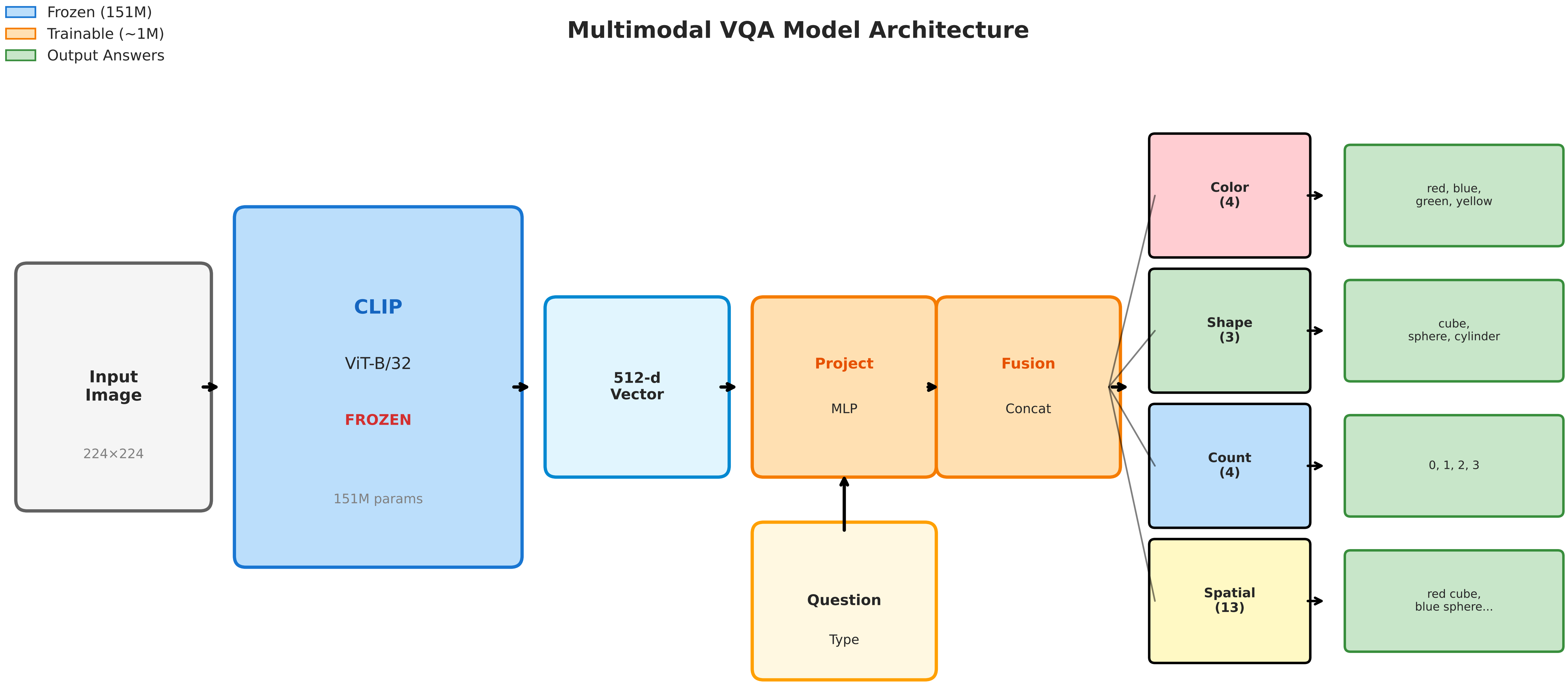
Multimodal VQA Model Architecture
Frozen (151M)
Trainable (~1M)
Output Answers
Input Image
224×224
CLIP
ViT-B/32
FROZEN
151M params
512-d Vector
Project
MLP
Question
Type
Fusion
Concat
Color (4)
red, blue, green, yellow
Shape (3)
cube, sphere, cylinder
Count (4)
0, 1, 2, 3
Spatial (13)
red cube, blue sphere...