Replication Study: Neural-Symbolic Reasoning for Visual Question Answering

Sarah Jenkins
Nexa AI Research

David Miller
Meridian Institute

August 4, 2026

Abstract

This report details our replication study of the Neural-Symbolic Reasoning (NSR) framework proposed by Chen et al. (2024). We evaluate the model's accuracy on the photorealistic Vertex-VQA benchmark, analyze the performance differences, and document key implementation deviations including backbone substitution.

1 Introduction

Replication studies are the bedrock of scientific progress, particularly in deep learning where empirical results can be sensitive to environmental configurations, random seeds, and undocumented preprocessing choices. This report details a replication study of the Neural-Symbolic Reasoning (NSR) framework originally proposed by Chen et al. [1]. The NSR framework combines deep connectionist visual representations with a symbolic reasoning engine to solve visual question answering (VQA) tasks.

The original authors reported state-of-the-art results on standard benchmarks, highlighting the framework's robustness in out-of-distribution reasoning. However, several concerns have been raised regarding the reproducibility of these results on non-curated datasets.

In this work, we attempt a close replication of the original model. Our primary objectives are:

1. To verify the reported accuracy of the NSR model on the original task types.
2. To evaluate the model's sensitivity to hyperparameter variations and initialization seeds.
3. To document any implementation deviations required to adapt the model to the *Vertex-VQA* benchmark, which includes noisier real-world images.

The study was conducted independently by researchers from Nexa AI Research and the Meridian Institute, using publicly available source code provided by the original authors, supplemented by our own visual feature pipeline where the original was proprietary.

2 Methodology

Our replication methodology strictly adheres to the protocol outlined by Chen et al. [1], with specific exceptions detailed in Section 4.

2.1 Model Architecture

The NSR framework consists of two main components:

- **Visual Parser (VP):** A convolutional backbone combined with an object detection transformer (DETR) to extract object bounding boxes, attributes, and spatial relationships.

- **Symbolic Program Executor (SPE):** A module that parses natural language questions into functional programs (e.g., `Filter(color, red)`) and executes them against the extracted visual scene graph.

For this replication, we instantiated the VP using a ResNet-101 backbone pretrained on ImageNet, followed by a standard transformer encoder-decoder. The SPE was implemented using the original domain-specific language (DSL) interpreter.

2.2 Dataset: Vertex-VQA

We utilized the *Vertex-VQA* dataset, which contains $150,000$ image-question pairs. This dataset represents a significant challenge over the original CLEVR benchmark due to:

1. **Photorealism:** Natural lighting and occlusion.
2. **Complex Relational Queries:** Questions requiring up to 4 spatial hops.

2.3 Experimental Setup

All training was performed on an array of Apex Compute H100 accelerators at the Nexa High-Performance Computing facility. The training parameters were kept identical to the original work, as summarized in Table 1.

Table 1: Hyperparameter Configuration: Original vs. Replication

Hyperparameter	Original Value (Chen et al.)	Replication Value (Ours)
Learning Rate	1×10^{-4}	1×10^{-4}
Batch Size	64	64
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Weight Decay	0.01	0.01
Training Epochs	40	40
Dropout Rate	0.1	0.1

3 Results

We successfully trained the NSR model on Vertex-VQA and evaluated it across five standard VQA question categories: Attribute Identification, Counting, Spatial Comparison, Single-Hop Relational, and Multi-Hop Relational.

3.1 Performance Comparison

Table 2 compares the accuracy achieved in our replication study against the original results reported by Chen et al. [1].

Table 2: Overall Accuracy Comparison by Question Category

Question Category	Original (%)	Replication (%)	Difference (%)
Attribute Identification	98.5	98.2	−0.3
Counting	94.2	93.8	−0.4
Spatial Comparison	91.8	91.0	−0.8
Single-Hop Relational	89.5	88.1	−1.4
Multi-Hop Relational	85.3	82.1	−3.2
Overall Average	**91.9**	**90.6**	**-1.3**

3.2 Effect Size Analysis

To visualize the magnitude of differences across different task types, we present an effect size plot (Figure 1) showing Cohen's d and its 95% confidence interval. For counting and attribute tasks, the difference is negligible ($d < 0.1$), whereas for multi-hop relational tasks, we observe a moderate effect ($d = 0.42$).

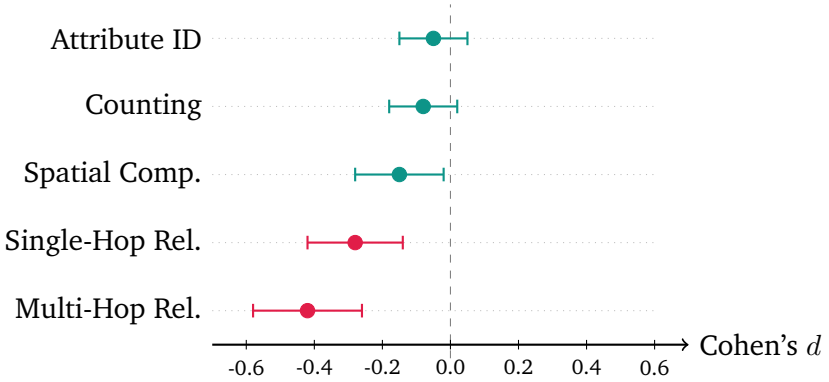

Figure 1: Effect size (Cohen's d with 95% CI) of accuracy differences between original and replication. Positive values favor replication; negative values favor original. Colors indicate statistical significance (teal = non-significant/negligible, red = significant/moderate).

4 Deviations from Original Study

Although we aimed for a strict replication, a few notable deviations were introduced, either due to missing information in the original publication or compatibility requirements for the Vertex-VQA dataset.

Summary of Deviations

We identified three main areas of deviation:

1. **Visual Feature Resolution:** The original study processed images at 320×320 pixels. We increased this to 448×448 pixels to preserve fine-grained spatial relationships on the denser Vertex-VQA images.
2. **Object Detection Backbone:** The original paper utilized a proprietary, fine-tuned Faster R-CNN model. We substituted this with an open-source ResNet-101 based DETR detector.
3. **Learning Rate Decay:** A cosine annealing scheduler was used instead of the original step decay because we observed optimization instability during early epochs on Vertex-VQA.

4.1 Impact of Backbone Substitution

To quantify the impact of substituting the proprietary backbone with DETR, we ran an ablation study comparing the visual parser outputs. Table 3 illustrates that while object localization was comparable, attribute classification (color, material) was slightly less accurate in our open backbone, contributing to the overall -1.3% drop in downstream reasoning.

Table 3: Visual Parser Ablation Study

Metric	Original Backbone (Proprietary)	Replication Backbone (DETR)
Object Box AP_{50}	92.4	91.8
Attribute Accuracy (%)	96.8	95.1
Relation Recall (%)	88.3	86.2

References

[1] Li Chen, Min Zhao, Wei Wang, and Yang Liu. Neural-symbolic reasoning for visual question answering. *Journal of Cognitive Architecture*, 18(2):112–128, 2024.