  operation retains only up to  $k$  proofs with the highest probabilities.

These definitions thus allow us to take advantage of the benefits of the DTKP provenance while enabling efficient computation on the GPU. To calculate the probability of the entire tag, DTKP-AM adds the probabilities of the individual proofs and clamps it at 1. We provide a detailed discussion of DTKP-AM in Appendix A.

### 3.3. Building the DOLPHIN Program

The programmer specifies the neurosymbolic task using a Python program  $P$ , which integrates neural components with symbolic operations via DOLPHIN’s interface. Given a dataset  $\mathcal{D}$  and one or more neural networks  $M_1, \dots, M_k$ , DOLPHIN constructs a computation graph where symbolic transformations occur on the CPU, and probabilistic computations, including neural network inference, are efficiently executed on the GPU. All computations leverage distribution objects  $D_i$ , enabling end-to-end differentiability and scalability. Training optimizes the objective function

$$\phi(\theta) = \min_{\theta} \sum_{(x,y) \in \mathcal{D}} \mathbb{L}(P(M_{\theta}(x)), y),$$

where  $\mathbb{L}$  is the loss function (e.g., binary cross entropy).

## 4. Experiments

We evaluate DOLPHIN on a set of 13 benchmarks of varying complexity and scale across 5 neurosymbolic tasks. Our evaluation addresses the following research questions:

- **RQ1: Scalability.** Can DOLPHIN scale to tasks and datasets beyond the scope of existing SOTA frameworks?
- **RQ2: Accuracy.** Do models written in DOLPHIN converge to SOTA accuracies in less training time?
- **RQ3: Provenance Comparisons.** Which provenances are most effective for each benchmark?

### 4.1. Benchmarks

We describe the benchmarks used to evaluate DOLPHIN and give additional information about the experiment setup and DOLPHIN code for each benchmark in Appendix D.

**MNIST SumN.** The MNIST SumN (or briefly, SumN) task from (De Smet et al., 2024) takes as inputs  $N$  handwritten

digits from the MNIST dataset and returns their sum. We consider three versions of this task: **SumN-5** ( $N = 5$ ), **SumN-10** ( $N = 10$ ), and **SumN-15** ( $N = 15$ ).

**Hand-Written Formula (HWF).** The HWF task from (Li et al., 2020) takes as input a set of images of handwritten digits and arithmetic operators representing a formula. The task is to evaluate the formula and return the result. We consider three versions based on formula length: **HWF-7** (up to 7), **HWF-15** (up to 15), and **HWF-19** (up to 19).

**PathFinder.** PathFinder (or Path) (Tay et al., 2021) tests the ability of an agent to reason over long-range dependencies within an image of two dots and a sequence of curved and dashed lines. The task is to identify whether the two dots are connected via the lines. We consider three versions based on the image size in pixels: **Path-32** ( $32 \times 32$ ), **Path-128** ( $128 \times 128$ ), and **Path-256** ( $256 \times 256$ ).

**CLUTRR.** In this task from (Sinha et al., 2019), given some text containing information about several individuals and some of their relationships, the model must infer the relationship between two given individuals, which is not explicitly provided in the input. We consider two versions, where the training data contains relation chains of lengths up to 3 (**CLUTRR-3**) or 4 (**CLUTRR-4**).

**Mugen.** In this task from (Hayes et al., 2022), given a 3.2 second long video of gameplay footage and text captioning the video, the goal is to measure how aligned the text is with the video. There are two variants: Mugen-TV, where the model retrieves the video that best aligns with the text, and Mugen-VTR, where the model retrieves the text that best aligns with the video. We consider two versions of this task: **1K** and **5K** comprising 1000 and 5000 training samples.

## 4.2. Experimental Setup and Baselines

**Setup.** All experiments, except CLUTRR, were run on machines with two 20-core Intel Xeon Gold 6248 CPUs, four NVIDIA GeForce RTX 2080 Ti (11 GB) GPUs, and 768 GB RAM. Since CLUTRR demands more GPU memory due to running the RoBERTa model with a standard batch size of 16, all programs for this benchmark were run with a NVIDIA A100 40GB GPU. We ran each tool thrice until convergence or until a soft timeout of 10 hours was reached and report the average best accuracy and training time. For HWF/MNIST, we use the same CNN architecture as Scallop (Appendix D). For CLUTRR, we use Scallop’s Roberta configuration: a pretrained model (roberta-base) finetuned while training the classification head.

**Baselines.** We select Scallop (Li et al., 2023), a contemporary state-of-the-art neurosymbolic framework supporting differentiable programming optimized to run on the CPU in parallel using multiple cores. We also choose two sampling-based gradient approximation methods, ISED (Solko-

Table 2: Training times (in seconds) for DOLPHIN and Scallop on all benchmarks. Training times more than 10 hours are highlighted in red. The scaling factor  $\alpha$  is the ratio of the total training times of Scallop to DOLPHIN.

Task	DOLPHIN	Scallop	
	$T_{\text{total}}$	$T_{\text{total}}$	$\alpha$
SumN-5	53.86	923.78	17.15
SumN-10	104.91	3.42e3	32.56
SumN-15	157.05	7.41e3	47.18
HWF-7	2.45e3	9.99e3	4.08
HWF-15	9.78e3	<b>1.66e5</b>	16.97
HWF-19	1.63e4	<b>1.82e5</b>	11.16
Path-32	1.29e4	2.2e4	1.71
Path-128	1.67e4	<b>4.17e4</b>	2.49
Path-256	1.97e4	<b>1.14e5</b>	5.78
CLUTRR-3	807.12	4.29e3	5.32
CLUTRR-4	923.86	7.83e3	8.48
Mugen-1K	2.39e3	6.71e3	2.81
Mugen-5K	1.15e4	3.59e4	3.12

Breslin et al., 2024) and IndeCateR+ (De Smet et al., 2024). We also include Logic Tensor Networks (LTN) (Serafini & Garcez, 2016), which combines first-order logic with continuous optimization by compiling logical constraints into a computation graph on the GPU. We compare DOLPHIN against Scallop on all benchmarks, and against ISED and IndeCateR+ on SumN and HWF. We compare SumN with LTN, but were unable to write HWF in LTN (explained in Appendix H.1). Since LTN also runs out of memory for a simpler benchmark like SumN-10, we do not compare against it for the more complex benchmarks of Path, CLUTRR, and Mugen. Similarly, we do not evaluate ISED and IndeCateR+ on these benchmarks, as ISED already fails to scale for simpler tasks like SumN-10 and HWF-7, while IndeCateR+ does not support their recursive structures.

## 4.3. RQ1: Scalability

Table 2 presents the total training times ( $T_{\text{total}}$ ) till convergence in seconds for DOLPHIN and Scallop across all benchmarks, alongside the scaling factor  $\alpha$  (the ratio of the total training times of the baselines to DOLPHIN). Table 3 shows the same for the remaining baselines over SumN and HWF. We set a soft timeout of 10 hours, though we still report the training times that run over highlighted in red. The results demonstrate that DOLPHIN advances the state-of-the-art in neurosymbolic learning by scaling to more complex problems, e.g., larger versions of HWF, Path, and Mugen, that are beyond the reach of the other baselines which time out within 10 hours. Even for the other benchmarks where baselines do not time out, DOLPHIN achieves a scalability factor of up to 47x, 62x, and 3.49x against Scallop, ISED, and IndeCateR+ with an average speed up of 13.95x across all baselines for all benchmarks.

Table 3: Training time (in seconds) for DOLPHIN, LTN, ISED, and IndeCateR+ on SumN and HWF. Training times more than 10 hours are highlighted in red.  $\alpha$  is the ratio of training times of the baselines to DOLPHIN. OOM occurred on an NVIDIA GeForce RTX 2080 Ti (11 GB).

Task	DOLPHIN	LTN		ISED		IndeCateR+	
	$T_{\text{total}}$	$T_{\text{total}}$	$\alpha$	$T_{\text{total}}$	$\alpha$	$T_{\text{total}}$	$\alpha$
SumN-5	53.86	92.54	1.72	299.63	5.56	416.78	7.74
SumN-10	104.91	<b>OOM</b>	-	2.17e3	20.64	385.65	3.68
SumN-15	157.05	<b>OOM</b>	-	9.8e3	62.41	548.28	3.49
HWF-7	2.45e3	N.A.		1.58e4	6.45	1.35e4	5.51
HWF-15	9.78e3			<b>5.99e4</b>	6.13	2.51e4	2.57
HWF-19	1.63e4			<b>9.34e4</b>	5.73	<b>6.27e4</b>	3.85

We also use the training times per epoch to calculate a scalability factor  $\alpha_{\text{epoch}}$  (Table 5, Appendix D). We see that among cases where baselines timeout,  $\alpha_{\text{epoch}}$  is up to 280x for HWF-19 and is 40.6x faster on average. This results in DOLPHIN effectively training for more epochs in less time compared with the other baselines, which also allows it to converge to higher accuracies, as we see in Figure 5. We expand on these results in the next RQ.

#### 4.4. RQ2: Accuracy

Figure 5 presents the accuracy of DOLPHIN and the baselines on the different benchmarks trained for up to 10 hours. DOLPHIN accuracies are marked in blue. In all cases, for DOLPHIN, we report the accuracies of the best-performing provenance. We use the DAMP provenance for MNIST, CLUTRR, and Mugen benchmarks, and the DTKP-AM provenance for the HWF and PathFinder benchmarks.

We observe that in all cases, DOLPHIN achieves state-of-the-art accuracy among general-purpose neurosymbolic frameworks, except in CLUTRR, where DOLPHIN’s accuracy is slightly lower than Scallop’s. As we scale up to larger versions of the benchmarks, DOLPHIN achieves better accuracy, because the baselines either report lower accuracy due to the complexity of the benchmark (e.g., black-box sampling techniques such as ISED on HWF) or fail to converge within 10 hours due to slower per-epoch train time (e.g., Scallop on PathFinder-256, IndeCateR+ on HWF-19). As a result, DOLPHIN significantly outperforms the second-best benchmark on the largest dataset versions, achieving up to a 20% gain on HWF and 33% on Path. In some cases, given no timeout, Scallop and IndeCateR+ eventually converge to accuracies comparable to DOLPHIN, as we show in Appendix D.2. However, doing so requires significantly more training time, as discussed earlier in Section 4.3.

These results show that not only do DOLPHIN’s scalability improvements not come at the cost of accuracy, but DOLPHIN enables SOTA accuracy when previously unattainable.

#### 4.5. RQ3: Provenance Comparisons

We perform ablation studies to compare the effectiveness of the DAMP and DTKP-AM provenances for each benchmark. We share the graphs in Figure 9 (Appendix E). In all cases, training with the DAMP provenance takes around 24.19 seconds per epoch less than with DTKP-AM on average.

However, the effectiveness of each provenance varies across benchmarks. For all variations of Path, CLUTRR, and Mugen, both provenances achieve comparable accuracies, with DTKP-AM having a slight edge. For SumN, DAMP provenance is more effective than the DTKP-AM by 72% points on average, since the top-k proofs cannot capture all the possible ways in which sums of digits can be computed.

In contrast, for HWF, DTKP-AM is more effective than DAMP by an average of 42.2% points. Each step of the HWF program, shown in Appendix H, involves both a concatenation operation and a partial parsing operation before the final expression is evaluated to produce a result. As such, it is difficult for the tags in DAMP to capture the semantics of the symbolic program. In the case of DTKP-AM, each tag is a collection of proofs over input symbols corresponding to logits derived from the neural model. Therefore, any calculated gradients can be directly backpropagated to the logits that most influenced the output, making this a more effective provenance for this task.

### 5. Discussion

Dolphin can be used for any task where the output of a model can be cast as a distribution over probabilities. This abstraction naturally encompasses a wide range of discriminative models in machine learning, such as classifiers, structured prediction systems, and vision-language models. By associating each symbolic object with a probability distribution over its possible values (tags), Dolphin enables downstream symbolic reasoning over uncertain predictions made by neural networks.

For instance, consider an autonomous driving scenario where a standard object detector such as Faster R-CNN



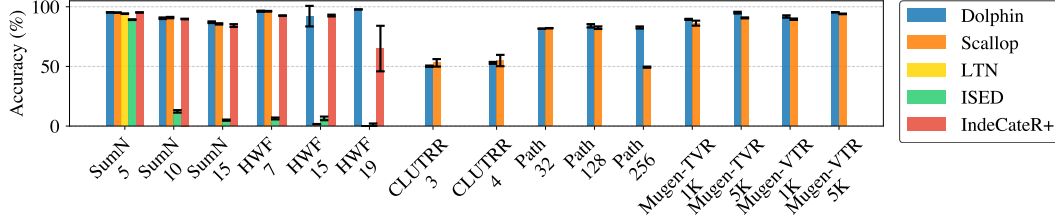


Figure 5: Accuracy of DOLPHIN and baselines trained for up to 10 hours across all benchmarks.

outputs bounding boxes and class probabilities. One can represent each detection as a symbolic object whose attributes include coordinates and a Distribution over class probabilities:

```
1 CLASSES = ['car', 'person', 'truck', ...]
2
3 class DetectedObject:
4     def __init__(self, coords, score, class_logits):
5         self.coords = coords
6         self.score = score
7         self.distr = Distribution(CLASSES, class_logits)
```

Now, consider checking whether a detected person is inside a detected car using symbolic reasoning:

```
1 def is_inside(coord_a, coord_b):
2     ...
3
4     person_inside_car = apply(
5         o1.distr, o2.distr, lambda c1, c2:
6         c1 == "person" and c2 == "car" and
7         is_inside(o1.coords, o2.coords)
8     )
```

This computation yields a Distribution over True and False, representing the likelihood that two detected objects constitute a person inside a car. This paradigm can generalize to many domains where neural predictions must be interpreted symbolically—for example, relationship inference in vision scenes, symbolic post-processing over structured outputs, or probabilistic parsing in NLP. Dolphin thus offers a principled and composable way to integrate symbolic logic over deep model outputs, opening the door for broader real-world applicability.

## 6. Related Work

**Neurosymbolic programming frameworks.** Apart from existing frameworks like Scallop (Li et al., 2023), Deep-ProbLog (Manhaeve et al., 2018), and ISED (Solko-Breslin et al., 2024), there exist domain-specific tools such as NeurASP (Yang et al., 2021) for phrase alignment. These approaches often suffer from inefficiencies due to CPU-bound symbolic reasoning.

**Scaling techniques.** Various methods exist to scale differentiable reasoning. LYRICS (Marra et al., 2019), Logic Tensor Networks (Badreddine et al., 2022), and Tensorlog (Cohen et al., 2020) compile first-order logic constraints into GPU-compatible computation graphs. Other techniques, such as Greedy NTP (Minervini et al., 2020a) and the conditional theorem prover (Minervini et al., 2020b), optimize

proof search using heuristics. SLASH (Skryagin et al., 2024) integrates neural networks and probabilistic circuits with Answer Set Programming, achieving scalability by dynamically pruning stochastically insignificant parts of programs. A-NESI (van Krieken et al., 2023) uses learned neural models to approximate the exact probabilistic semantics of WMC, boosting scalability. However, these methods are often task-specific and lack generalizability to broader neurosymbolic learning, or they struggle to scale due to memory constraints when grounding symbolic computations on GPU.

**Specialized neurosymbolic solutions.** There are many specialized solutions for various neurosymbolic tasks. For instance, NGS (Li et al., 2020) uses a hand-coded syntax to specify the structure of mathematical expressions for HWF. More general solutions, such as NS-CL (Mao et al., 2019) includes a framework for visual question answering that learns symbolic representations for text and images. NeRd (Chen et al., 2021) transforms questions in natural language into executable programs based on symbolic information extracted from text. (Orvieto et al., 2023) proposes a recurrent neural network architecture that achieves 95% accuracy on Path-32 and 94% on Path-128. In contrast, DOLPHIN is a general system that tries to scale diverse neurosymbolic programs.

## 7. Conclusion and Limitations

We proposed DOLPHIN, a framework for scaling neurosymbolic learning. DOLPHIN provides abstractions for writing symbolic programs along with pluggable vectorized provenances to compute symbolic gradients. This allows users to write differentiable symbolic programs in Python within PyTorch pipelines that can scale to complex programs and large datasets. We show that DOLPHIN scales significantly better than existing neurosymbolic frameworks while achieving state-of-the-art performance on a variety of tasks.

A limitation of DOLPHIN is that it needs the user to write programs in a batched manner. This is a common pattern within deep learning but may be restrictive to users new to batched programming. Also, while DOLPHIN works well with most models, the representation needed by generative models (e.g., Causal LLMs) has not been investigated yet. DOLPHIN also lacks support for non-deterministic symbolic programs. We leave these for future work.

## Acknowledgements

We thank the reviewers for their insightful feedback that helped to improve this paper. This research was supported by the ARPA-H program on Safe and Explainable AI under the award D24AC00253-00, the NSF award #2313010, and a Google PhD Fellowship.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## References

- Badreddine, S., d’Avila Garcez, A., Serafini, L., and Spranger, M. Logic tensor networks. *Artificial Intelligence*, 303:103649, 2022. ISSN 0004-3702. doi: <https://doi.org/10.1016/j.artint.2021.103649>. URL <https://www.sciencedirect.com/science/article/pii/S0004370221002009>.
- Chaudhuri, S., Ellis, K., Polozov, O., Singh, R., Solar-Lezama, A., Yue, Y., et al. Neurosymbolic programming. *Foundations and Trends® in Programming Languages*, 7(3):158–243, 2021.
- Chen, Z., Gao, Q., and Moss, L. S. NeuralLog: Natural language inference with joint neural and logical reasoning. In Ku, L.-W., Nastase, V., and Vulić, I. (eds.), *Proceedings of \*SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pp. 78–88, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.starsem-1.7. URL <https://aclanthology.org/2021.starsem-1.7>.
- Cohen, W. W., Yang, F., and Mazaitis, K. Tensorlog: A probabilistic database implemented using deep-learning infrastructure. *J. Artif. Intell. Res.*, 67:285–325, 2020. URL <https://api.semanticscholar.org/CorpusID:211263674>.
- Dang, M., Khosravi, P., Liang, Y., Vergari, A., and den Broeck, G. V. Juice: A julia package for logic and probabilistic circuits. In *AAAI Conference on Artificial Intelligence*, 2021. URL <https://api.semanticscholar.org/CorpusID:235363700>.
- Darwiche, A. An advance on variable elimination with applications to tensor-based computation. In *ECAI 2020*, pp. 2559–2568. IOS Press, 2020.
- De Smet, L., Sansone, E., and Zuidberg Dos Martires, P. Differentiable sampling of categorical distributions using the catlog-derivative trick. *Advances in Neural Information Processing Systems*, 36, 2024.
- Green, T. J., Karvounarakis, G., and Tannen, V. Provenance semirings. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 31–40, 2007.
- Hayes, T., Zhang, S., Yin, X., Pang, G., Sheng, S., Yang, H., Ge, S., Hu, Q., and Parikh, D. Mugen: A playground for video-audio-text multimodal understanding and generation. In *European Conference on Computer Vision*, pp. 431–449. Springer, 2022.
- Huang, J., Li, Z., Chen, B., Samel, K., Naik, M., Song, L., and Si, X. Scallop: From probabilistic deductive databases to scalable differentiable reasoning. *Advances in Neural Information Processing Systems*, 34:25134–25145, 2021.
- Kambhampati, S., Valmeekam, K., Guan, L., Verma, M., Stechly, K., Bhambri, S., Saldyt, L. P., and Murthy, A. B. Position: Llms can’t plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning*.
- Li, Q., Huang, S., Hong, Y., Chen, Y., Wu, Y. N., and Zhu, S.-C. Closed loop neural-symbolic learning via integrating neural perception, grammar parsing, and symbolic reasoning. In *International Conference on Machine Learning*, pp. 5884–5894. PMLR, 2020.
- Li, Z., Huang, J., and Naik, M. Scallop: A language for neurosymbolic programming. *Proceedings of the ACM on Programming Languages*, 7(PLDI):1463–1487, 2023.
- Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., and De Raedt, L. Deepproblog: Neural probabilistic logic programming. *Advances in neural information processing systems*, 31, 2018.
- Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., and Wu, J. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=rJgMlhRctm>.
- Marra, G., Giannini, F., Diligenti, M., and Gori, M. Lyrics: A general interface layer to integrate logic inference and deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pp. 283–298, Berlin, Heidelberg, 2019. Springer-Verlag. ISBN 978-3-030-46146-1.

- doi: 10.1007/978-3-030-46147-8\_17. URL [https://doi.org/10.1007/978-3-030-46147-8\\_17](https://doi.org/10.1007/978-3-030-46147-8_17).
- Minervini, P., Bošnjak, M., Rocktäschel, T., Riedel, S., and Grefenstette, E. Differentiable reasoning on large knowledge bases and natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5182–5190, 2020a.
- Minervini, P., Riedel, S., Stenetorp, P., Grefenstette, E., and Rocktäschel, T. Learning reasoning strategies in end-to-end differentiable proving. In *International Conference on Machine Learning*, pp. 6938–6949. PMLR, 2020b.
- Naik, A., Stein, A., Wu, Y., Naik, M., and Wong, E. Torchql: A programming framework for integrity constraints in machine learning. *Proc. ACM Program. Lang.*, 8(OOPSLA1), April 2024. doi: 10.1145/3649841. URL <https://doi.org/10.1145/3649841>.
- Orvieto, A., Smith, S. L., Gu, A., Fernando, A., Gulcehre, C., Pascanu, R., and De, S. Resurrecting recurrent neural networks for long sequences. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org, 2023.
- Sanh, V. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- Scallop Language Group. Scallop and neuro-symbolic programming: Tags, instrumentation, and provenance. Eleventh Summer School on Formal Techniques, 2022. URL <https://www.scallop-lang.org/ssft22/lectures/lecture-2.pdf>.
- Serafini, L. and Garcez, A. d. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. *arXiv preprint arXiv:1606.04422*, 2016.
- Sinha, K., Sodhani, S., Dong, J., Pineau, J., and Hamilton, W. L. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In Inui, K., Jiang, J., Ng, V., and 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. 4506–4515, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1458. URL <https://aclanthology.org/D19-1458>.
- Skryagin, A., Ochs, D., Dhami, D. S., and Kersting, K. Scalable neural-probabilistic answer set programming. *J. Artif. Int. Res.*, 78, January 2024. ISSN 1076-9757. doi: 10.1613/jair.1.15027. URL <https://doi.org/10.1613/jair.1.15027>.
- Solko-Breslin, A., Choi, S., Li, Z., Velingker, N., Alur, R., Naik, M., and Wong, E. Data-efficient learning with neural programs. *arXiv preprint arXiv:2406.06246*, 2024.
- Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., and Metzler, D. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=qVyeW-grC2k>.
- van Krieken, E., Thanapalasingam, T., Tomczak, J., van Harmelen, F., and Ten Teije, A. A-nesi: A scalable approximate method for probabilistic neurosymbolic inference. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 24586–24609. Curran Associates, Inc., 2023.
- Yang, Z., Ishay, A., and Lee, J. Neurasp: embracing neural networks into answer set programming. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI’20*, 2021. ISBN 9780999241165.

## A. DTKP-AM Provenance

We clarify and expand on some aspects of the DTKP-AM provenance.

### A.1. WMC approximation

In this section, we emphasize that DTKP-AM does not perform precise weighted model counting (WMC) and address possible shortcomings that could arise. A hardware-efficient vectorization of exact WMC is beyond the scope of this paper, and is itself an active area of research. Instead, we use the following add-mult approximation of WMC:

$$\Pr(t) = \sum_i \Pr(t_i) = \sum_i \prod_j \text{norm}(t_{ij})$$

We note that this approximation upper bounds the result from DTKP-WMC: the coarseness arises from the summation, which may double count models that satisfy more than one of the proofs. However, add-mult achieves significant computational speedup since it simplifies the exponential enumeration over all possible models into a linear pass over the tag’s elements.

We further claim that this approximation does not destroy *all* the semantics from DTKP-WMC due to DTKP-AM’s faithful implementation of the semiring operations  $\oplus$  and  $\otimes$  for tracking top-k proofs. DTKP-AM tags therefore remain similar to DTKP-WMC tags at every intermediate symbolic reasoning step. By contrast, the imprecise add-mult is a one-time transformation of the final tags into probabilities, performed only after the tags have been propagated through the entire symbolic program. Crucially, we show there exists information that is uniquely captured by top-k tag operations, and is not lost when fuzzily converting the tags to probabilities.

As a simple illustrative example, consider using APPLY with the following toy function:

$$f(a, b) = \begin{cases} \mathbb{T} & a = b \\ \mathbb{F} & \text{otherwise} \end{cases}$$

For any distribution  $D$  of mutually exclusive input symbols (e.g. the digit classification of a CNN), we intuitively would like the distribution  $f(D, D)$  to assign a probability of 1 to symbol  $\mathbb{T}$  and a probability of 0 to symbol  $\mathbb{F}$ . According to our semantics, the tag for  $\mathbb{T}$  is actually given by:

$$f(D, D)(\mathbb{T}) = \bigoplus_i (D(i) \otimes D(i))$$

However, if we were to use DAMP to compute the tags for  $f(D, D)$ , the provenance treats the two input distributions as independent when they are the exact same distribution! Thus, the probability assigned to  $\mathbb{T}$  by  $f(D, D)$  is incorrectly calculated as:

$$f(D, D)(\mathbb{T}) = \sum_i (D(i))^2$$

On the other hand, consider any top-k provenance that satisfies:

$$t \otimes t' = \text{top}_k(\{t_i \cup t'_j \mid (t_i, t'_j) \in t \times t'\})$$

where  $\times$  is the set Cartesian product. Note that DTKP-AM does satisfy this condition, where the set union is implemented with an element-wise minimum. Now assuming  $D(i)$  is initialized in the natural way (i.e. a tag consisting of a single proof containing just the input symbol  $i$ ), then  $D(i) \otimes D(i) = D(i)$  and therefore:

$$f(D, D)(\mathbb{T}) = \bigoplus_i D(i)$$

Under both DTKP-WMC and DTKP-AM, the probability of  $\mathbb{T}$  is:

$$\sum_i \prod_j \text{norm}(D(i)_{ij}) = \sum_i D(i)_{ii} = \sum_i \Pr(i) = 1$$

for any normalized  $D$  with at most  $k$  symbols. Even if the number of symbols exceeds  $k$ , we note that the distributions we seek to learn are often skewed (an accurate model should assign a probability to the ground truth that significantly outweighs



the other symbols). For such distributions, DTKP and DTKP-AM would still yield the same probability for  $\mathbb{T}$ , and it is much closer to 1 than the sum of squares result from DAMP.

While this example may seem contrived, it still suggests the smaller role a “correct” WMC can have on the final answer compared to  $\oplus$  and  $\otimes$  implemented with proper set-based semantics. We even hypothesize that in most cases, the add-mult approximation does not meaningfully affect the final result compared to DTKP-WMC. This is empirically demonstrated by our benchmark results, which shows DTKP-AM achieving similar accuracy to Scallop’s implementation of DTKP-WMC. In fact, DAMP can be considered as a sort of ablation, where both the WMC and semiring operations use fuzzy add-mult semantics instead of a set-based one, and indeed, its accuracy often performs worse than both DTKP-WMC and DTKP-AM.

### A.2. Role of $+\infty$ and $-\infty$

In this section, we motivate the use of  $+\infty$  and  $-\infty$  in DTKP-AM’s tensor representation of tags. Because tensors  $t$  are rectangular where every proof  $i$  and symbol  $j$  must have an entry  $t_{ij}$ , we require a way to denote the absence of an input symbol from a proof, and the absence of a proof from a tag. Importantly, an absent symbol should not influence the probability of a proof (i.e. its normalized value should contribute 1 to the probability’s product), and an absent proof should not influence the probability of a tag (i.e. it should contribute 0 to the sum during add-mult WMC). Indeed this is captured by our definition of norm, which clamps  $+\infty$  to 1 (representing absent symbols) and  $-\infty$  to 0 (representing absent proofs) during any probability calculation. While this introduces clamping operations, PyTorch’s implementation of clamp backpropagation ensures a gradient of 1 everywhere, even on the clamp boundaries (source: <https://github.com/pytorch/pytorch/pull/7049>).

Since  $\hat{0}$  corresponds to the tag consisting of no proofs (i.e. a tag with probability 0), we initialize it to be a tensor where every proof is absent (all  $-\infty$ ). Likewise, since  $\hat{1}$  corresponds to the tag consisting of a single empty proof (i.e. a tag with probability 1), we initialize it to be a tensor where every symbol is absent from the first row / proof (all  $+\infty$ ), while the remaining rows / proofs are absent (all  $-\infty$ ).

### A.3. Further reading

For a more in-depth explanation of provenances in general, including the formalization of DTKP semantics with Boolean formulae, see Section 4 of (Li et al., 2023). For worked examples of provenance computation with comparisons of top-k provenances to DAMP, we refer the reader to (Scallop Language Group, 2022).

## B. Control Flows and Recursion in DOLPHIN

Table 4: Time taken by the symbolic program for the HWF task split by the time spent on the CPU and GPU. UDFs refer to user-defined functions where control flows reside for HWF. The times annotated with C and G indicate time spent on the CPU and GPU, respectively.

Config	Time for UDF (s)	Time for Tag Computations (s)	Total Time (s)
No Parallelism	36.24 (C)	461.02 (C)	497.26
Parallelized Tag Computations	14.13 (C)	75.125 (G)	89.25

In this section, we provide a more detailed explanation of how DOLPHIN handles control flows and recursion. In DOLPHIN, control flows largely exist within the lambda functions supplied to the ‘Apply’, ‘ApplyIf’, and ‘Filter’ operations, which can be arbitrary Python functions over the symbols in the Distributions. As discussed in Section 3.1.2, these functions can include complex operations like if-then-else branches, loops, and even recursions. We do assume that divergent control flows are resolved within the lambda function itself. The nature of these functions means that they cannot be parallelized over the GPU. Instead, they are executed sequentially on the CPU, while the associated tags are computed parallelly on the GPU. We optimize the design of the Distribution class so that there is one set of CPU-based computations for the entire batch of samples rather than one set of computations for each sample, which is typical of other neurosymbolic frameworks. This allows DOLPHIN to maintain the benefits of parallelism even while the user-defined functions are executed sequentially.

```

1 Coord = namedtuple('Coord', ['x', 'y'])
2
3 def compute_paths(paths, edges):
4     new_paths = apply_if(paths, edges, lambda p1, p2: Coord(p1.x, p2.y), lambda p1, p2: p1.
5         y == p2.x)
6     merged = union(paths, new_paths)
7     # checking for convergence via fix-point
8     if merged.symbols == paths.symbols:
9         return merged
10    else:
11        return compute_paths(merged, edges)
12
13 edges = Distribution(model(img), points)
14 paths = compute_paths(edges, edges)

```

Figure 6: Example of a transitive closure computation in DOLPHIN.

### B.1. Control Flow in HWF

We demonstrate this by showing the time taken by the symbolic program for the HWF task split by the time spent on the CPU and GPU in Table 4. The first row shows the time taken when the neurosymbolic model is run sequentially on the CPU with no parallelism. The second row shows the time taken when tag computations are parallelized on the GPU over batches of 64 samples each. The times annotated with C and G indicate time spent on the CPU and GPU, respectively. We only show the time taken in the forward pass in the table.

Observe that the time, both for UDF computation and for Tag computation, decreases as we move from sequential CPU evaluation to the batched evaluation. Due to DOLPHIN’s design, increases in batch size result in fewer CPU operations, since the set of CPU operations is shared for the entire batch, while parallelizing more tag computations over the entire batch.

### B.2. Recursion

In order to write recursive computations in DOLPHIN, one has two choices: either supply a recursive user-defined function to the DOLPHIN primitives, or write a more fine-grained program in Python that uses DOLPHIN primitives in the base case as well as the recursive case, set to terminate once a condition is met. Here, the diverging control flows can be merged using the UNION primitive. We follow the latter approach for tasks involving recursion, such as Path and CLUTRR. The crux of those programs involves performing a transitive closure computation over a graph, represented by a set of edges for Path or relations for CLUTRR. We show an example of a transitive closure computation in Figure 6.

Here, let’s say that `model` is a neural model that predicts the edges between each pair of points in a graph, represented by `points`. The `compute_paths` function computes the transitive closure of the graph by iteratively applying the edges to the paths. The `APPLYIF` function applies the edges to the paths if the end of the first path is the same as the start of the second path. The `UNION` function merges the new paths with the existing paths. The function `compute_paths` is called recursively until a fixpoint is reached, specifically until no new paths can be added. This is a simple example of a recursive computation in DOLPHIN, and also forms the core program needed for the PathFinder task. We perform a similar recursive computation for the CLUTRR task, where we find the transitive closure of a graph representing relations between people in a passage.

## C. On the Language and Semantics

### C.1. Language

To develop the operations provided by DOLPHIN, we studied several neurosymbolic tasks to determine the most common operations needed for these tasks. We found that the main operation needed in most programs is to apply a function to symbols from different input models and relations. This is primarily achieved via the join operation in Datalog, but we introduce the `Apply` or `ApplyIf` primitives for a more Pythonic approach. `Filters` are used to remove symbols violating conditions, similar to Datalog selections, while `Union` mimics the disjunction operation in Datalog, typically needed for writing recursive programs as described in Appendix B.

## C.2. Semantics

We designed DOLPHIN to be a general-purpose neurosymbolic framework able to support various semantics, as long as they can be expressed as operations over tags tracked via the Distribution class. DOLPHIN assumes that the provenance supplied to it offers both the conjunction and disjunction operations that operate over combinations of tags from input symbols, as well as a way to translate tags to probabilities. As long as these assumptions are satisfied, the primitives of DOLPHIN preserve the semantics offered by the provenances.

As such, supplying the DAMP provenance to the DOLPHIN program introduces basic fuzzy semantics which are preserved by DOLPHIN. However, there are cases where the independence assumptions may not hold and fuzzy semantics may not be appropriate.

The DTKP-AM provenance, on the other hand, offers an alternative without the assumption of variable independence, except on the input variables. At each step of the program, each symbol is associated with the tags of the input symbols that produce it via the proofs. Again, since DTKP-AM satisfies the aforementioned assumptions, the top-k semantics of the provenance are preserved.

These tags are then translated into probabilities by performing an add-mult operation over the proofs. This approximation of the WMC operation is more complex and results in a more precise translation of tags to probabilities. However, as we see in the experiments where Scallop uses DTKP-WMC, the accuracies achieved by DTKP-AM and DTKP-WMC are comparable.

## D. DOLPHIN Experiment Details for Benchmarks

### D.1. Comparison of Per Epoch Training Times

Table 5: Comparison of training times (in seconds) taken by each baseline. The Timeout (TO) is set at 10 hours.  $\alpha$  is the scaling factor, which is the ratio of the per epoch training times of the baselines and DOLPHIN.

Task	DOLPHIN	Scallop		LTN		ISED		IndeCateR+		
	T <sub>epoch</sub>	T <sub>epoch</sub>	α <sub>epoch</sub>	T <sub>epoch</sub>	α <sub>epoch</sub>	T <sub>epoch</sub>	α <sub>epoch</sub>	T <sub>epoch</sub>	α <sub>epoch</sub>	
SumN-5	10.77	184.76	17.16	4.63	0.43	59.93	5.56	59.54	5.52	
SumN-10	10.49	341.57	32.56	OOM	–	216.54	20.64	32.14	3.06	
SumN-15	10.47	493.87	47.17	OOM	–	653.39	62.41	23.84	2.28	
HWF-7	152.87	499.57	3.27	N.A.		790.04	5.17	540.26	3.53	
HWF-15	858.8	1.49e4	17.35			6.84e3	7.97	2.51e3	2.93	
HWF-19	1.4e3	3.92e5	280			1.05e4	7.5	4.18e3	2.99	
Path-32	1.29e3	2.2e3	1.71	N.A.						
Path-128	1.67e3	4.18e3	2.5							
Path-256	1.97e3	1.13e4	5.74							
CLUTRR-3	152.21	429.97	2.82	N.A.						
CLUTRR-4	165.13	783.11	4.74							
Mugen-1K	165.74	133.68	0.81	N.A.						
Mugen-5K	826.31	634.86	0.77							

The following are more details on the experiment setup. For each experimental trial, we report the highest evaluation accuracy over all epochs. Unless otherwise noted, each trial was run on a machine with Intel Xeon Gold 6248 (2.50 GHz) CPUs and NVIDIA GeForce RTX 2080 Ti (11 GB) GPUs. Additionally, for Scallop, the accuracy of the best-performing provenance is reported. Table 5 shows the training time per epoch for all of the baselines in each of the benchmarks.

### D.2. Accuracy without Timeouts

While Figure 5 shows the accuracy of baselines run until a timeout of 10 hours, Figure 7 shows their accuracy when run until convergence or until the test-time accuracy plateaus. We see that only IndeCateR+ for HWF-19 and Scallop for Path-256 are able to match DOLPHIN’s accuracies. Scallop and ISED are unable to converge for the larger versions of HWF, while ISED is unable to converge for the larger versions of SumN.

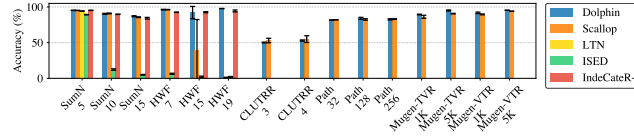


Figure 7: Accuracy of DOLPHIN and baselines across all benchmarks without any timeout.

### D.3. Effect of $K$ on Accuracy and Runtime

We report preliminary results on the effect of the top- $K$  value in DTKP-AM across different tasks. As shown in Tables 6 and 7, increasing  $K$  generally has little effect on final accuracy or per-epoch training time. HWF-19 is the only task where increasing  $K$  offers noticeable gains. For all other benchmarks, accuracy remains stable and runtime scales sub-linearly due to DTKP-AM’s vectorized implementation.

Table 6: Accuracy (%) of DTKP-AM with different values of  $K$  across five benchmarks.

Benchmark	$K = 1$	$K = 3$	$K = 5$	$K = 7$
Sum-15	9.61	10.81	10.51	10.21
HWF-19	8.94	99.15	96.89	95.75
Path-256	81.39	82.14	80.86	82.38
CLUTRR-4	53.62	48.52	50.35	48.17
Mugen-5K	(94.1 / 95.7)	(95.4 / 95.7)	(95.3 / 95.4)	(95.4 / 95.4)

Table 7: Training time per epoch (T/ep in seconds) for DTKP-AM with different values of  $K$ .

Benchmark	$K = 1$	$K = 3$	$K = 5$	$K = 7$
Sum-15	37.21	47.70	53.52	58.54
HWF-19	1.21e3	1.40e3	1.33e3	1.46e3
Path-256	1.97e3	2.34e3	2.10e3	2.12e3
CLUTRR-4	240.50	257.89	261.31	257.99
Mugen-5K	460.38	464.68	470.05	465.10

#### D.4. MNIST Sum-N

For this task, the base neural network model is a standard CNN (a 3-layer convolutional network with ReLU activations) classifying each image into 10 classes of digits (0, 1, ..., 9). The symbolic module sums the Distribution objects over the logits output by the neural model for each image.

Each of the MNIST Sum-N tasks had a batch train size of 64 samples, a learning rate of 0.001, and a top-k value of 1. Each of the tasks were trained on a dataset size of the original MNIST dataset divided by N of Sum-N. Sum5’s dataset consisted of 12000 train samples and 2000 test samples. Sum10’s dataset consisted of 6000 train samples and 1000 test samples. Sum15’s dataset consisted of 4000 train samples and 666 test samples. Figure 7 is a high-level overview of the Sum-N model’s architecture.

### D.5. Hand-Written Formula

For Hand-Written Formula, the perception model is again a standard CNN that classifies images into 14 classes: 10 digits (0, 1,  $\dots$ , 9), and 4 operations (+, -,  $\times$ , and  $/$ ). The DOLPHIN program for this task builds strings of formulae from the outputs of the neural model and evaluates them using Python’s `eval` function, demonstrating the ability of DOLPHIN to support black-box functions.

We trained each task with a batch train size of 64 samples. The learning rate was 0.0001, the global sampling value was 7, and top-k value was 3. Length 7’s dataset consisted of 9600 samples for training, 2400 samples for testing. Length 15 consisted of 24000 training samples and 6000 testing samples. Length 19 consisted of 32000 training samples and 8000 testing samples.



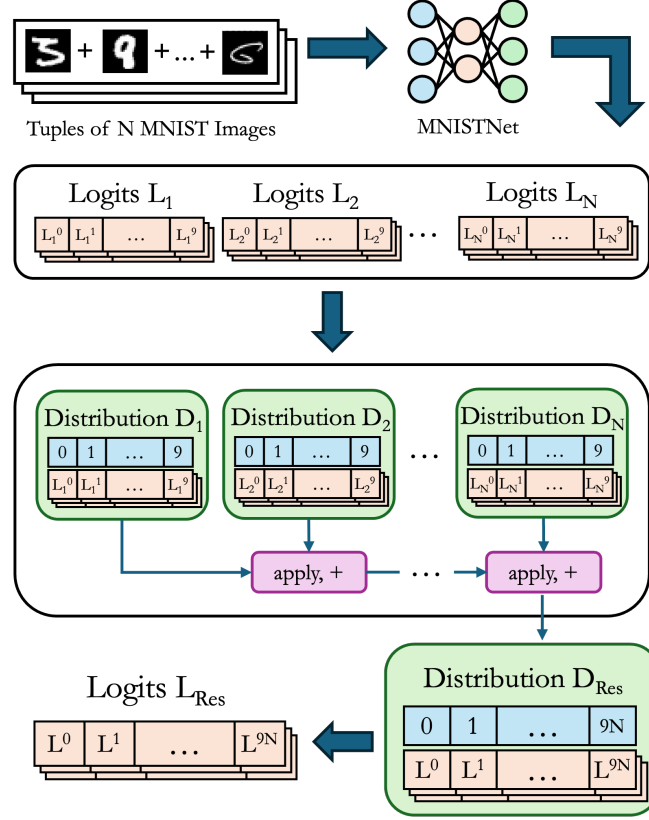


Figure 8: Components of the SumNNNet model written in Figure 2.

#### D.6. PathFinder

For this task, the perception model is also a CNN, but it predicts edges between pairs of nodes (denoted by dashes) as well as the end points depicted in the image of the maze. The DOLPHIN program for this task is recursive since it must search for paths between the two dots.

For each of the PathFinder tasks, we used a batch train size of 64 samples, a learning rate of 0.0001, and a top-k value of 1. Each task’s dataset consisted of 539459 images for training and 59940 images for testing. Each task had its own dataset of images with dimensions of the task’s pixel size.

#### D.7. CLUTRR

For each CLUTRR task, we used a single A100 GPU (40 GB), with a learning rate of 0.00001 and use a batch size of 16. The length of the training dataset for CLUTRR (Small) was 11,093 and that of the test set was 1146. The training set for CLUTRR (Medium) contained 15,083 samples and the test set contained 1048 samples.

The DOLPHIN program for CLUTRR receives as inputs pairs of entities from the input paragraph along with the logits for each pair over 21 possible relations produced by the classification head of the Roberta-base (Liu, 2019) model. The program then recursively derives relations over the graphs these pairs represent until no new relations can be derived. After that, it returns the Distribution over relations for the target pair of entities.

#### D.8. Mugen

For each Mugen task, we use a batch size of 3 and a learning rate of 0.0001. From the full Mugen dataset, we sample a training set of 5000 examples for Mugen (Medium), and from that set, we sample a training set of 1000 for Mugen (Small). Both Small and Medium are evaluated on a fixed holdout set of 1000 samples. We train and evaluate for up to 100 epochs.

We use a combination of DistilBert (Sanh, 2019) and S3D (De Smet et al., 2024) as the perception model for the text and video inputs respectively. The DOLPHIN program for both Mugen tasks computes the temporal alignment of a given text-video pair. The inputs (extracted from the text and video by neural components) are pairs of IDs and actions, where the ID order corresponds to the action sequence (e.g. the IDs for video actions are the frame numbers). The program finds the Distribution of all valid mappings between text event IDs and video frame IDs that preserve the order of actions.

## E. Graph Results of RQ3

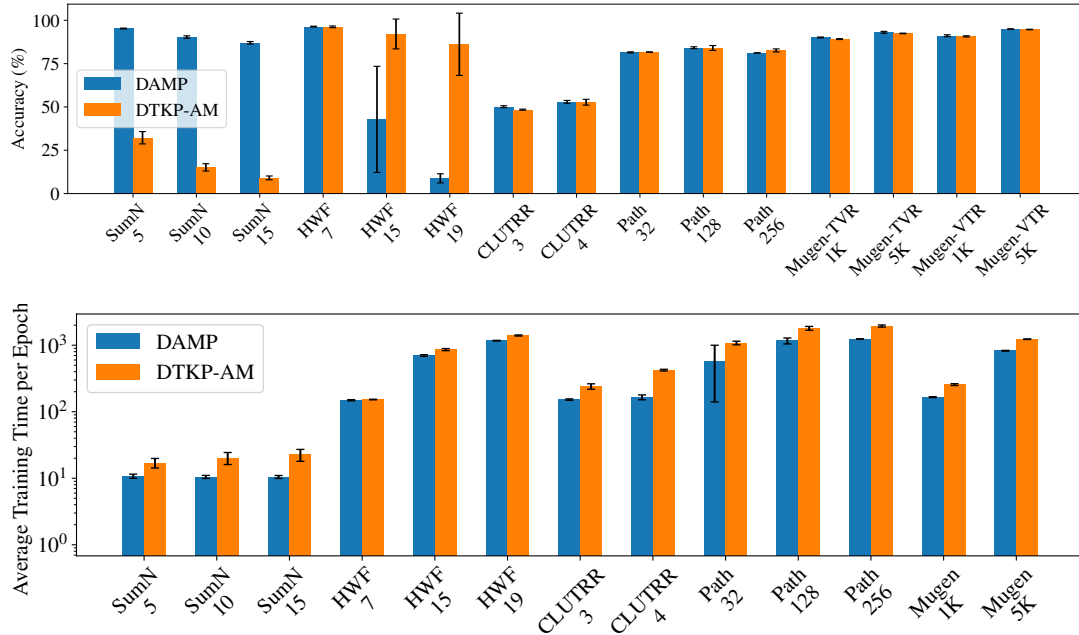


Figure 9: Accuracy and average training time per epoch for DAMP and DTKP-AM.

We show the results of the provenance comparison experiments (RQ3) in Figure 9. The graph on the top shows the accuracies achieved by each provenance over all tasks, while the bottom graph shows the average training time per epoch required for each provenance over all tasks.

## F. Comparison with Tensor-based Neurosymbolic Frameworks

Systems like LYRICS (Marra et al., 2019), Logic Tensor Networks (LTNs) (Badreddine et al., 2022), and Tensorlog (Cohen et al., 2020) all have limited expressivity, which is one of the obstacles DOLPHIN aims to overcome. Specifically, they restrict the symbolic programs to first order logic and require users to specify low-level information such as how variables are grounded and what their domains are. They also restrict the symbols to be in the form of tensors and the user defined functions to consist of TensorFlow operations. These restrictions allow such systems to use TensorFlow to compile these programs into highly efficient computational graphs, but at the cost of expressivity. These frameworks also exclusively support simpler provenances and t-norms which are not sufficient for complex neurosymbolic programs.

On the other hand, DOLPHIN allows the user to track tags for symbols which can be arbitrary Pythonic objects. DOLPHIN programs further allow the user to manipulate Distributions over such symbols using arbitrarily complex code which may not necessarily translate to a computational graph. As such, there is a fine balance between the probabilistic computations, that happen over a GPU, and the symbolic computations, that take place on a CPU, all while maintaining a mapping between the two. This fundamental design choice is also what allows DOLPHIN to be more expressive and flexible than existing systems. We also design DOLPHIN to be modular so that users can easily extend it to support new provenances and t-norms. As such, the t-norms used in LYRICS and LTN can be trivially added in a vectorized manner to DOLPHIN.

For instance, assume the case of MNIST Sum-2, where ‘model’ is the neural model. This is how it needs to be expressed in LTN:

```

1  ### Predicates
2  Digit = ltn.Predicate.FromLogits(model, activation_function="softmax")
3  ### Variables
4  d1 = ltn.Variable("digits1", range(10))
5  d2 = ltn.Variable("digits2", range(10))
6  ### Operators
7  Not = ltn Wrapper_Connective(ltn.fuzzy_ops.Not_Std())
8  And = ltn Wrapper_Connective(ltn.fuzzy_ops.And_Prod())
9  Or = ltn Wrapper_Connective(ltn.fuzzy_ops.Or_ProbSum())
10 Implies = ltn Wrapper_Connective(ltn.fuzzy_ops.Implies_Reichenbach())
11 Forall = ltn Wrapper_Quantifier(ltn.fuzzy_ops.Aggreg_pMeanError(), semantics="forall")
12 Exists = ltn Wrapper_Quantifier(ltn.fuzzy_ops.Aggreg_pMean(), semantics="exists")
13
14
15 # mask
16 add = ltn.Function.Lambda(lambda inputs: inputs[0]+inputs[1])
17 equals = ltn.Predicate.Lambda(lambda inputs: inputs[0] == inputs[1])
18
19 ### Axioms
20 @tf.function
21 def axioms(images_x, images_y, labels_z, p_schedule=tf.constant(2.)):
22     images_x = ltn.Variable("x", images_x)
23     images_y = ltn.Variable("y", images_y)
24     labels_z = ltn.Variable("z", labels_z)
25     axiom = Forall(
26         ltn.diag(images_x, images_y, labels_z),
27         Exists(
28             (d1, d2),
29             And(Digit([images_x, d1]), Digit([images_y, d2])),
30             mask=equals([add([d1, d2]), labels_z]),
31             p=p_schedule
32         ),
33         p=2
34     )
35     result_logits = axiom.tensor
36     return result_logits

```

Note that the FOL semantics of the Real Logic language used in LTN requires the user to specify the tracking of the probabilities with the symbols denoted by the ‘digits\*’ variables.

On the other hand, DOLPHIN’s design allows the user to write the same program in a more intuitive way:

```

1  d1 = Distribution(model(img[0]), range(10))
2  d2 = Distribution(model(img[1]), range(10))
3
4  result_logits = GetProbs(Apply(d1, d2, lambda x, y: x + y))

```

## F.1. Optimizing Probabilistic Computations

Other works such as (Dang et al., 2021) and (Darwiche, 2020), focus on solely on probabilistic computations rather than neurosymbolic frameworks. For instance, Juice (Dang et al., 2021) is a Julia package for logic and probabilistic circuits, which is not designed to be integrated with deep learning frameworks. On the other hand, (Darwiche, 2020) focuses on variable elimination with applications to optimize tensor-based computation. It will be interesting to see how DOLPHIN can be integrated with such systems to further improve the scalability and efficiency of neurosymbolic learning, and will include a discussion on this in the revised manuscript. However, we still believe that DOLPHIN’s novelty lies in its design that allows for the seamless integration of general purpose neurosymbolic programs within deep learning frameworks, which is not addressed by the existing systems.

Table 8: MNIST ProductN Training Epoch Times in Seconds.

N	B = 64		B = 256	
	Time per Epoch (s)	Accuracy	Time per Epoch (s)	Accuracy
4	11.42	0.96	8.92	0.97
8	12.55	0.95	9.15	0.95
16	27.45	0.94	15.71	0.89
20	36.59	0.92	18.73	0.85

## G. On Combinatorial Explosions

The nature of the APPLY and APPLYIF primitives imply the possibility of combinatorial ballooning of computations in cases where either the number of symbols is large or where there are several distributions over which the function is applied. This is indeed a fundamental challenge in neurosymbolic frameworks as a whole. DOLPHIN mitigates this by leveraging the Distribution class, which condenses symbols into a single collection stored in CPU RAM while maintaining tags as a GPU tensor ( $b \times N \times T$ , where  $b$  is the batch size,  $N$  is the number of symbols and  $T$  is the shape of the tag). As shown in Figure 8, this approach reduces symbolic overhead by avoiding redundant evaluations for each batch sample, unlike frameworks like Scallop, where each sample in a batch is independently evaluated. While tag evaluations still involve all combinations across all samples in a batch, they are computed in a vectorized manner on the GPU.

To see the effect of such computations even on larger experiments, we consider MNIST ProductN, where we multiply digits classified by the MNIST CNN as opposed to adding them in SumN. We show the per epoch training times in Table 8 for batch sizes of 64 and 256. In both cases, the DOLPHIN program is able to achieve high accuracies even for  $N=20$  while running in reasonable amounts of time. The scaling gets even better for larger batch sizes (e.g. 256) since it increases the number of parallelized operations executed at any given time.

## H. The HWF Model

We show the neurosymbolic model written in DOLPHIN for the HWF task along with the base neural model. In the HWF task, the neural model simply classifies each input image into 14 symbols: 10 digits and 4 operations.

```

1 class SymbolNet(nn.Module):
2     def __init__(self):
3         super(SymbolNet, self).__init__()
4         self.conv1 = nn.Conv2d(1, 32, 3, stride = 1, padding = 1)
5         self.conv2 = nn.Conv2d(32, 64, 3, stride = 1, padding = 1)
6         self.fc1 = nn.Linear(30976, 128)
7         self.fc1_bn = nn.BatchNorm1d(128)
8         self.fc2 = nn.Linear(128, 14)
9
10    def forward(self, x):
11        x = self.conv1(x)
12        x = F.relu(x)
13        x = self.conv2(x)
14        x = F.max_pool2d(x, 2)
15        x = F.dropout(x, p=0.25, training=self.training)
16        x = torch.flatten(x, 1)
17        x = self.fc1(x)
18        x = self.fc1_bn(x)
19        x = F.relu(x)
20        x = F.dropout(x, p=0.5, training=self.training)
21        x = self.fc2(x)
22        return F.softmax(x, dim=1)

```

This neural model is then used in the DOLPHIN program as follows:

```

1 class HWFNet(nn.Module):
2     def __init__(self):
3         super(HWFNet, self).__init__()

```



```

4
5 # Symbol embedding
6 self.symbol_cnn = SymbolNet()
7 self.operators = [("+", ), ("-", ), ("*", ), ("/", )]
8 self.symbols = [ (str(i),) for i in range(10)] + self.operators
9
10 self.db = torchql.Database()
11
12 def forward(self, img_seq, img_seq_len):
13     batch_size, formula_length, _, _, _ = img_seq.shape
14     length = [l.item() for l in img_seq_len]
15
16     inp = img_seq.flatten(start_dim=0, end_dim=1)
17     symbol = self.symbol_cnn(inp).view(batch_size, -1, 14)
18
19     def eval_formula(s):
20         try:
21             return eval("".join(s))
22         except:
23             return math.nan
24
25     def concat_symbol(formula, symbol):
26         if formula[-1] == " ":
27             return formula
28         else:
29             if not isinstance(symbol, tuple):
30                 symbol = (symbol,)
31             formula += symbol
32             if len(formula) %
33                 if formula[-2] in ["*", "/"]:
34                     eval_result = str(eval_formula(formula[-3:]))
35                     formula = formula[:-3] + (eval_result,)
36             return formula
37
38     def infer_expression(length, *symbols):
39         res = symbols[0]
40         for i in range(1, len(symbols)):
41             res = Apply(res, symbols[i], concat_symbol)
42         x = (Apply(res, eval_formula), )
43         return x
44
45     def reorg(symbols, lengths):
46         distrs = []
47         for i in range(symbol.shape[1]):
48             if i < lengths:
49                 distrs.append(Distribution(symbols[i, :].view(-1, 14), self.symbols))
50             if i %
51                 distrs[-1] = distrs[-1].filter(lambda s : s not in self.operators)
52             else:
53                 distrs[-1] = distrs[-1].filter(lambda s : s in self.operators)
54             else:
55                 distrs.append(Distribution(torch.ones(1, device=device), [("("), )])
56
57     res = (lengths, *distrs)
58     return res
59
60 q = torchql.Query("hwf", base="symbols").join("lengths") \
61     .project(lambda symbols, lengths: reorg(symbols, lengths)) \
62     .project(infer_expression, batch_size=batch_size)
63
64 res = q(db, tensors={"symbols": symbol, "lengths": length}, disable=True).rows
65
66 stacked = Distribution.stack(res)
67 return GetProbs(stacked)

```

The `HWFNet` class is the neurosymbolic model. It takes in a sequence of images, `img_seq`, and their lengths, `img_seq_len`. Note that within a single batch there may be image sequences of varying lengths. The neural model, `symbol_cnn`, is used to classify each image in the sequence into one of the 14 symbols. Since we know that each number in the expression is a single digit, the `reorg` function is used to filter out relevant symbols based on their position in the sequence (operators in even places, digits in odd places). This function also pads sequences of smaller lengths with empty strings, written as Distributions with a single element and a probability of 1. Once reorganized, the `infer_expression` function is used to infer the expression from the symbols. It does so by first concatenating Distributions using the `concat_symbol` function, which also performs partial evaluations whenever possible. Once all the symbols are concatenated, the expression is evaluated using the `eval_formula` function. The final expression is then returned as a Distribution. As a sidenote, while optional, we use the TorchQL (Naik et al., 2024) library to help write certain parts of the program. This shows the ease with which Distributions can be used with existing machine learning frameworks.

For such a complex DOLPHIN program, using a simple provenance like DAMP proves insufficient for longer sequences since the tags of all possible combinations of symbols are collated into a single number. On the other hand, DTKP-AM is able to track the top-k proofs for each symbol, pruning out the less probable proofs. Furthermore, since each proof is a collection of *input* symbols leading to a specific output, once the loss is calculated, gradients can be backpropogated directly to the input symbols that had the most influence on the output. On the other hand, the gradients may be distributed across all symbols in DAMP as it backpropogates through each intermediate computation regardless of their role in the computation of the output, resulting in slower convergence.

### H.1. Writing HWF using LTNs

The crux of the HWF program written in DOLPHIN relies on string concatenations and using Python’s `eval` function. In order to write HWF using LTNs, we need a different approach, since LTNs require all symbols to be grounded as tensors and all functions to be operations over those tensors. While one can ground all the possible strings (representing  $0, 1, \dots, 9$  and operators  $+, -, \times, /$ ) as real-value tensor encodings, one cannot execute functions like `eval` over those encodings since it cannot be compiled onto the tensorflow computation graph.

This leaves us with the option of trying to write the HWF program as a probabilistic parser, as seen in Scallop. However, such a parser needs to be able to recursively parse arbitrarily long expressions, since HWF expressions are not of a fixed length even within the same task. This means it needs the ability to evaluate subtrees of the expression AST and produce intermediate results as new constants that can be populated within relations and operated over.

Based on our understanding of LTNs, the LTN paper does not provide any information on the dynamic creation of new constants or adding new domain elements at inference time, and instead focuses on examples and experiments where the domains are static and grounded prior to the evaluation of the logic formulae. We also spent considerable time perusing the official LTN repository, but could not find examples that could guide us towards writing an implementation of HWF.

We try to simulate the creation of a constant at inference time. In the following code, we attempt to write the program for MNIST Sum-3 by adding the first two digits in one LTN function, and then adding the result to the third digit, as follows (where the MNIST model is represented as `model`):

```

1  ### Predicates
2  Digit = ltn.Predicate.FromLogits(model, activation_function="softmax")
3  ### Variables
4  d1 = ltn.Variable("digits1", range(10))
5  d2 = ltn.Variable("digits2", range(10))
6  d3 = ltn.Variable("digits3", range(10))
7  ### Operators
8  Not = ltn Wrapper_Connective(ltn.fuzzy_ops.Not_Std())
9  And = ltn Wrapper_Connective(ltn.fuzzy_ops.And_Prod())
10 Or = ltn Wrapper_Connective(ltn.fuzzy_ops.Or_ProbSum())
11 Implies = ltn Wrapper_Connective(ltn.fuzzy_ops.Implies_Reichenbach())
12 Forall = ltn Wrapper_Quantifier(ltn.fuzzy_ops.Aggreg_pMeanError(), semantics="forall")
13 Exists = ltn Wrapper_Quantifier(ltn.fuzzy_ops.Aggreg_pMean(), semantics="exists")
14
15 # mask
16 add = ltn.Function.Lambda(lambda inputs: inputs[0]+inputs[1])
17 equals = ltn.Predicate.Lambda(lambda inputs: inputs[0] == inputs[1])
18

```

```
19 ### Axioms
20 @tf.function
21 def axioms(images_x, images_y, images_w, labels_z, p_schedule=tf.constant(2.)):
22     images_x = ltn.Variable("x", images_x)
23     images_y = ltn.Variable("y", images_y)
24     images_w = ltn.Variable("w", images_w)
25     labels_z = ltn.Variable("z", labels_z)
26     diagonal = ltn.diag(images_x, images_y, images_w, labels_z)
27     formula = And(
28         Digit([images_x, d1]),
29         And(
30             Digit([images_y, d2]),
31             Digit([images_w, d3])
32         )
33     )
34     exists = Exists(
35         (d1, d2, d3),
36         formula,
37         mask>equals([add([add([d1, d2]), d3]), labels_z]),
38         p=p_schedule
39     )
40     axiom = Forall(
41         diagonal,
42         exists,
43         p=2
44     )
45     sat = axiom.tensor
46     return sat
```

Here, rather than adding all three digits  $d1$ ,  $d2$ , and  $d3$ , we write the following mask: `equals([add([add([d1, d2]), d3]), labels_z])`. However, this line results in a typecheck error, since functions can only take terms, but `add([d1, d2])` returns a formula.

Given the lack of resources on such functionalities in LTN both in the paper, tutorials, and the examples in their repository, we were not able to write an implementation of HWF to evaluate DOLPHIN against.