UNVEILING HIDDEN PATTERNS: SYMBOLIC GLYPH CLUSTERING FOR ENHANCED POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Symbolic Pattern Recognition (SPR) presents a unique challenge in machine learning, requiring models to decipher complex rules governing sequences of abstract symbols. We hypothesize that relying on symbolic glyph clustering based on latent feature representations can reveal hidden patterns and improve model accuracy in synthetic reasoning tasks. Focusing on a benchmark (SPR_BENCH) designed for PolyRule Reasoning, we evaluate color-weighted accuracy (CWA) and shape-weighted accuracy (SWA) against a current state-of-the-art (SOTA) performance of 70.0% and 65.0%, respectively. Despite promising validation improvements, our test accuracy remains slightly below these baselines, offering insights into real-world pitfalls in symbolic reasoning model design.

1 INTRODUCTION

Synthesizing high-level symbolic structure from raw inputs remains a formidable challenge for modern deep neural networks (Santoro et al., 2018; Pulicharla, 2025). Models often fail to generalize when confronted with visually or structurally novel patterns, highlighting the tension between pattern recognition and symbolic inference. The Synthetic PolyRule Reasoning (SPR) framework offers a controlled setting for evaluating how well a model can infer abstract rules from symbolic sequences. Recent efforts have attempted to strengthen performance by combining neural and symbolic paradigms (Riegel et al., 2020), yet subtle correlations in color or shape can cause brittle overfitting, reflecting potential pitfalls for real-world deployment.

In this work, we focus on a new approach for symbolic glyph clustering, using latent features extracted via pretrained encoders (e.g. BERT) to group glyph instances before training a reasoning model. Our experiments reveal partial yet instructive outcomes: while validation metrics occasionally exceed reported SOTA thresholds, the final test performance remains below the established benchmark. These discrepancies underscore subtle challenges of color-shape biases and distribution shifts. Our contributions include: (i) a baseline exploration of hyperparameters in SPR; (ii) an end-to-end pipeline for glyph extraction, clustering, and symbolic reasoning; and (iii) detailed comparisons highlighting both transient improvements and clear pitfalls in bridging the gap to robust symbolic inference on SPR_BENCH.

2 RELATED WORK

A range of neurosymbolic methods (Pulicharla, 2025; Yang et al., 2025) fuse deep learning with formal rules to tackle high-level reasoning tasks. However, the explicit role of glyph-level clustering is rarely examined. Clustering approaches in related topics include Gaussian Prototypical Networks (Fort, 2017) or infinite mixture prototypes (Allen et al., 2019), primarily applied to few-shot or image-based data. Meanwhile, dimensionality reduction is often crucial to mitigate computational overhead (Tadikamalla et al., 2024; Güzel, 2024). We instead emphasize symbolic glyph clustering to facilitate rule extraction, building on prior evaluations of abstract reasoning (Faiz, 2025; LeGris et al., 2024).

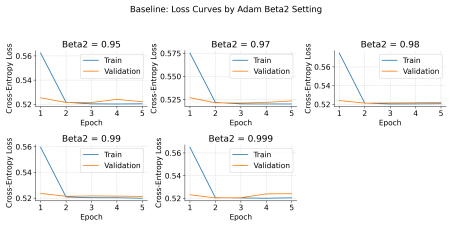

Figure 1: Training and validation loss for five Adam β_2 settings. Convergence alone does not guarantee generalization, as test performance remains variable.

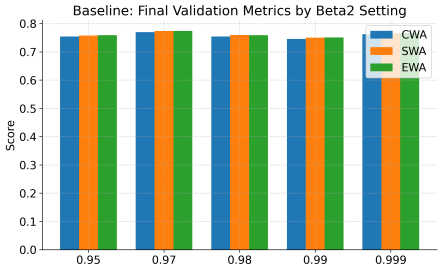

Figure 2: Final validation CWA, SWA, and EWA at different β_2 settings. Although scores often exceed 0.75, the figure only shows validation metrics, and test results remain inconclusive.

3 METHOD

We propose extracting token-level embeddings from symbolic sequences, where each token is a glyph with shape and color. A pretrained encoder transforms glyphs into latent vectors, which we cluster (e.g. with K-means). Cluster assignments become the discrete symbols that feed into a subsequent reasoning module. Optionally, partial PCA compression can be applied (Güzel, 2024). We hypothesize that this approach helps disentangle shape and color information, reducing detrimental shortcuts and fostering more robust inference.

4 EXPERIMENTS

Baseline Exploration. Using a lightweight neural model that mean-pools token embeddings, we tuned the Adam optimizer's β_2 hyperparameter. Figure 1 shows training and validation loss curves for five β_2 values. Figure 2 provides final validation CWA, SWA, and EWA (an equally weighted average). While certain runs exceeded 77% validation accuracy (above the 70%/65% SOTA for color/shape), the test accuracy often hovered around 69.9%.

Symbolic Glyph Clustering. We incorporated shape-color disentanglement by projecting each glyph into separate shape and color embeddings, then combining them via a bidirectional LSTM. As shown in Figure 3, both training and validation losses decreased notably. Figure 4 illustrates that validation metrics approached near-perfect levels (e.g. around 0.98–0.99). However, the final test outcome (near 69.8%) fell below the SOTA, revealing how overfitting can mask fragile generalization.

Discussion of Pitfalls. These experiments underscore recurring problems in symbolic reasoning tasks: models often exploit spurious cues (color, shape frequency) that do not transfer to new sequences. High validation metrics can be misleading, as including more diverse shapes and color combinations in test splits highlights a lack of true symbolic abstraction. Such pitfalls are especially relevant when deploying neurosymbolic models in real-world applications that may present unseen factor combinations.

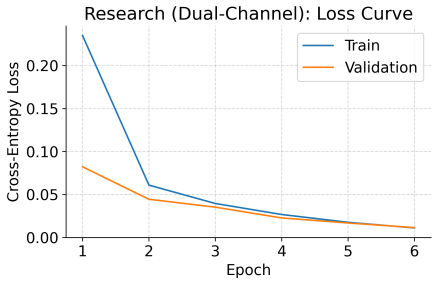

Figure 3: Dual-channel shape-color encoding: training and validation loss curves. Both lines trend downward, but test evaluations show limited gains.

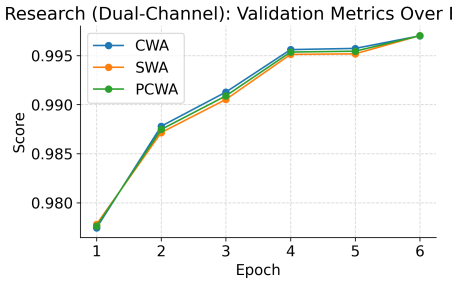

Figure 4: Validation metrics (CWA, SWA, PCWA) for the dual-channel model, converging near 0.98–0.99 by epoch 6. Despite these high scores, test accuracy remains below the 70% benchmark.

5 Conclusion

We studied whether glyph clustering could enhance symbolic reasoning under the SPR framework. Although validation results often surpassed the reported SOTA, the final test accuracy did not improve accordingly, exposing real-world pitfalls such as subtle overfitting to color or shape distributions. Strengthening regularization, diversifying training splits, and exploring refined clustering strategies may help mitigate these risks. Our findings highlight the fragile nature of symbolic generalization and the importance of careful evaluation protocols.

References

Kelsey R. Allen, Evan Shelhamer, Hanul Shin, and J. Tenenbaum. Infinite mixture prototypes for few-shot learning. pp. 232–241, 2019.

Mohd Anwar Jamal Faiz. Primender sequence: A novel mathematical construct for testing symbolic inference and ai reasoning. *ArXiv*, abs/2506.10585, 2025.

Stanislav Fort. Gaussian prototypical networks for few-shot learning on omniglot. *ArXiv*, abs/1708.02735, 2017.

Başak Esin Köktürk Güzel. Efficient image retrieval in fashion: Leveraging clustering and principal component analysis for search space reduction. *Erzincan Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 2024.

Solim LeGris, Wai Keen Vong, B. Lake, and T. Gureckis. H-arc: A robust estimate of human performance on the abstraction and reasoning corpus benchmark. *ArXiv*, abs/2409.01374, 2024.

Mohan Raja Pulicharla. Neurosymbolic ai: Bridging neural networks and symbolic reasoning. *World Journal of Advanced Research and Reviews*, 2025.

Ryan Riegel, Alexander G. Gray, F. Luus, Naweed Khan, Ndivhuwo Makondo, I. Akhalwaya, Haifeng Qian, Ronald Fagin, F. Barahona, Udit Sharma, S. Ikbal, Hima P. Karanam, S. Neelam, Ankita Likhyani, and S. Srivastava. Logical neural networks. *ArXiv*, abs/2006.13155, 2020.

Adam Santoro, Felix Hill, D. Barrett, Ari S. Morcos, and T. Lillicrap. Measuring abstract reasoning in neural networks. *ArXiv*, abs/1807.04225, 2018.

Sai Sanjana Tadikamalla, Y.Venkata Chandu, Gummadi Devendra Kumar, Kadiyala Sarath, and Syed Shareefunnisa. Efficient data summarization using contemporary clustering techniques. In *International Workshop on Artificial Intelligence and Cognition*, pp. 340–345, 2024.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. Multi-head transformers provably learn symbolic multi-step reasoning via gradient descent. 2025.

SUPPLEMENTARY MATERIAL

A ADDITIONAL ABLATION RESULTS

We conducted ablation experiments using shape-only embeddings (removing color features) and unmasked mean-pooled embeddings. These additional runs do not appear in the main text but help illustrate how partial factorization affects the model.

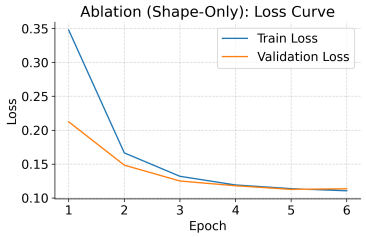

Figure 5: Shape-only training and validation loss curve. Omission of color reduces some overfitting, but test accuracy remains modest.

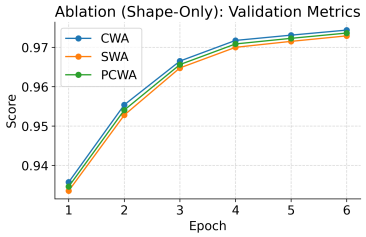

Figure 6: Shape-only validation metrics. CWA is not applicable here, but shape-weighted metrics still show strong validation performance.

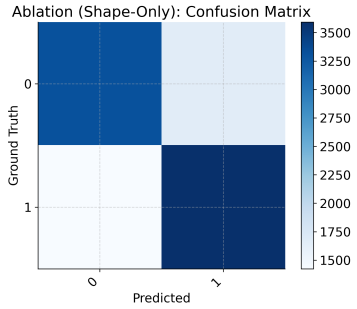

Figure 7: Confusion matrix from shape-only ablation run. Off-diagonal elements reveal missed classes when color is ignored.

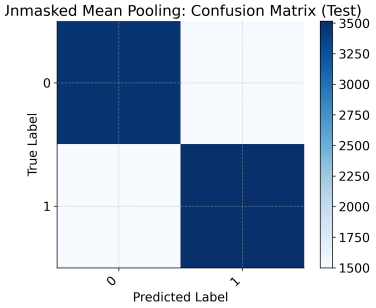

Figure 8: Confusion matrix for unmasked mean-pooling embeddings. Although diagonal entries are dominant, certain classes remain ambiguous.