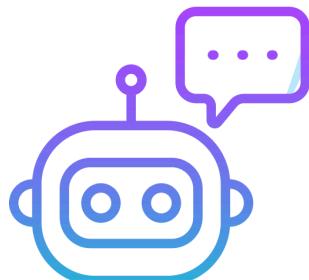


## Stage I: Coarse-grained Box Localization



MLLM  
(QwenVL, InternVL,  
LLaVA etc.)

## Stage II: Fine-grained Pixel Classification

### Scale-Aware Sampling

$$\begin{matrix} N_x \propto W \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ N_y \propto H \end{matrix}$$

Question: Candidate points  
[x, y]: {"points": [[333, 415],  
[333, 485], [333, 555] ...  
[596, 694], [596, 764]]}. For  
each point, output 1 if it lies  
on the target object,  
otherwise 0.



### Chain-of-Thought

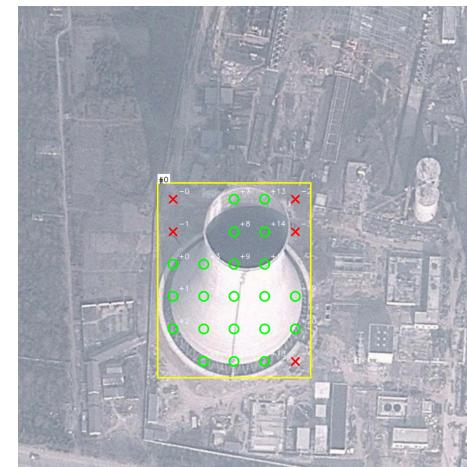
<think>The image contains a large structure that appears to be a chimney. It is cylindrical with a wide base and a narrower top, forming a frustum of a cone shape. The structure is white and located in the center of the image, surrounded by other industrial buildings.</think>



<answer> [292, 381, 639, 801] </answer>

### Chain-of-Thought

<think>Step 1: Identify the target object. The target object is a large frustum of a cone chimney, which is the large, white.....  
Step 2: Evaluate each point in relation to the target object. Point1 [333, 415]: Located above the visible part of the chimney, this point does not lie on the chimney's surface. The area around this coordinate seems to be either sky or another structure not part of the chimney. Decision: 0. Point 2: [333, 485]... Point3 [333, 555]: This point starts entering the region where the chimney's surface begins, possibly at the edge of its top. As it aligns with the chimney's structure, it can be considered part of the target object. Decision: 1 ...</think>



<answer>[0,0,1,1,1,0,0,0,  
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,  
1,1,0,0,1,1,1,0]</answer>

## Stage III: Prompting Segmentation

### SAM

Vision  
Backbone

Image

box prompt:  
[292, 381, 639, 801]

positive point: [  
[333, 555], ... ]

negative point:[  
[333, 415], ... ]

Decoder

