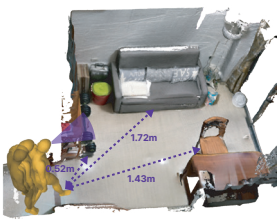
(a) Previous Methods

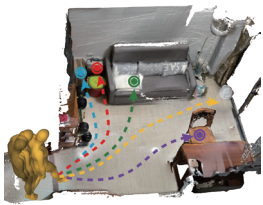
Gaze/Distance based Methods



No explicit interaction target
Geometry alone is ambiguous
Gaze requires extra sensing

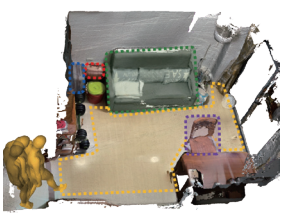
(b) Ours (SGIP)

Ambiguous future grounding

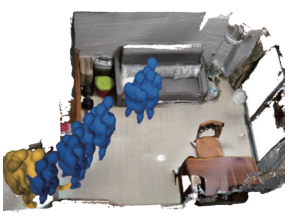


Where to go?

Explicit Interaction Prior



Scene-Grounded Prediction



Explicit interaction prior
Semantic+geometric grounding
No extra sensing at inference