1.Text Grounder Module

Text Input:

Picture hanging directly above the laptop.

Image Input:

Text Grounder

Refined Query: *Picture.*

2. Candidate Positioning & Setting Marks Module

Refined Query: *Picture.*

Image Input

Candidate Positioning

Candidate Bounding Boxes

Setting Marks

3. Visual Grounder Module

Marked Image

Original Text Input

Visual Grounder

Final Target: [2]

Reference