

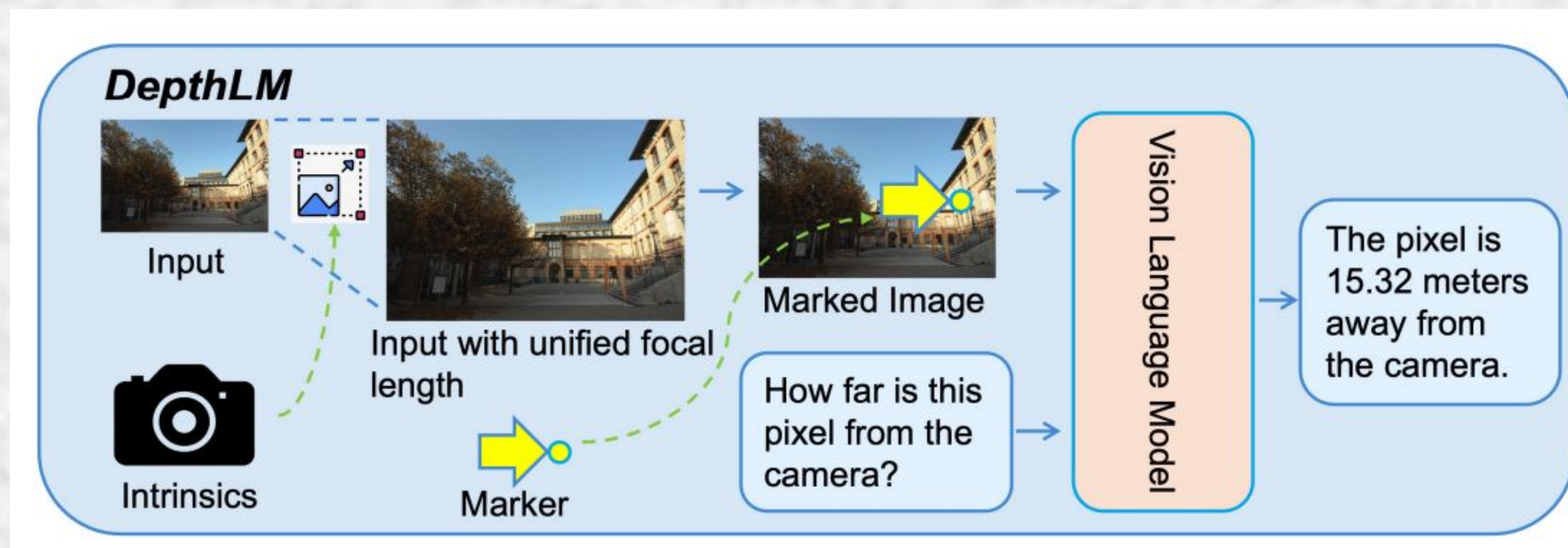
DepthLM: Metric Depth From Vision Language Models

(Oral)



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

No more task-specific models,
just standard VLMs!

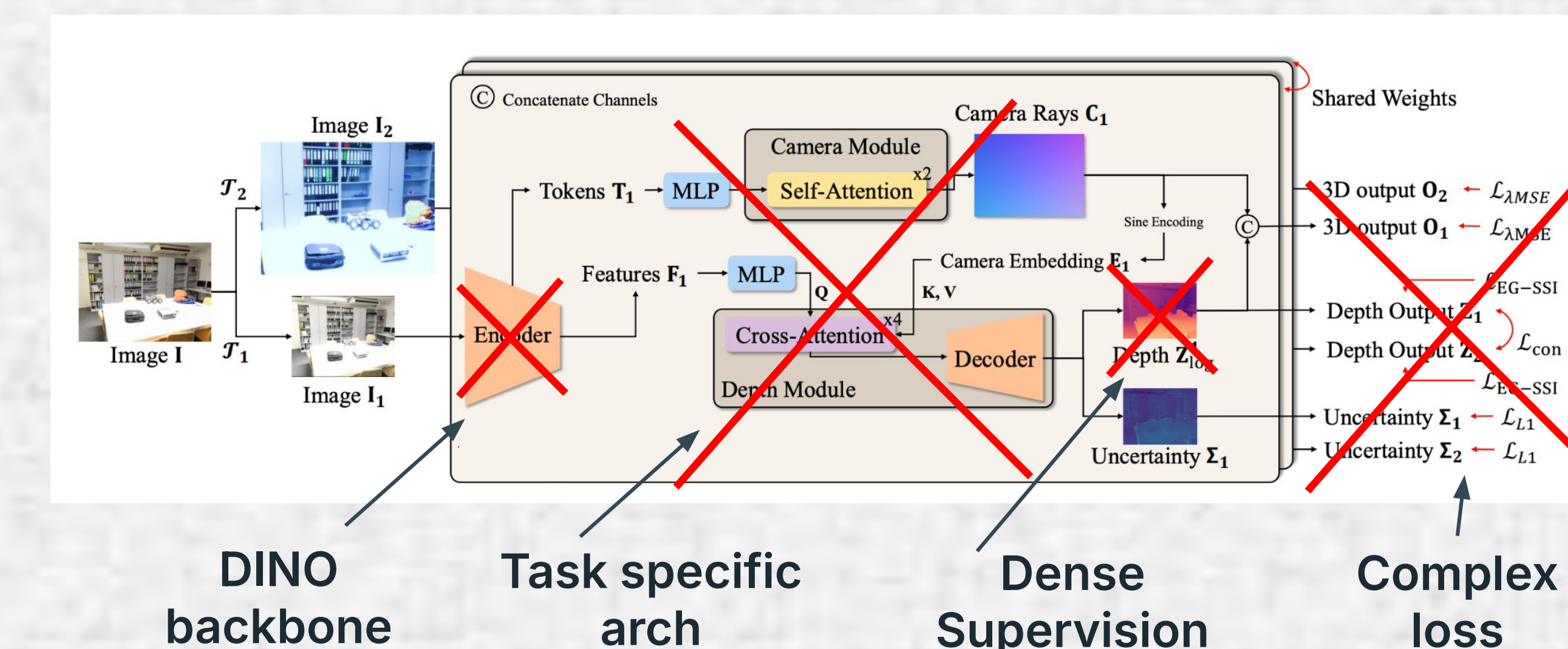


Question: How far is this point from the camera?

- Ground Truth: 15.45m
- Seed1.5-VL: 5.38m
- GPT-5: 8.5m
- Ours: 14.37m

Simple

1 pixel * 7M images



Scalable

Principal axis distance: How far is this point from the camera in the forward backward direction?

• Ground Truth: 4.40m
• GPT-5: 2.7m
• Ours: 4.30m

Speed: How many meters per second should we move in order to reach this point in exactly 4.0 seconds?

• Ground Truth: 1.76m/s
• GPT-5: 250m/s
• Ours: The point is around 7.23 meters away. Hence, the speed should be around 7.23 / 4.0 = 1.81m/s

Time: How many seconds do we need to reach this point if we move towards it with the speed of 6.0m/s?

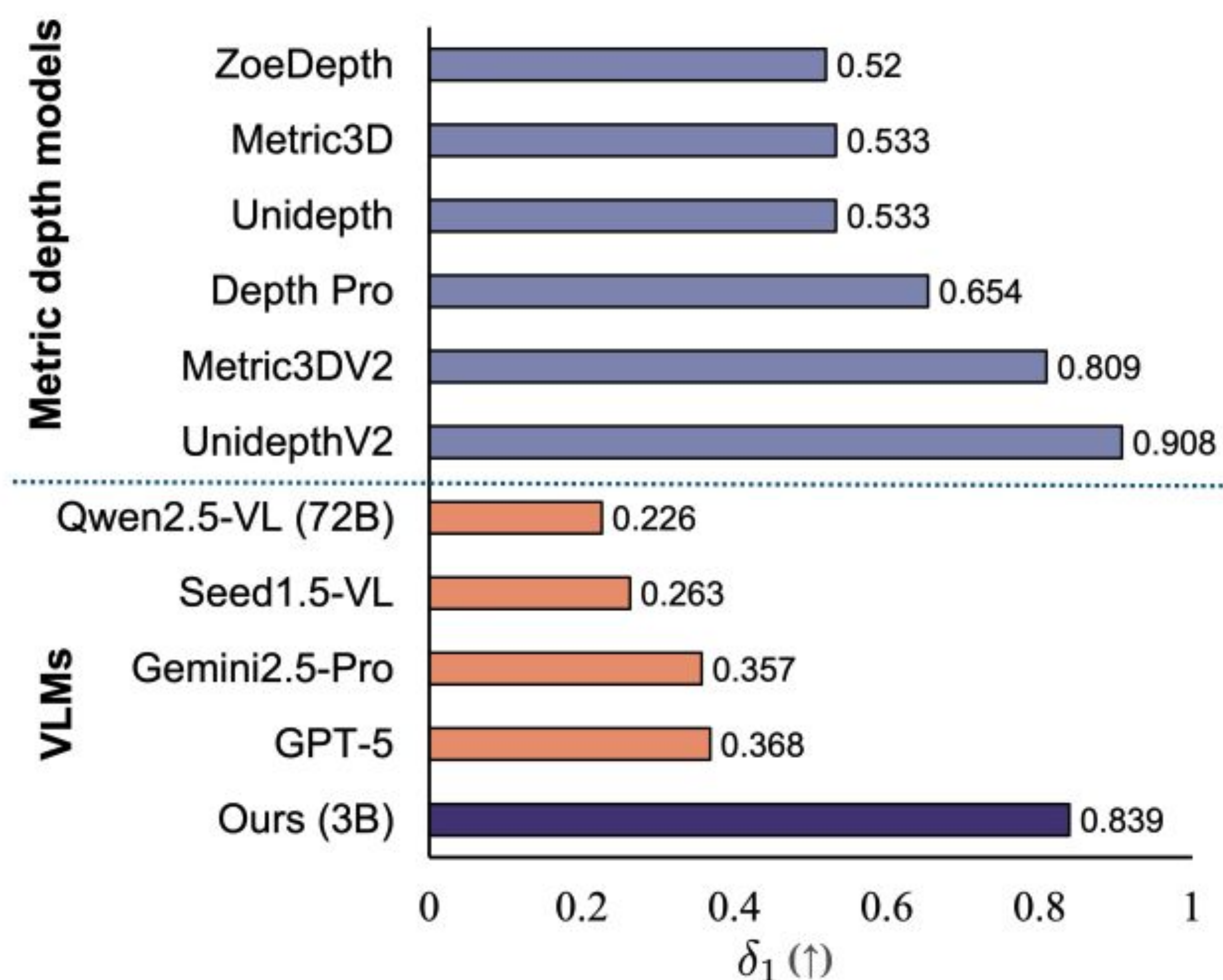
• Ground Truth: 7.75s
• GPT-5: 2.5s
• Ours: The point is around 48.28 meters away. Hence, we need around 48.28 / 6.0 = 8.05s

Two point distance: How far are these 2 points from each other?

• Ground Truth: 2.75m
• GPT-5: 9.87m
• Ours: 2.58m

Metric Scale Camera Pose: How many meters has the camera moved between these 2 images?

• Ground Truth: 5.94m
• GPT-5: 0m
• Ours: 5.62m



$\delta_1 (\uparrow)$ of different methods	Out		Out+In	In		vs Ours (\uparrow)
	DDAD	Nuscenes	ETH3D	sunRGBD	ibims1	
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	-

