Q: What do the orange letters say on the bus

I_g:

$\oplus T_{plan}$

Stage 1: Planner

Input: Q、I_g、T_{plan}

Output: Detailed Plan is

Step1: Find object: bus

Step2: Find object: orange letters (conditioned on bus)

Step3: Read attribute: text content (conditioned on orange letters)

Stage 2: Perceiver&Verifier

I_g:

S_i: {Step $_n$}

SG_{prompt}: {Scene Graph Generation Prompt}

LMM

N_{cand}

{
"identity":"orange letters",
"bbox":[x1,y1,x2,y2],
"attributes":{"color":"orange","text_content":"SELECT"},
"relations":{"anchor_node":"node_01_bus","relationship":"attached to"}
}

Crop

I_{crop}:

Visual QA:
Does this cropped region provide clear visual evidence of {Target Object} with {Attribute}?

Level 1:
Expand BBox → Re-verify

Level 2:
Exclude → New Candidate

Level 3:
Reflection → Re-plan

Fail ≥ N

Correct

Stage 3: Response

User: Use the image and scene graph as context and answer the following question

I_g:

SG_{prompt}:

[{"identity": "bus", "bbox": "[x_bus1,y_bus1,x_bus2,y_bus2]", "attributes": {"color": "blue and white", "type": "public transit"}, "relations": {}}, {"identity": "orange letters", "bbox": "[x1,y1,x2,y2]", "attributes": {"color": "orange", "text_content": "SELECT"}, "relations": {"anchor_node": "bus", "relationship": "attached to"}}]

Q: What do the orange letters say on the bus?

E: Answer in the format specified by the answer to the question

P_{all} = "$[User][I_g][SG][Q][E]$"

LMM → Response: *Answer*