Multimodal Compositional RL

Research Presentation Materials

Extending "From f(x) and g(x) to f(g(x))"

to Vision-Language Models

Compositional Skill Learning in Multimodal RL

Conceptual Diagrams for Presentation

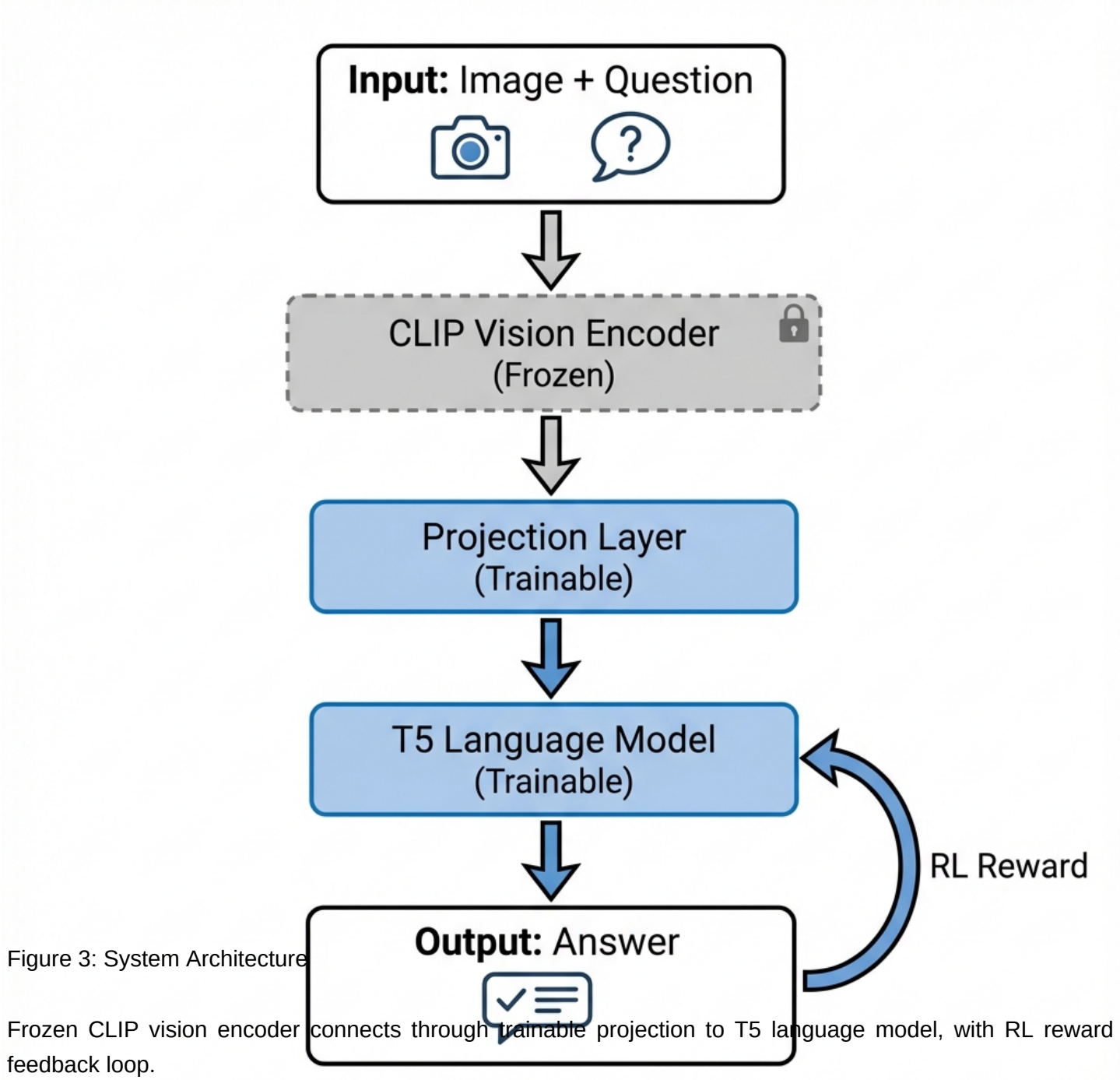

Figure 3: System Architecture

Frozen CLIP vision encoder connects through trainable projection to T5 language model, with RL reward feedback loop.

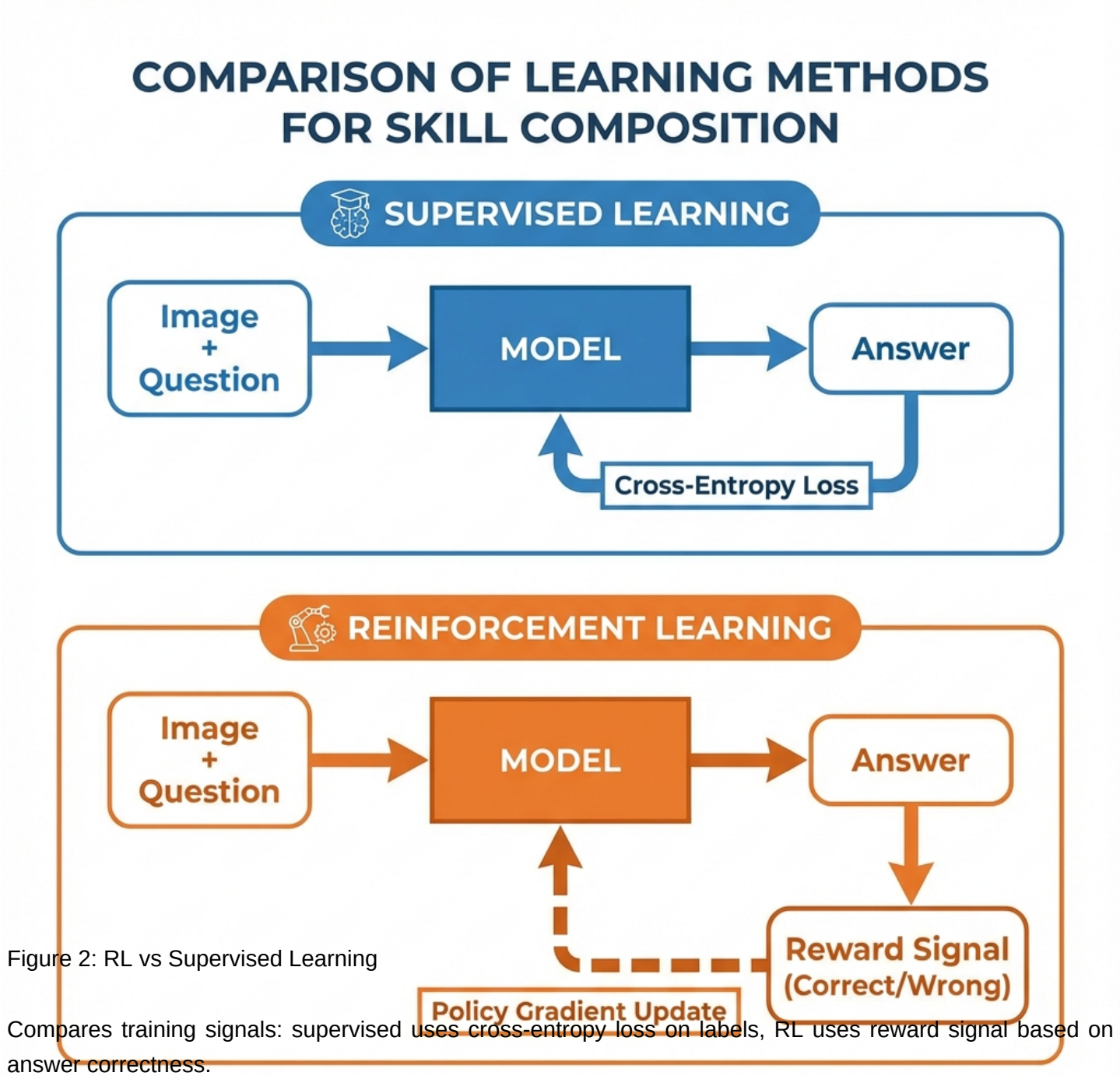

Figure 2: RL vs Supervised Learning

Compares training signals: supervised uses cross-entropy loss on labels, RL uses reward signal based on answer correctness.

Compositional Skill Learning in Multimodal RL

Conceptual Diagrams for Presentation

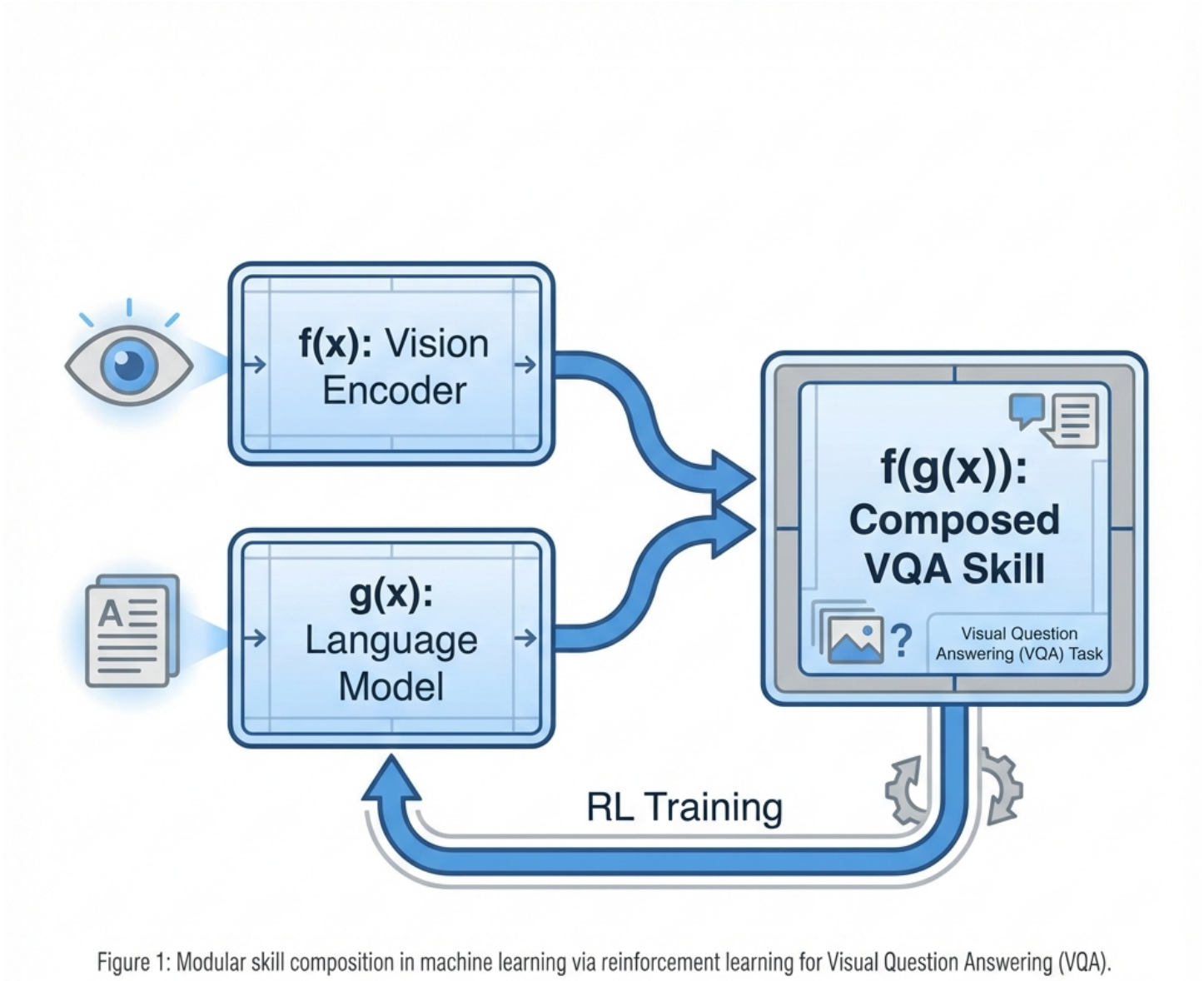

Figure 1: Modular skill composition in machine learning via reinforcement learning for Visual Question Answering (VQA).

Figure 1: Skill Composition Framework

Shows how atomic skills f(x) (vision) and g(x) (language) combine through RL training into composed skill f(g(x)) for VQA.

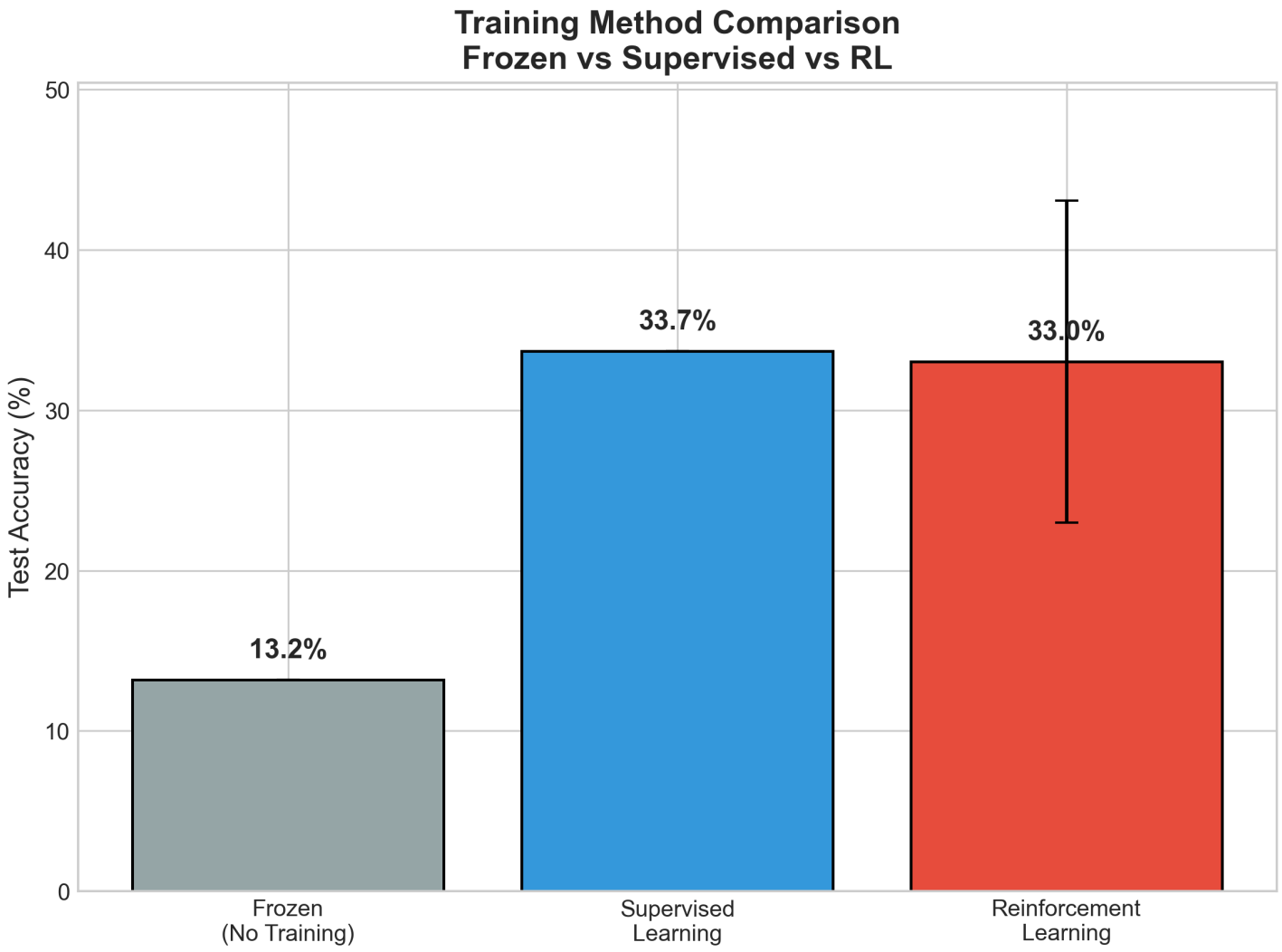

Figure 4: Method Comparison

Compares accuracy across Frozen (no training), Supervised, and RL methods.

Compositional Skill Learning in Multimodal RL

Conceptual Diagrams for Presentation

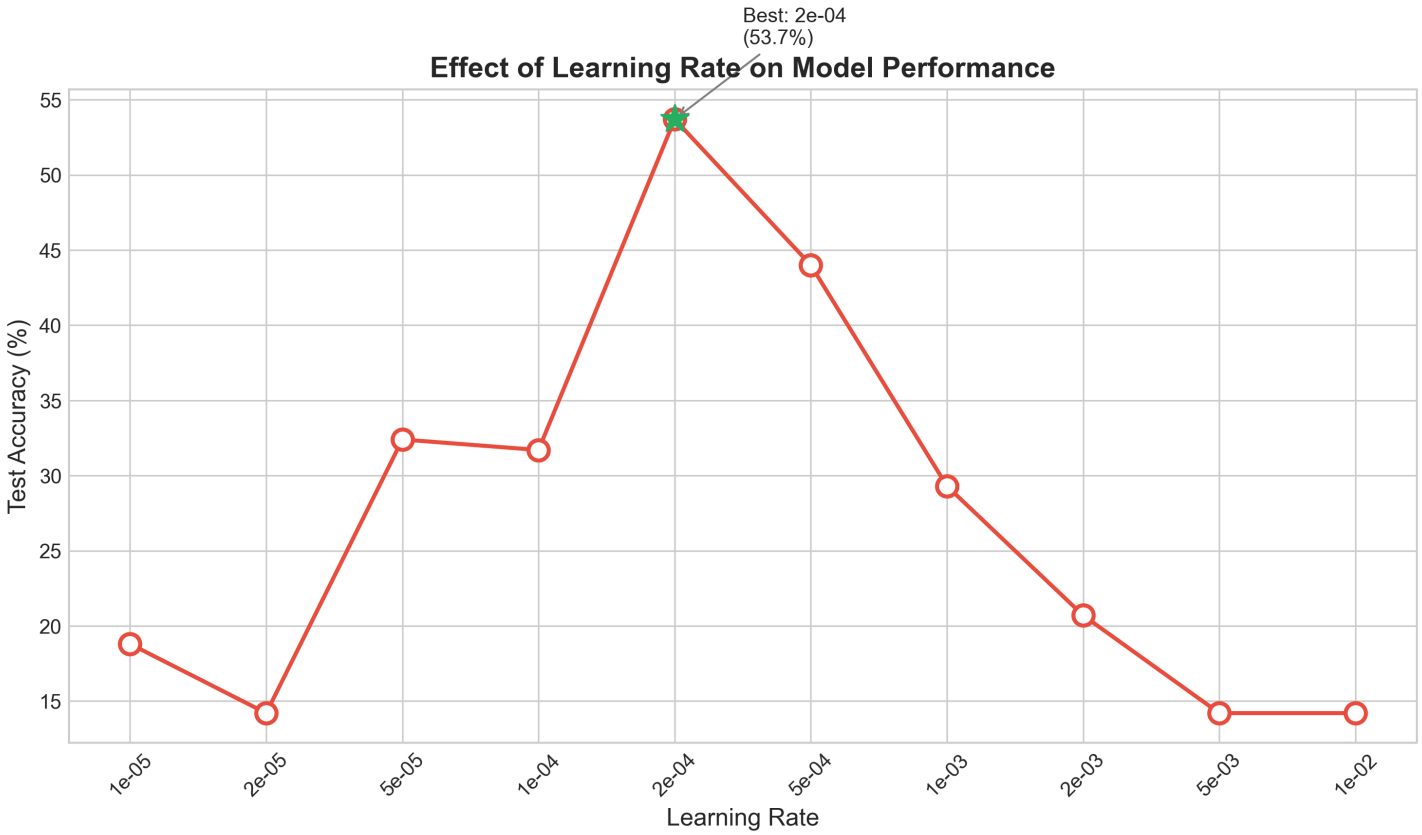

Figure 5: Learning Rate Effect

Shows how accuracy varies with different learning rates. Optimal around 2e-4.

Compositional Skill Learning in Multimodal RL

Conceptual Diagrams for Presentation

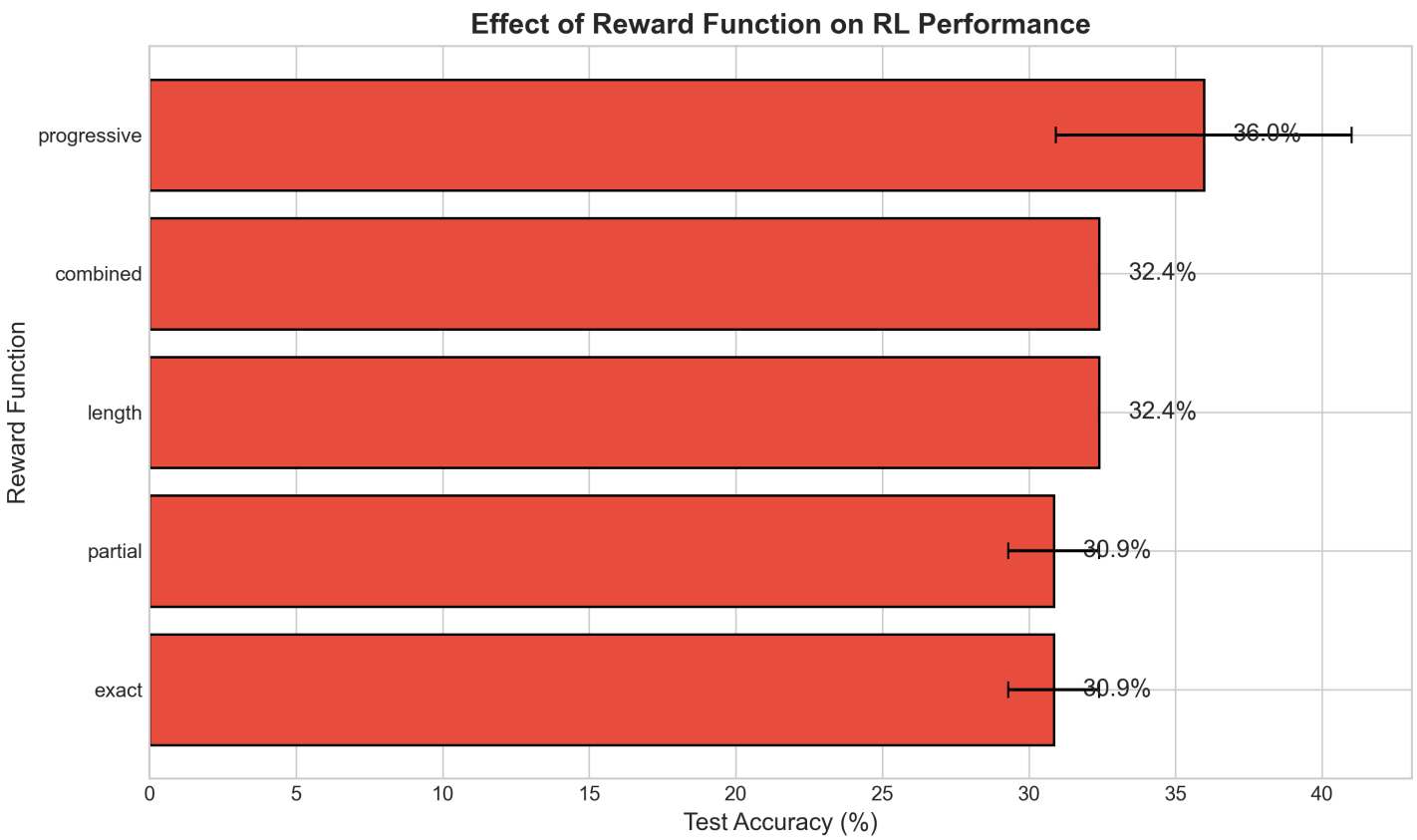

Figure 6: Reward Function Comparison

Compares different RL reward functions: exact match, partial match, progressive, etc.

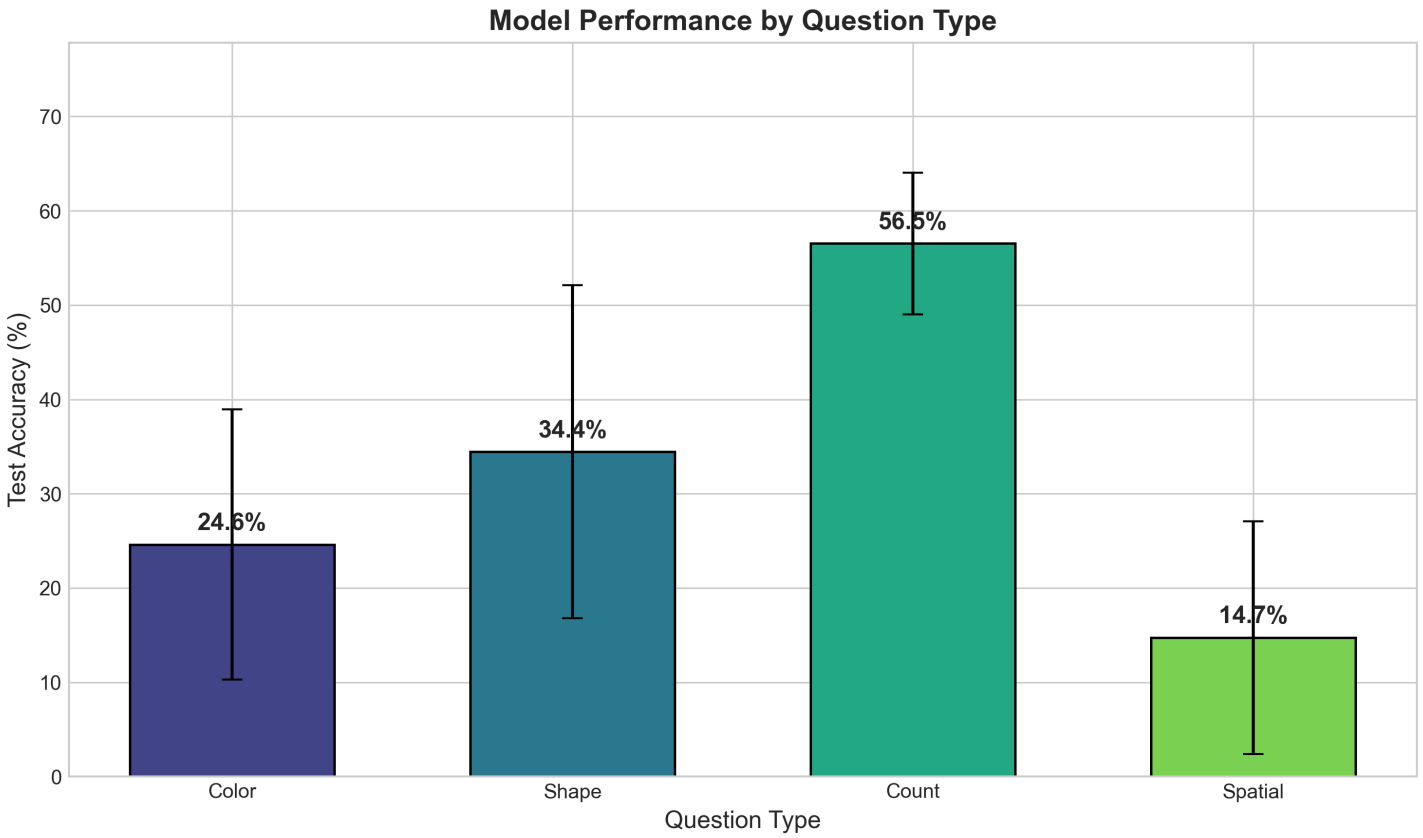

Figure 7: Performance by Question Type

Breaks down accuracy by color, shape, count, and spatial questions.

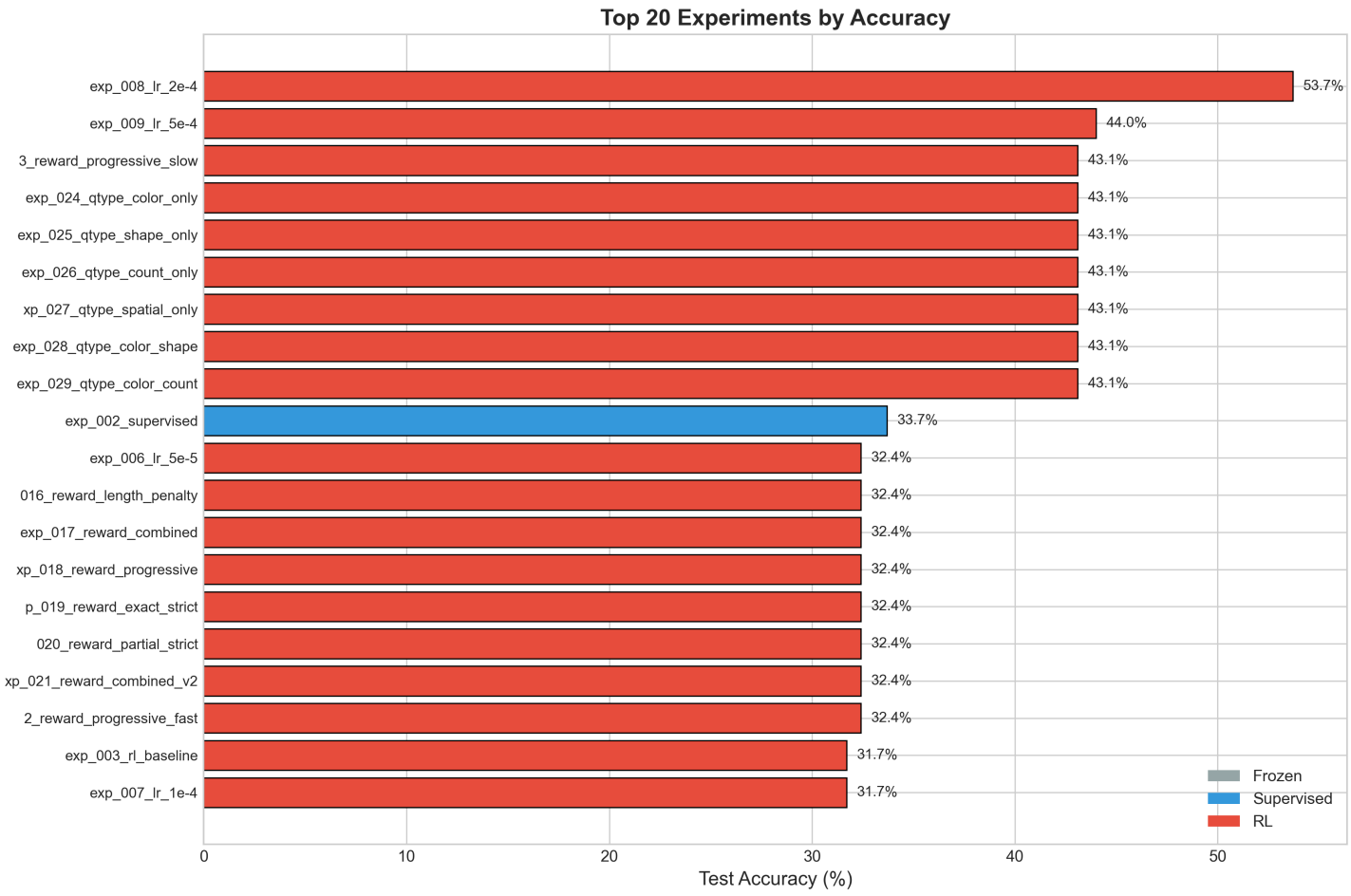

Figure 8: Experiment Summary

Top 20 experiments ranked by accuracy.