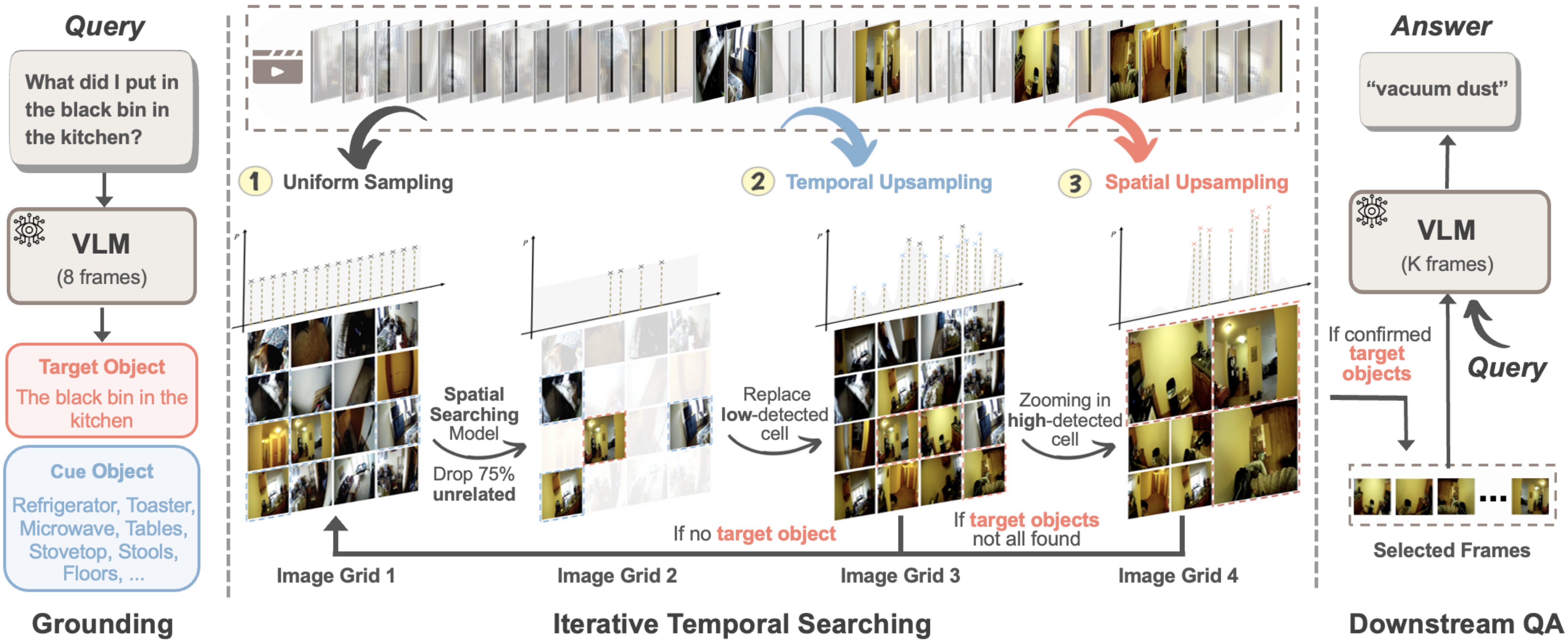
Query
What did I put in the black bin in the kitchen?
VLM
(8 frames)
Target Object
The black bin in the kitchen
Cue Object
Refrigerator, Toaster, Microwave, Tables, Stovetop, Stools, Floors, ...
Grounding
1 Uniform Sampling
2 Temporal Upsampling
3 Spatial Upsampling
Spatial Searching Model
Drop 75% unrelated
Replace low-detected cell
Zooming in high-detected cell
If no target object
If target objects not all found
Image Grid 1
Image Grid 2
Image Grid 3
Image Grid 4
Iterative Temporal Searching
Answer
"vacuum dust"
VLM
(K frames)
Query
If confirmed target objects
Selected Frames
Downstream QA