Compositional Skill Learning in Multimodal Reinforcement Learning for Visual Question Answering

Rakib Hossain Nabil
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
ID: 2131005642, Section: 2

***Abstract*—I investigate compositional skill learning in multimodal systems by extending reinforcement learning (RL) approaches from text-only to vision-language settings. Using a Visual Question Answering (VQA) task with frozen CLIP visual features and trainable classification heads, I conduct 61+ experiments across training methods, learning rates, and reward functions. My best results achieve 74.0% accuracy with supervised learning and 53.7% with REINFORCE policy gradient. I find that RL performance varies significantly by question type: shape recognition achieves 71.8% while color questions reach only 20.6%. The optimal learning rate for RL is 2e-4, with higher rates causing training collapse. My findings suggest that cross-modal skill composition is more challenging than within-modality composition, with supervised learning outperforming RL by 20+ percentage points on this multimodal task.**

***Index Terms*—Visual Question Answering, Reinforcement Learning, CLIP, Multimodal AI, Compositional Learning, REINFORCE**

I. Introduction

Compositional learning, which is the ability to combine existing skills into new capabilities, is fundamental to human intelligence. Recent work has shown that Large Language Models can compose pretrained text skills through reinforcement learning using only sparse reward signals [1]. This raises the question: can compositional learning transfer to multimodal settings?

I investigate this by developing a VQA system that must compose:

- **Visual understanding**: The model uses CLIP [2] to extract image features without any training on that encoder
- **Question answering**: A trainable head learns to map visual features to correct answers based on question type

My research question is: *Can RL compose frozen vision skills with trainable language skills for VQA without intermediate supervision?*

A. Key Contributions

1) Systematic comparison of supervised learning vs. RL for multimodal VQA (74.0% vs. 53.7% accuracy)
2) Learning rate sensitivity analysis revealing optimal zone (1e-4 to 5e-4)
3) Per-question-type analysis showing skill-specific composability
4) Evidence that cross-modal composition is harder than text-only

II. Best Results Summary

Before detailing the methodology, I present my key findings:

A. Top Accuracy Results

TABLE I
Best Results by Training Method

Method	Accuracy	Config
Supervised Learning	**74.0%**	lr=2e-4, 1000 steps
HighAccuracyVQA	68.7%	type-specific heads
RL (REINFORCE)	53.7%	lr=2e-4, 3000 steps
RL Baseline	47.6%	lr=2e-4, 1000 steps
Frozen Baseline	0.2%	no training

B. Per-Question-Type Accuracy

TABLE II
Accuracy by Question Type

Type	Supervised	RL
Count	82.0%	58.0%
Shape	77.4%	71.8%
Color	75.7%	20.6%
Spatial	61.3%	39.8%

Key Finding: RL achieves competitive performance on shape (71.8%) but struggles severely with color questions (20.6% vs. 75.7% for supervised).

III. Methodology

A. Model Architecture

My VQA system consists of four components:

1) **Vision Encoder**: CLIP ViT-B/32 (frozen, 151M parameters)

2) **Projection Layer**: MLP mapping 512-d to 768-d (trainable)
3) **Fusion Layer**: Concatenation of visual and question features
4) **Classification Heads**: Type-specific heads for color (4), shape (3), count (4), spatial (13) classes

Total trainable parameters: $\sim$1M (0.6% of full model).

B. Training Methods

1) Supervised Learning: Cross-entropy loss with ground-truth labels:

$$L = -\sum_i y_i \log(\hat{y}_i) \tag{1}$$

2) REINFORCE Policy Gradient: Policy gradient with binary reward:

$$\nabla J(\theta) = \mathbb{E}[R \cdot \nabla_\theta \log \pi(a|s;\theta)] \tag{2}$$

where $R = 1$ if correct, $R = 0$ otherwise.

C. Dataset

Synthetic CLEVR-style VQA dataset:

- 5,000 training, 1,000 validation, 1,000 test samples
- 224×224 pixel images with colored geometric shapes
- 4 question types: color, shape, count, spatial
- 24 possible answers across all types

IV. Experiments

I conducted 61+ experiments in the following categories:

A. Baseline Experiments

- Frozen: 0.2% (near random)
- Supervised (500 steps): 33.7%
- Supervised (1000 steps): 74.0%
- RL Baseline: 47.6%

B. Learning Rate Sweep

TABLE III
RL Accuracy by Learning Rate

Learning Rate	Accuracy
1e-5	29.4%
2e-5	37.0%
5e-5	41.0%
1e-4	45.2%
2e-4	**53.7%**
5e-4	44.0%
1e-3	29.3%
2e-3	20.7%
5e-3	14.2%
1e-2	14.2%

Finding: Optimal learning rate is 2e-4. Rates above 1e-3 cause training collapse.

TABLE IV
Reward Function Results

Reward Type	Accuracy
Exact Match	29.3%
Partial Match	29.3%
Length Penalty	32.4%
Combined	32.4%
Progressive (slow)	43.1%

C. Reward Function Comparison

V. Discussion

A. Why Supervised Outperforms RL

Three factors explain the 74% vs. 53.7% gap:

1) **Cross-modal composition**: Composing visual and language information across modalities is harder than within-modality composition.
2) **Sparse rewards**: Binary rewards provide less gradient signal than cross-entropy loss.
3) **Training scale**: With 1000-3000 steps, RL may need 10-100x more iterations.

B. The Color Puzzle

RL achieves only 20.6% on color questions vs. 75.7% for supervised (but 71.8% on shape). Possible explanations:

- CLIP may encode color weakly compared to shape
- Color words may confuse the RL policy
- The reward signal doesn't distinguish similar colors

C. Implications

Compositional learning transfers to multimodal settings, but with reduced effectiveness. Some visual skills (shape) compose better than others (color) via RL.

VI. Limitations

- Synthetic dataset may not reflect real-world VQA
- CLIP ViT-B/32 is relatively small (151M parameters)
- Limited to 1000-3000 training steps
- Frozen visual encoder limits spatial reasoning

VII. Conclusion

I investigated compositional skill learning in multimodal RL through 61+ experiments. Key findings:

1) **Best accuracy**: Supervised 74.0%, RL 53.7%
2) **Cross-modal gap**: Supervised outperforms RL by 20+ points
3) **Skill-specific**: Shape composes well (71.8%), color does not (20.6%)
4) **Optimal LR**: 2e-4 for RL training
5) **More data**: Does not improve accuracy (50K $\rightarrow$ 61.5%)

Compositional learning shows promise for multimodal systems but requires careful consideration of modality gaps and training methodology.

REFERENCES

[1] Anonymous, "From $f(x)$ and $g(x)$ to $f(g(x))$: LLMs Learn New Skills in RL by Composing Old Ones," arXiv:2509.25123, 2024.

[2] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML, 2021.

[3] R. J. Williams, "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning," Machine Learning, 1992.

[4] S. Antol et al., "VQA: Visual Question Answering," ICCV, 2015.

[5] J. Li et al., "BLIP-2: Bootstrapping Language-Image Pre-training," ICML, 2023.

[6] J. Johnson et al., "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning," CVPR, 2017.

[7] B. M. Lake et al., "Building Machines That Learn and Think Like People," Behavioral and Brain Sciences, 2017.

[8] C. Keysers et al., "Measuring Compositional Generalization," ICLR, 2020.