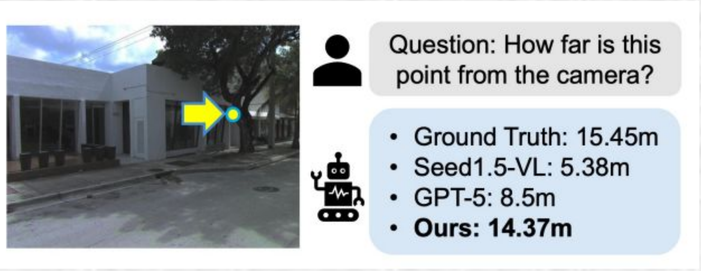
DepthLM: Metric Depth From Vision Language Models (Oral)

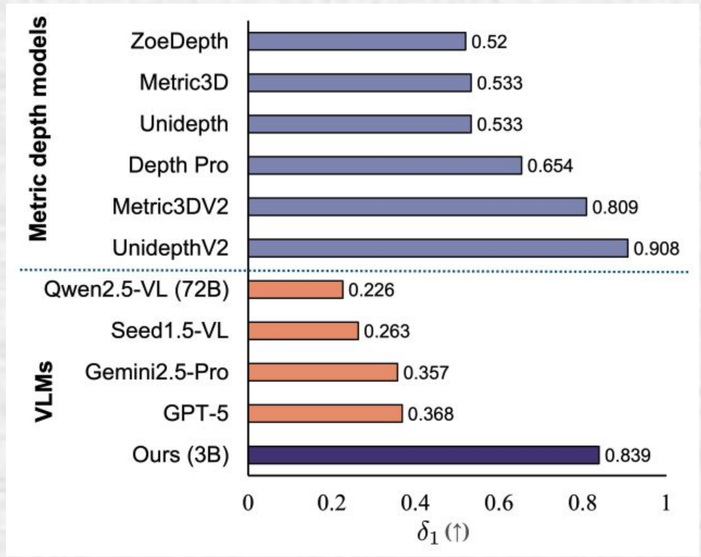
Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory P. Meyer , Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, Yangyang Shi

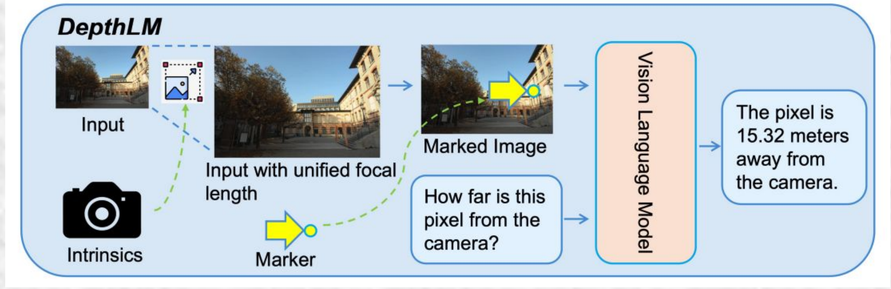
Meta

PRINCETON UNIVERSITY

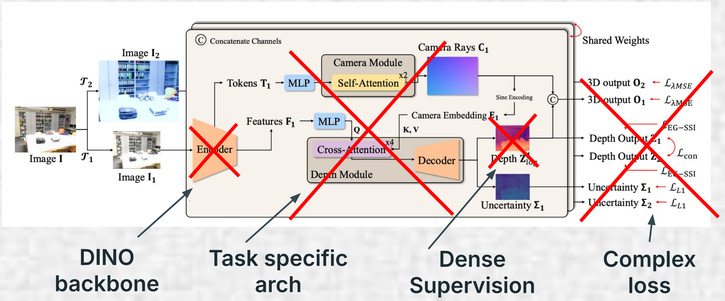
No more task-specific models, just standard VLMs!



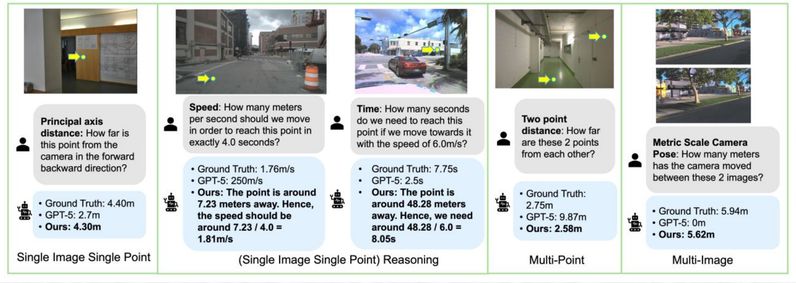




- **Simple**
  - 1 pixel * 7M iamges



- **Scalable**



δ_1(↑) of different methods	*Out* DDAD	*Out* Nuscenes	*Out+In* ETH3D	*In* sunRGBD	*In* ibims1	vs Ours (↑)
ZoeDepth	0.272	0.283	0.350	0.867	0.580	-42.8%
DepthAnything	-	0.354	0.093	0.850	0.714	-40.3%
DepthAnythingV2	-	0.171	0.363	0.724	-	-48.5%
Metric3D	-	0.723	0.456	0.154	0.797	-36.6%
Unidepth	0.858	0.846	0.185	0.943	0.157	-27.3%
Depth Pro	0.299	0.566	0.397	0.831	0.823	-29.1%
Metric3Dv2	-	0.841	0.900	0.812	0.684	-3.8%
UnidepthV2	0.882	0.870	0.852	0.964	0.945	+9.2%
Ours (7B)	0.747	0.865	0.718	0.859	0.920	-

Contact