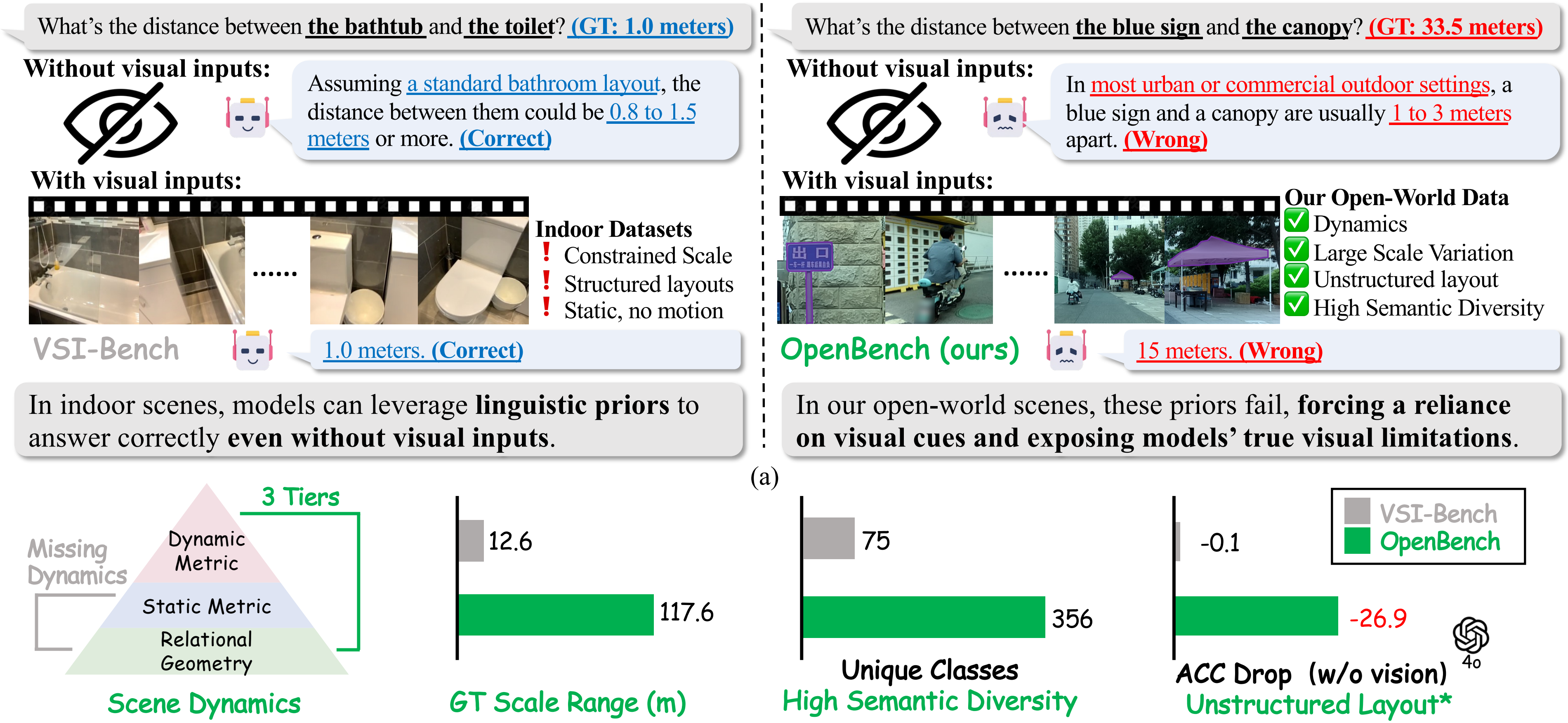

What's the distance between **the bathtub** and **the toilet**? (GT: 1.0 meters)

**Without visual inputs:**

Assuming a standard bathroom layout, the distance between them could be 0.8 to 1.5 meters or more. **(Correct)**

**With visual inputs:**

**Indoor Datasets**

! Constrained Scale

! Structured layouts

! Static, no motion

VSI-Bench

1.0 meters. **(Correct)**

In indoor scenes, models can leverage **linguistic priors** to answer correctly **even without visual inputs**.

What's the distance between **the blue sign** and **the canopy**? (GT: 33.5 meters)

**Without visual inputs:**

In most urban or commercial outdoor settings, a blue sign and a canopy are usually 1 to 3 meters apart. **(Wrong)**

**With visual inputs:**

**Our Open-World Data**

✓ Dynamics

✓ Large Scale Variation

✓ Unstructured layout

✓ High Semantic Diversity

OpenBench (ours)

15 meters. **(Wrong)**

In our open-world scenes, these priors fail, **forcing a reliance on visual cues and exposing models' true visual limitations.**

(a)

(b)

(c)

(d)

(e)