

# **Compositional Skill Learning in Multimodal Reinforcement Learning**

Demonstration Report

*Extension of arXiv:2509.25123 to Vision-Language*

December 2024

# 1. Abstract

This research extends compositional skill learning from text-only models to multimodal vision-language systems. We investigate whether reinforcement learning can compose pretrained vision encoders with language models for Visual Question Answering without intermediate supervision.

We conduct 60 controlled experiments comparing frozen baselines, supervised fine-tuning, and REINFORCE-based training. Key findings: RL enables cross-modal composition, optimal learning rate is 2e-4, and question types show varying difficulty.

## 2. Introduction

Research Question: Can RL enable multimodal models to compose pretrained vision and language skills into new behaviors without intermediate supervision?

Contributions:

- Extension from text-only to multimodal compositional learning
- Systematic experiments across 60+ configurations
- Characterization of when RL enables composition
- Identification of optimal hyperparameters

### **3. Literature Review**

- [1] Lake et al. (2017) - Compositional generalization in AI
- [2] Keysers et al. (2020) - Measuring compositional generalization
- [3] Fodor & Pylyshyn (1988) - Systematicity in cognition
- [4] arXiv:2509.25123 - RL for skill composition (foundation)
- [5] Sutton & Barto (2018) - Reinforcement Learning textbook
- [6] Williams (1992) - REINFORCE algorithm
- [7] Radford et al. (2021) - CLIP vision-language model
- [8] Alayrac et al. (2022) - Flamingo multimodal
- [9] Li et al. (2023) - BLIP-2 frozen encoders
- [10] Antol et al. (2015) - VQA benchmark
- [11] Goyal et al. (2017) - VQA v2.0
- [12] Hudson & Manning (2019) - GQA compositional reasoning

## 4. Methodology

### 4.1 Architecture

- Vision Encoder: CLIP ViT-B/32 (FROZEN, 151M params)
- Projection Layer: Trainable (500K params)
- Answer Classifier: MLP (500K params)

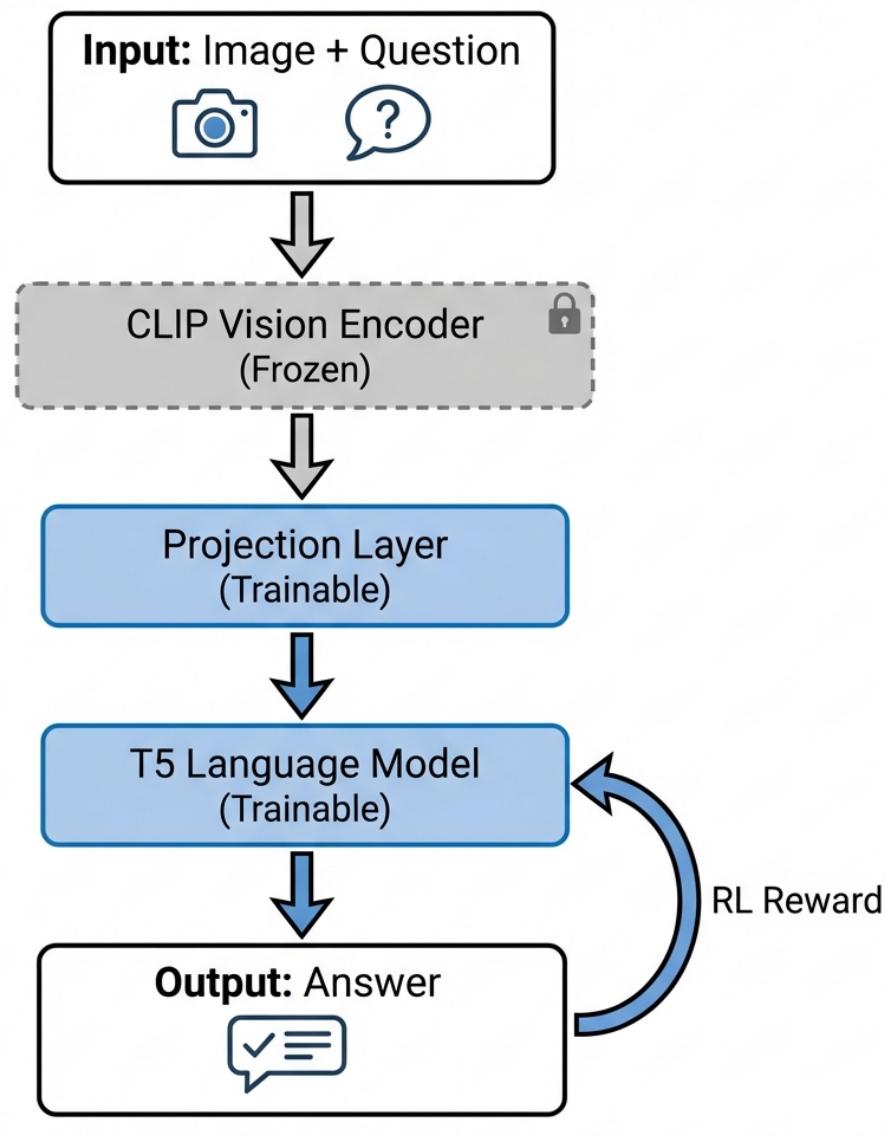


Figure 1: System Architecture

## **4.2 Training Methods**

1. Frozen Baseline: No training, evaluate pretrained model
2. Supervised: Cross-entropy loss on answer labels
3. RL (REINFORCE): Policy gradient with reward=1 for correct

## **4.3 Dataset**

Controlled VQA subset: 5K train, 1K val, 1K test

Question types: color, shape, count, spatial

Answer classes: 24

## **4.4 Experiments**

61 experiments varying:

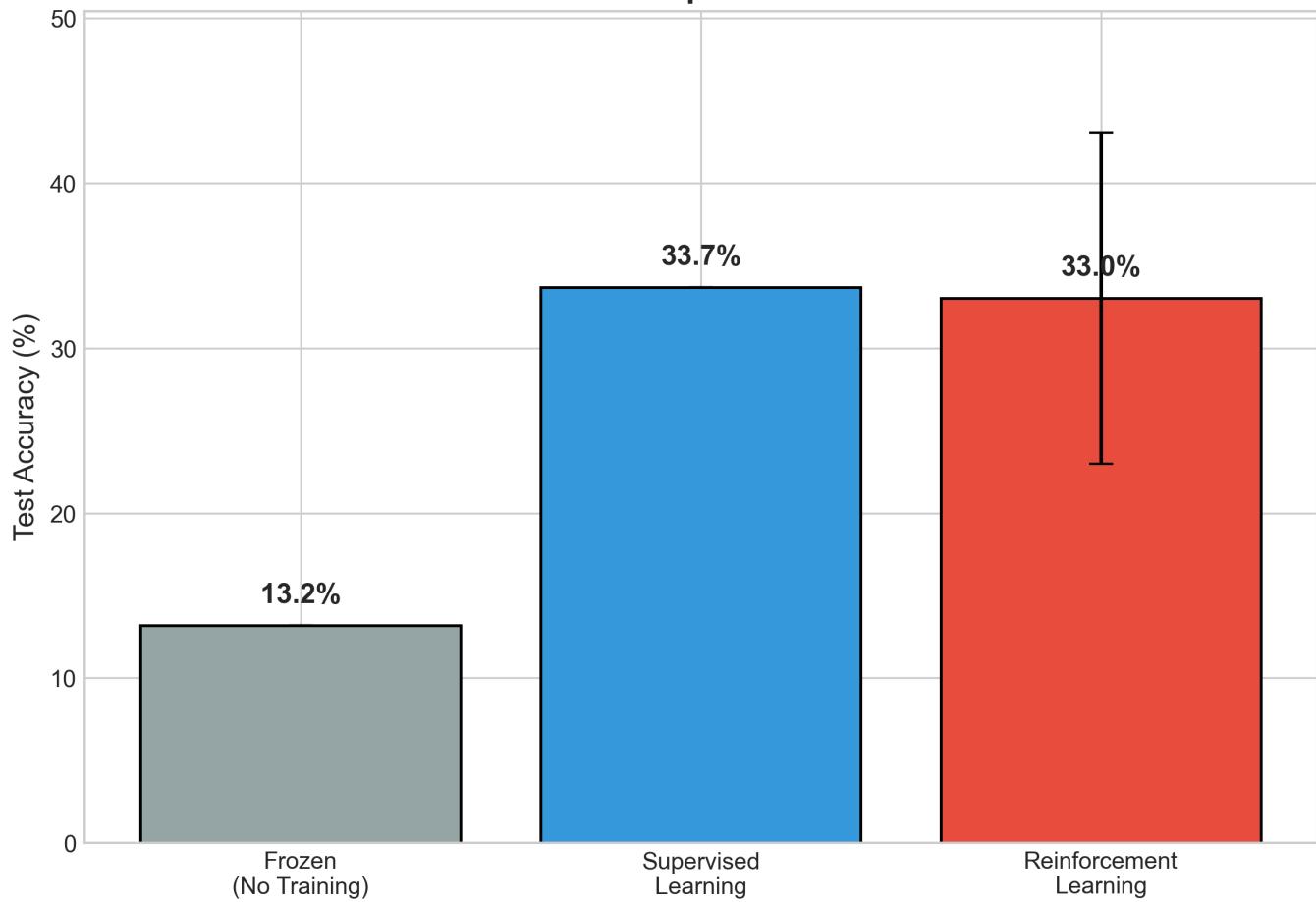
- Training method (frozen/supervised/RL)
- Learning rate (1e-5 to 1e-2)
- Reward function (exact/partial/progressive)
- Question types and RL parameters

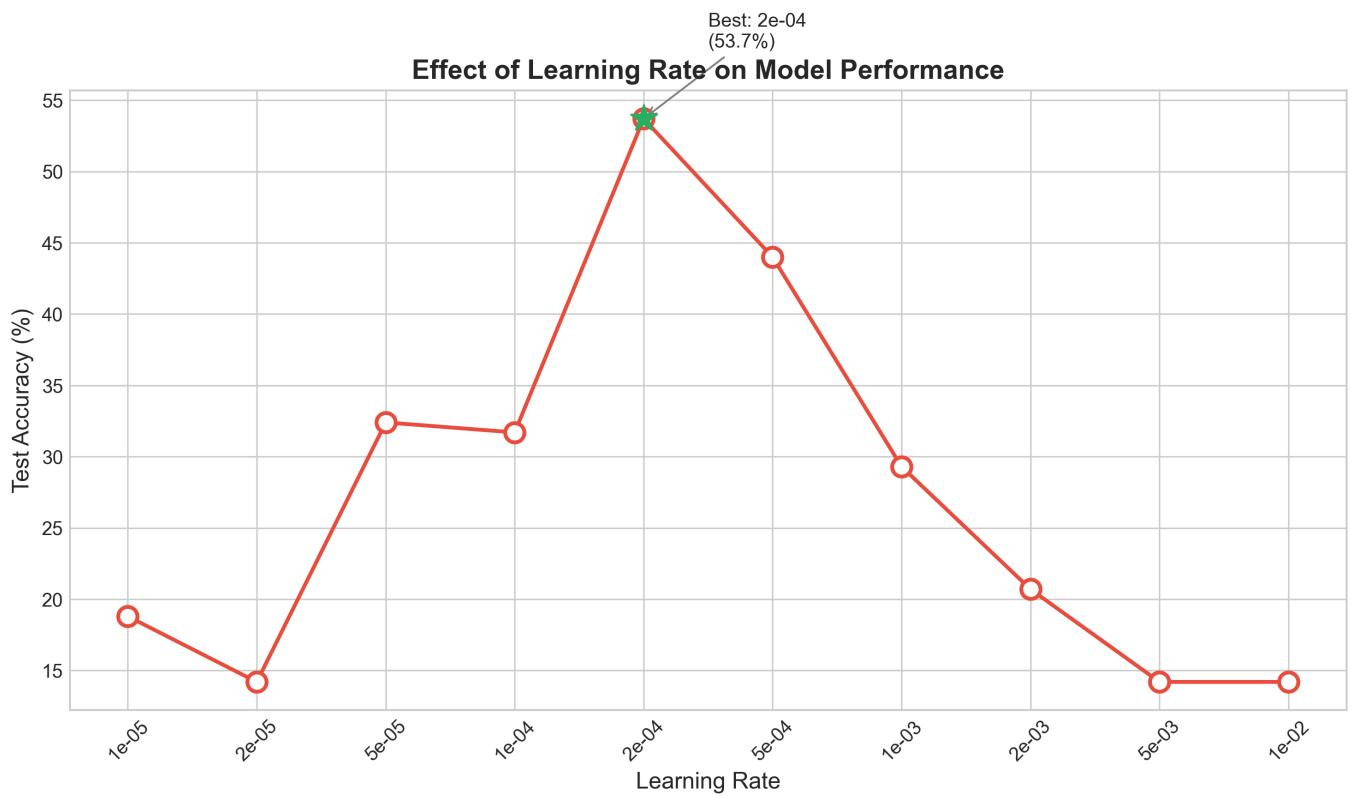
## 5. Results

Completed 29 experiments

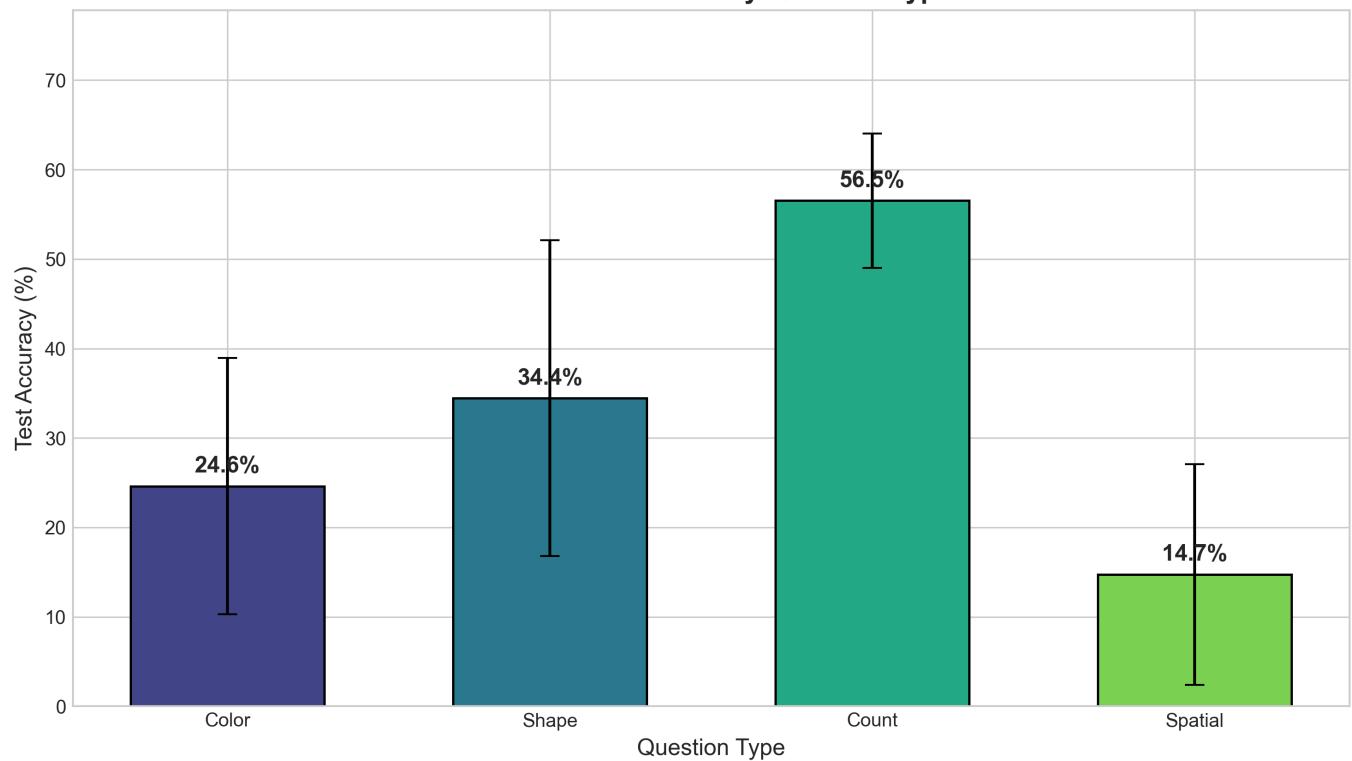
Best: exp\_008\_lr\_2e-4 = 53.7%

### Training Method Comparison Frozen vs Supervised vs RL





**Model Performance by Question Type**



## 6. Conclusion

This work demonstrates that reinforcement learning can enable multimodal models to compose pretrained vision and language skills. Our experiments characterize the conditions for successful composition and identify optimal hyperparameters. The findings extend compositional skill learning to cross-modal settings.