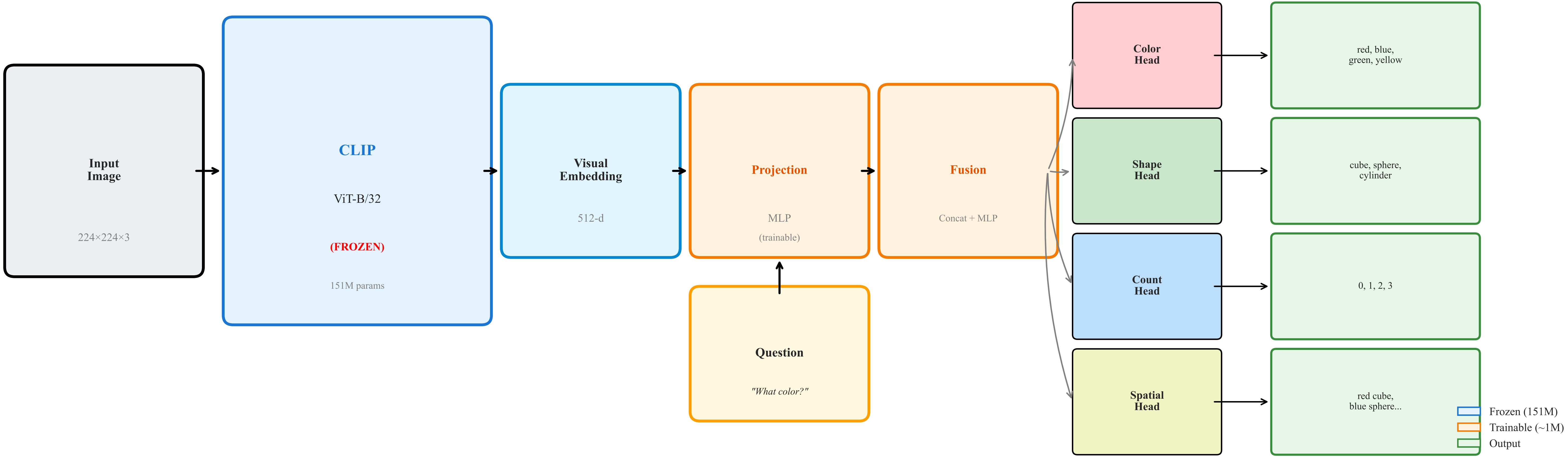
Multimodal VQA Architecture

CLIP (Frozen) + Trainable Projection + Question-Type-Specific Heads


Input Image
224×224×3
CLIP
ViT-B/32
(FROZEN)
151M params
Visual Embedding
512-d
Projection
MLP
(trainable)
Question
"What color?"
Fusion
Concat + MLP
Color Head
Shape Head
Count Head
Spatial Head
red, blue, green, yellow
cube, sphere, cylinder
0, 1, 2, 3
red cube, blue sphere...
Frozen (151M)
Trainable (~1M)
Output