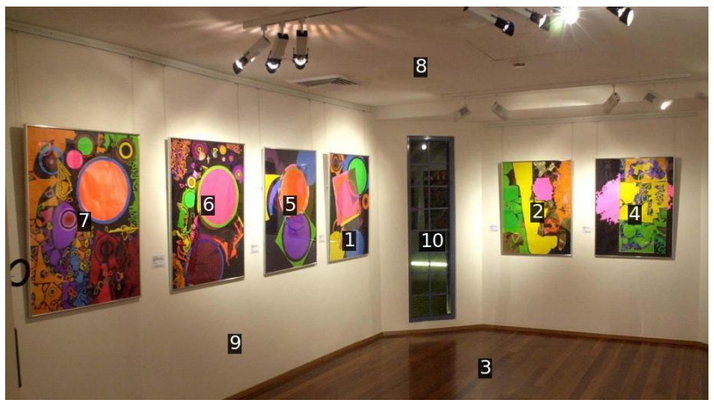
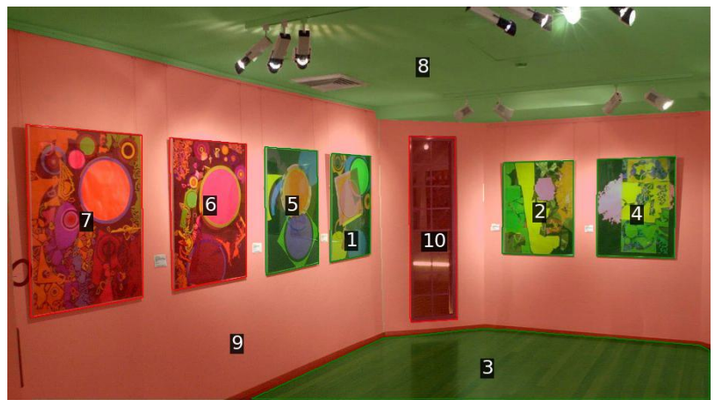
Input Image



Ground truths

1. Painting
2. Painting
3. Floor
4. Painting
5. Painting
6. Painting
7. Painting
8. Ceiling
9. Wall
10. Window

t=0 (base predictions)

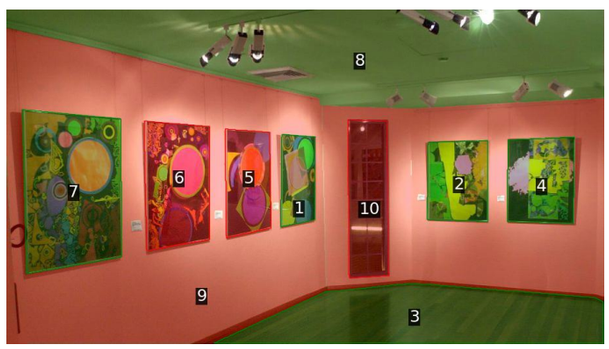


Initial predictions

1. Painting
2. Painting
3. Floor
4. Painting
5. Painting
6. Sculpture
7. Sculpture
8. Ceiling
9. Floor
10. Door

Intrinsic Self-Correction

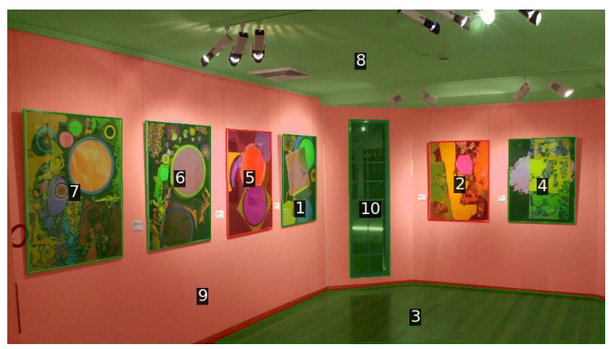
t=1



Revised predictions

1. Painting
2. Painting
3. Floor
4. Painting
5. Sculpture
6. Sculpture
7. Painting
8. Ceiling
9. Painting
10. Door

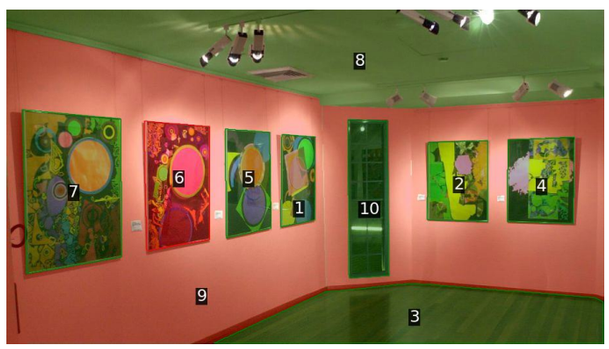
t=2



Revised predictions

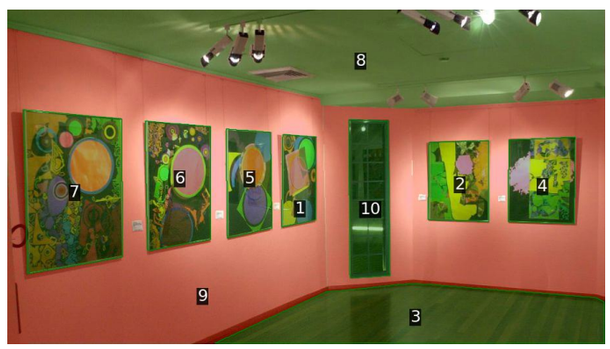
1. Painting
2. Sculpture
3. Floor
4. Painting
5. Sculpture
6. Painting
7. Painting
8. Ceiling
9. Sculpture
10. Window

VLM Binary Verification (**Ours**)



Revised predictions

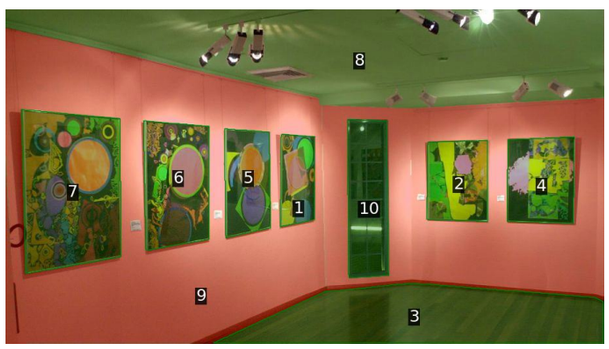
1. Painting
2. Painting
3. Floor
4. Painting
5. Painting
6. Sculpture
7. Painting
8. Ceiling
9. Floor
10. Window



Revised predictions

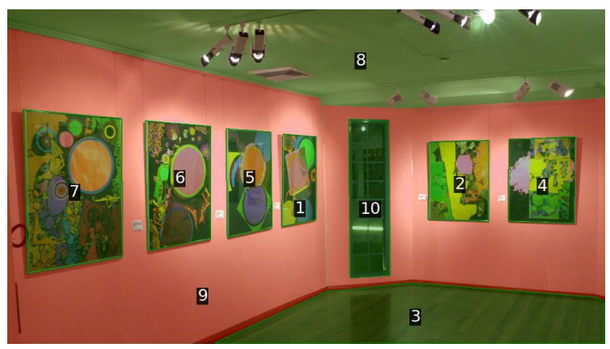
1. Painting
2. Painting
3. Floor
4. Painting
5. Painting
6. Painting
7. Painting
8. Ceiling
9. Floor
10. Window

Noise-Free



Revised predictions

1. Painting
2. Painting
3. Floor
4. Painting
5. Painting
6. Painting
7. Painting
8. Ceiling
9. Floor
10. Window



Revised predictions

1. Painting
2. Painting
3. Floor
4. Painting
5. Painting
6. Painting
7. Painting
8. Ceiling
9. Floor
10. Window