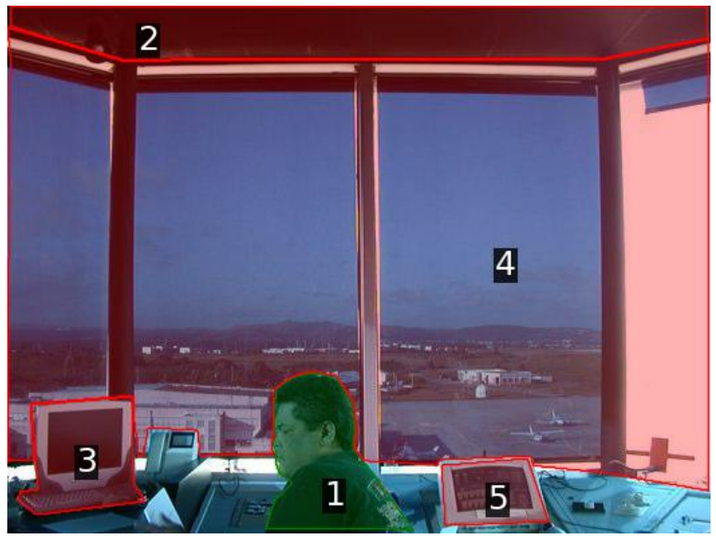
Input Image


Ground truths
1. Person
2. Ceiling
3. Computer
4. Window
5. Monitor

t=0 (base predictions)




Initial predictions
1. Person
2. Window
3. Monitor
4. Desk
5. Desk

---

Intrinsic Self-Correction

t=1


Revised predictions
1. Trash can
2. Window
3. Monitor
4. Desk
5. Desk

t=2

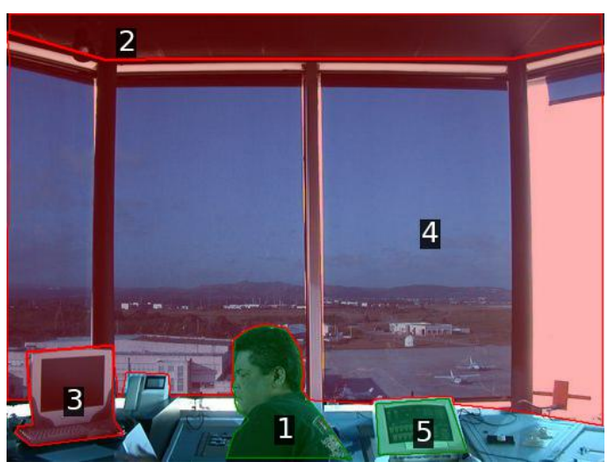

Revised predictions
1. Trash can
2. Window
3. Monitor
4. Monitor
5. Monitor

VLM Binary Verification (**Ours**)

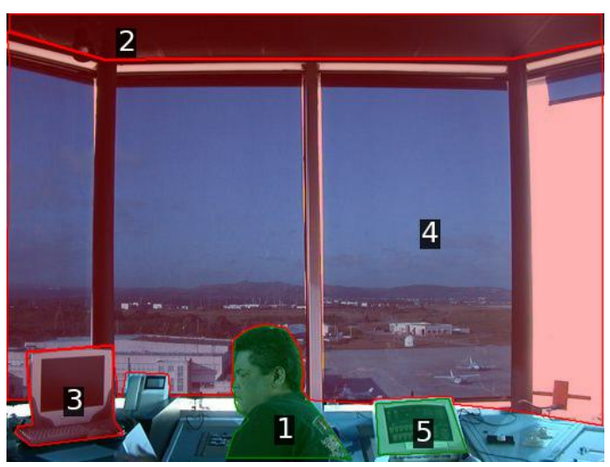

Revised predictions
1. Person
2. Window
3. Monitor
4. Window
5. TV


Revised predictions
1. Person
2. Window
3. Monitor
4. Window
5. Monitor

Noise-Free




Revised predictions
1. Person
2. Window
3. Monitor
4. Runway
5. Monitor




Revised predictions
1. Person
2. Window
3. Monitor
4. Runway
5. Monitor