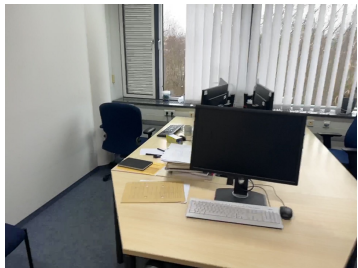


User Input



Q1: Given this image, generate its **metric depth map** (in meters).

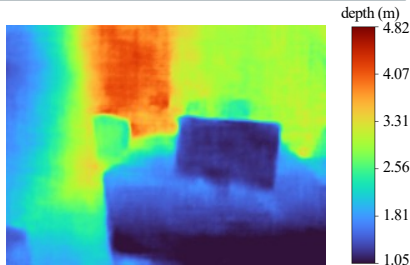


Q2: Which one is **closer** to me? The *black monitor* or the *office chair*?



Dense Geometry Prediction

Depth Map



Point Cloud



: Here is the metric depth map.



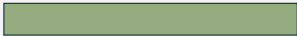
3D Spatial Reasoning



: The *black monitor* is closer.

Inference Time

(a 256×192 input image)

DepthLM  13h

Youtu-VL  2.48s

Ours  0.42s