DepthLM

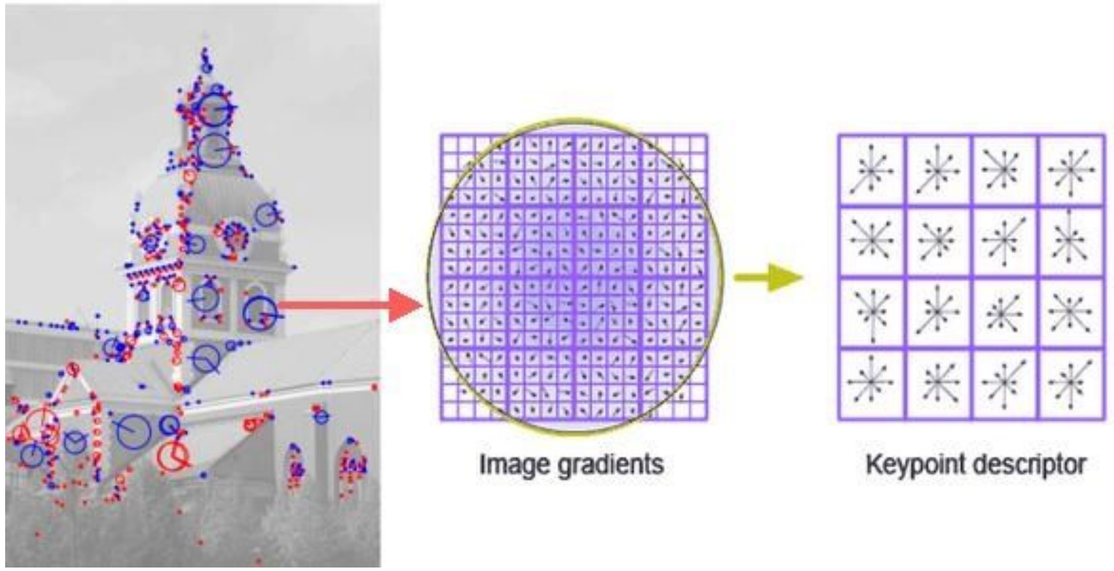
Metric Depth From Vision Language Models

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

Meta PRINCETON UNIVERSITY

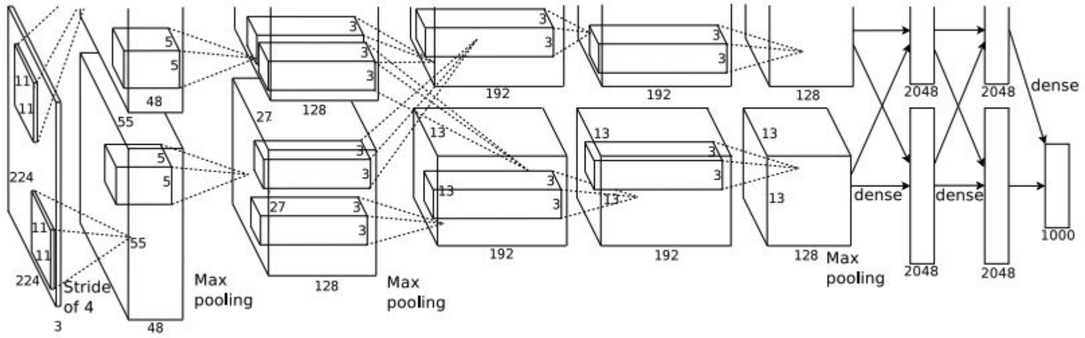
Code & Model

History

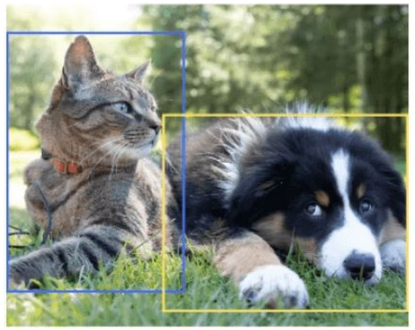


Lowe et al., 2004

History

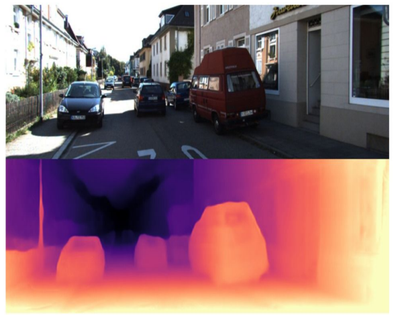


classification



detection

segmentation

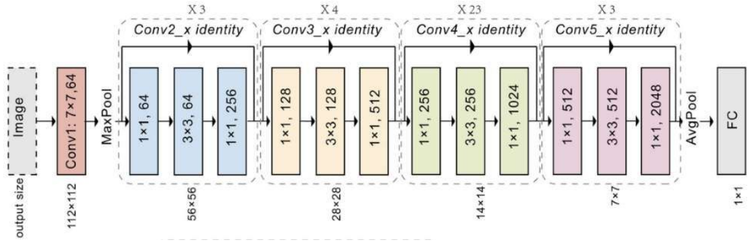


depth

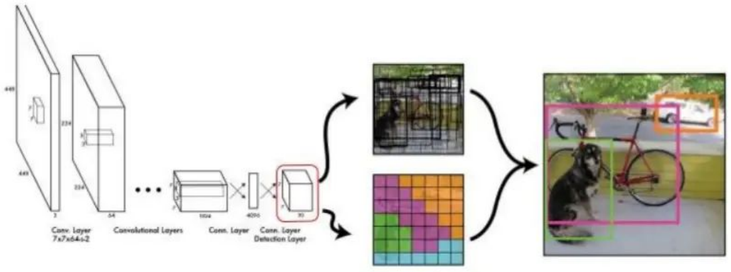
Krizhevsky et al., 2012

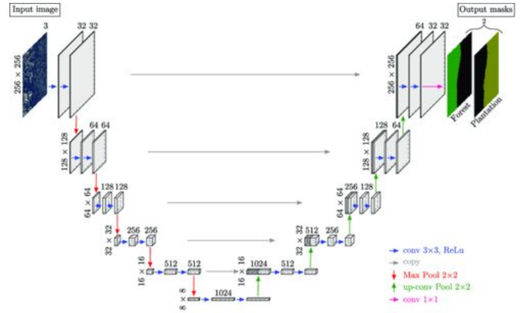
History

Classification
(He et al. 2016)

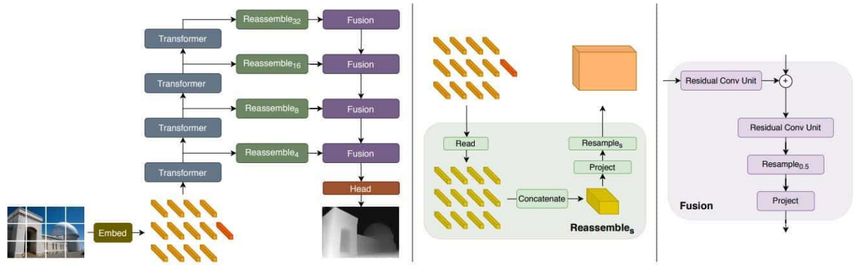


Detection
(Redmon et al. 2016)



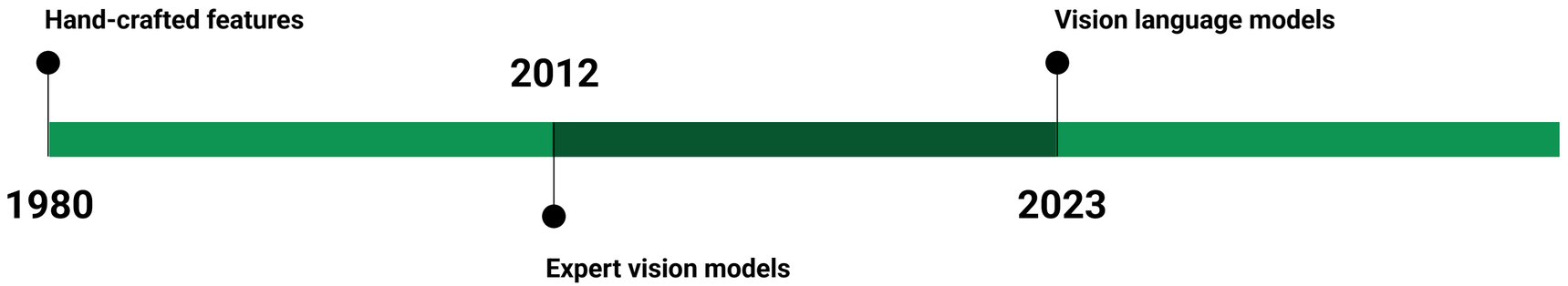


Segmentation
(Ronneberger et al. 2015)

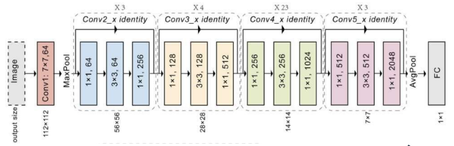


Depth
(Ranftl et al. 2012)

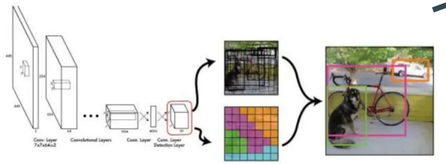
History

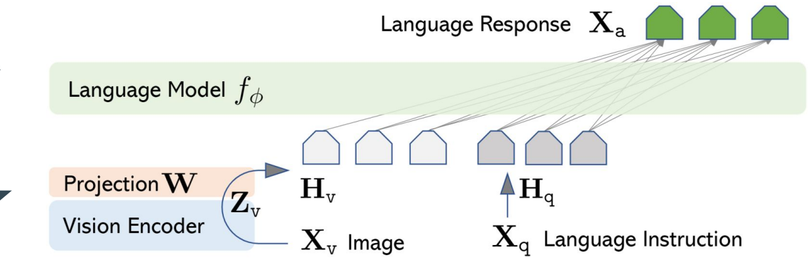


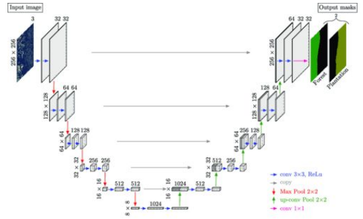
Classification
(He et al. 2016)



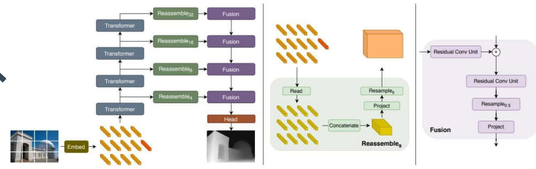
Detection
(Redmon et al. 2016)







Segmentation
(Ronneberger et al. 2015)



Depth
(Ranftl et al. 2012)

Liu et al., 2023

3D understanding is still dominated by expert vision models

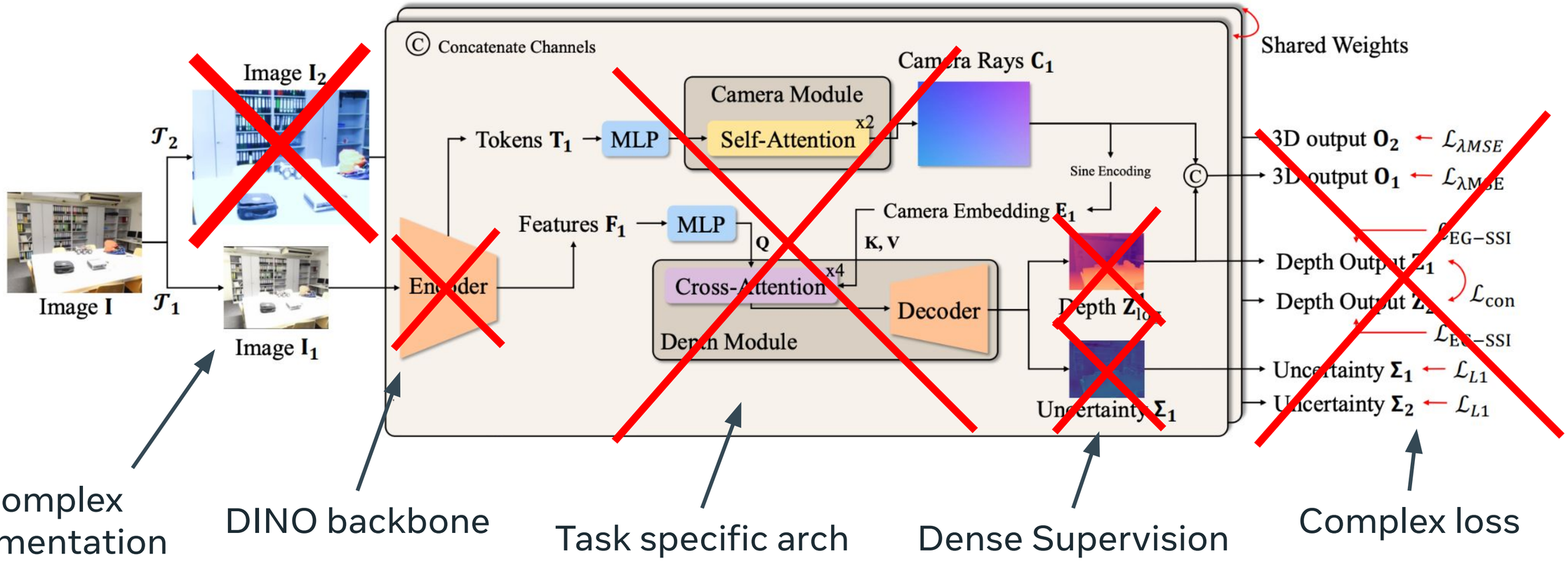
- Depth Estimation:
  - Metric3D, Unidepth, Moge, DepthAnything
- Pose Estimation:
  - VGGT, MapAnything
- Pixel Correspondence:
  - DKM, RoMa
- ...

Key Message

VLMs can learn 3D understanding as accurate as expert vision models.

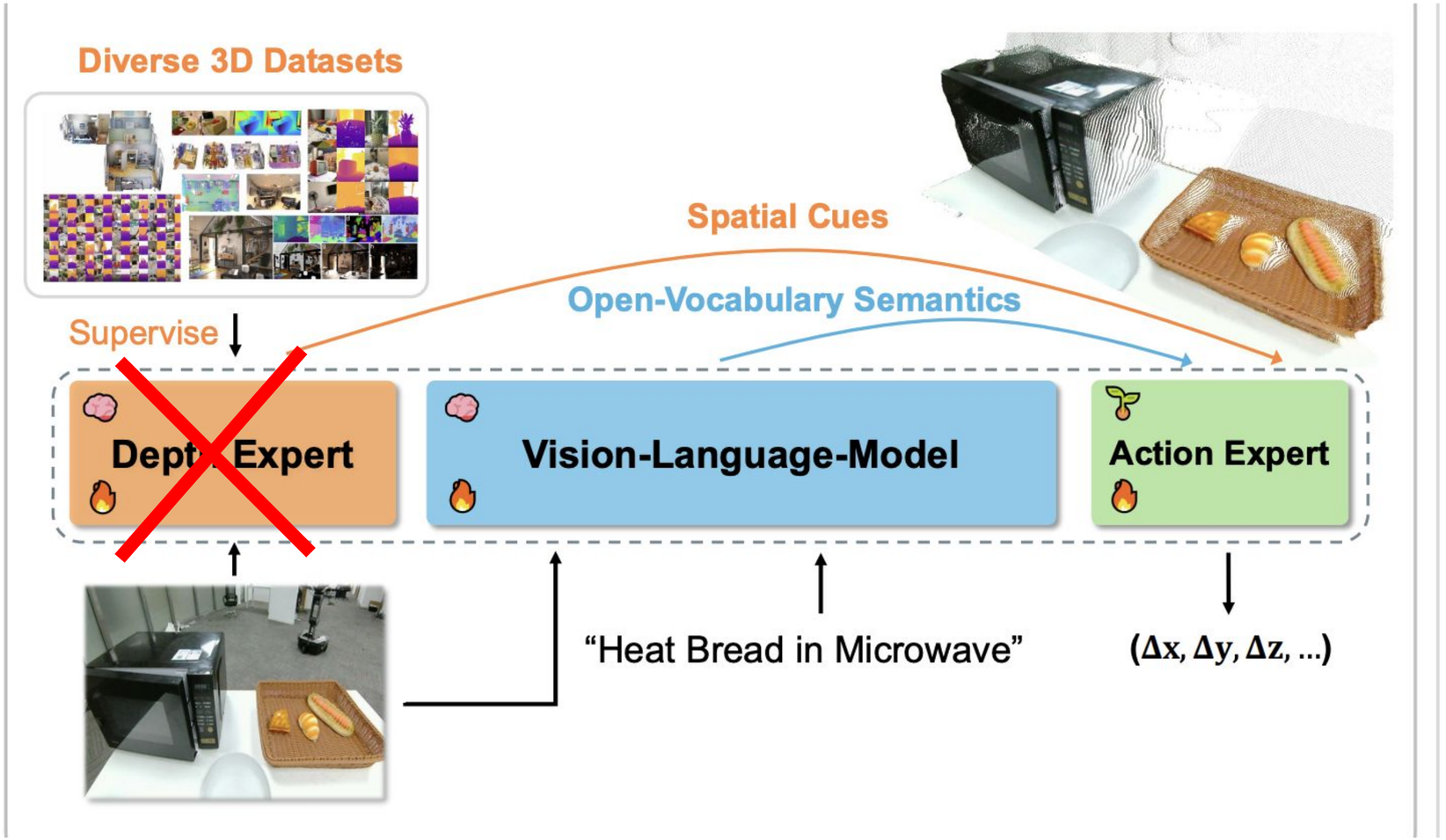
- **No architecture change**
- **Standard text supervision**

Why use VLMs? —> Simplicity



Luigi et al., 2025

Why use VLMs? —> Simplicity



Yuan et al. 2025

What Is Missing?

Can existing VLMs understand 3D?

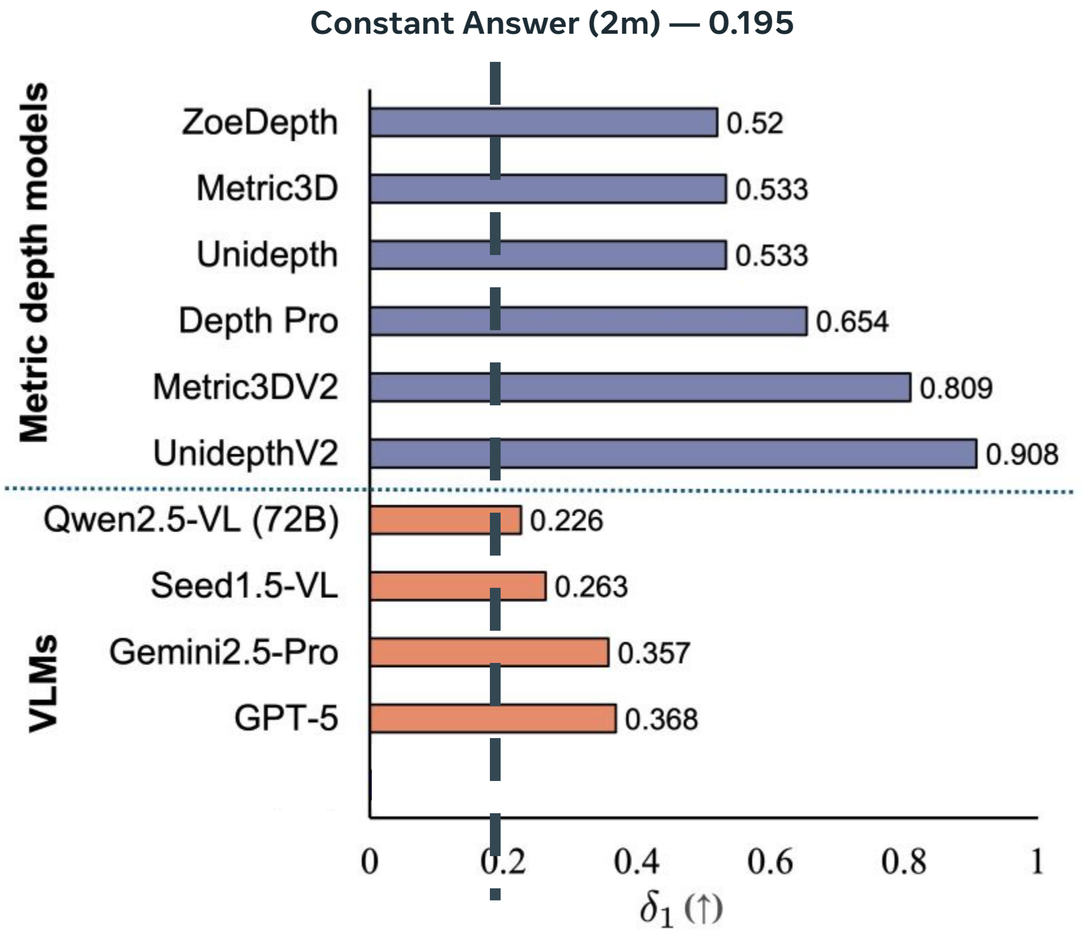
Question: How far is this point from the camera?

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

Can existing VLMs understand 3D?

Question: How far is this point from the camera?

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



DepthLM: standard VLM, expert model performance

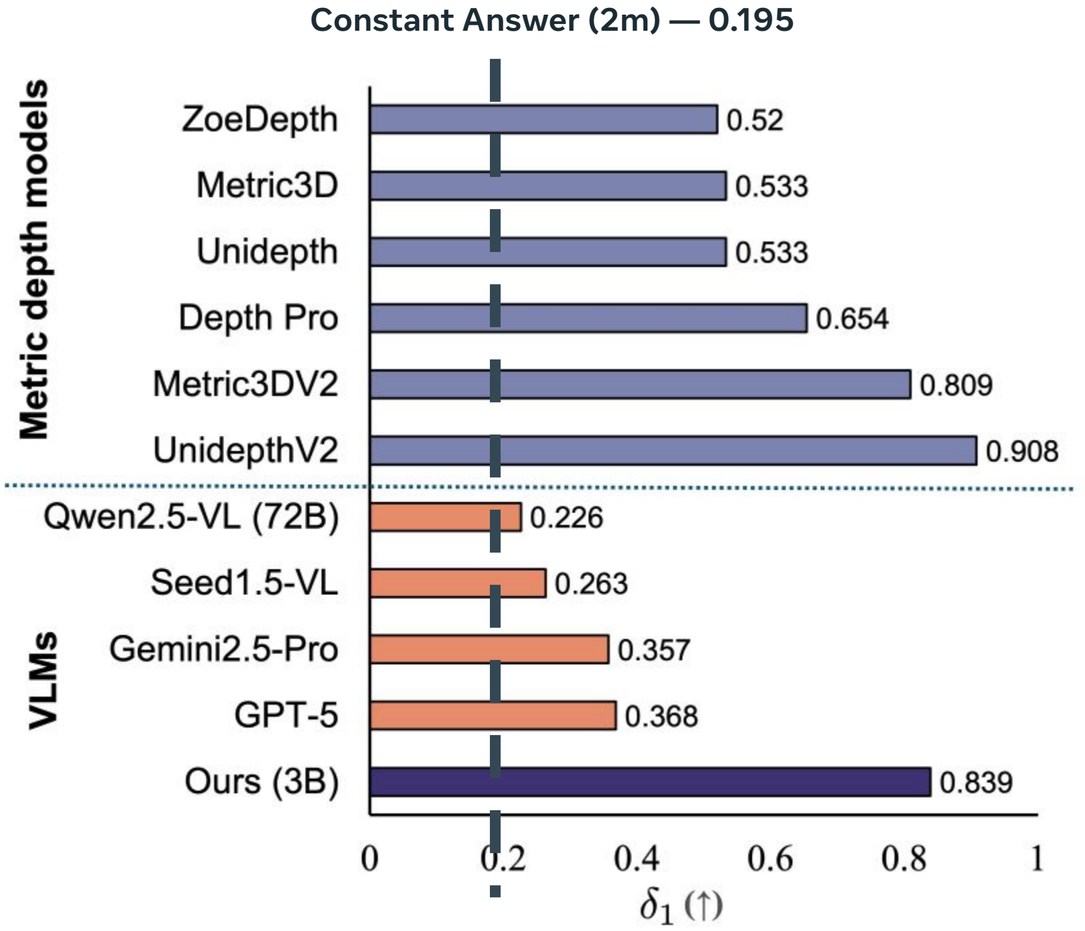
Question: How far is this point from the camera?

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

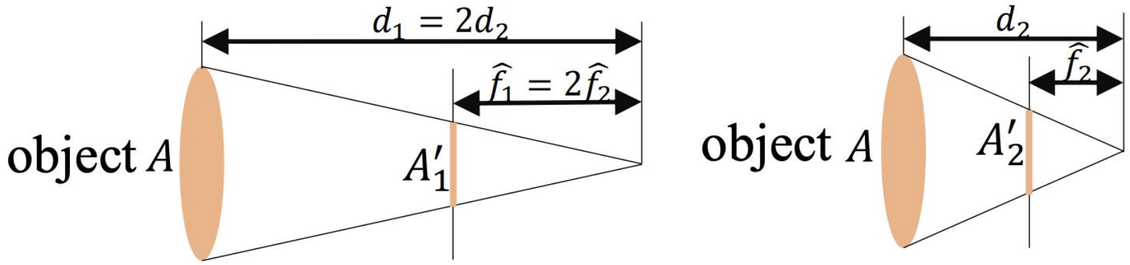
Input

Query each pixel independently

Ours



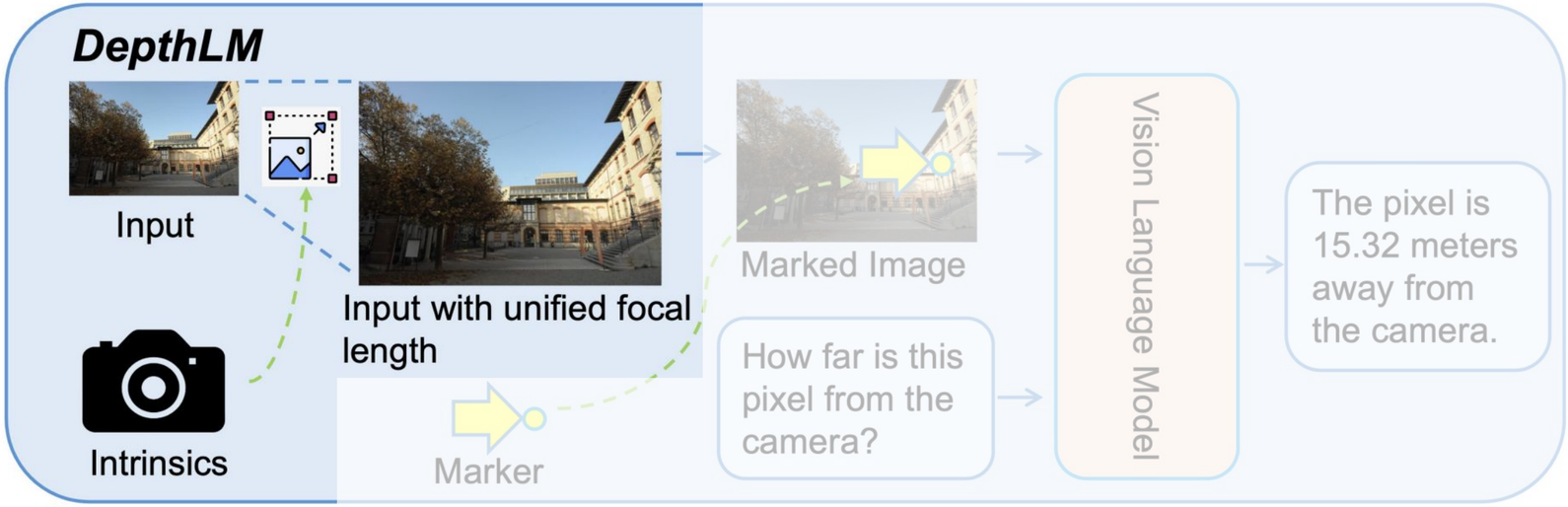
Bottleneck 1: camera ambiguity

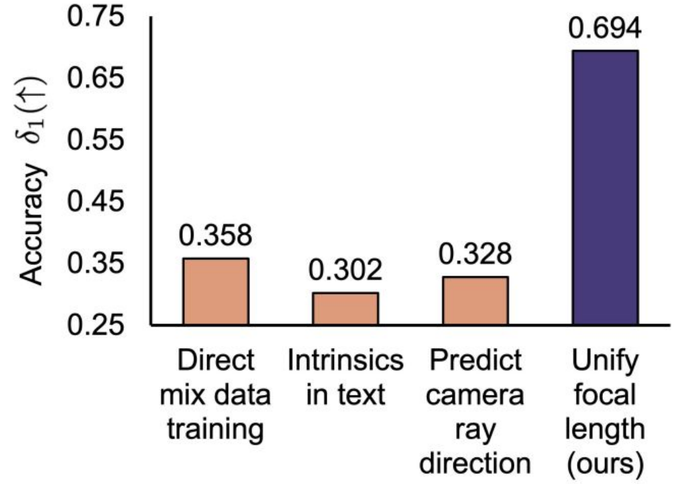


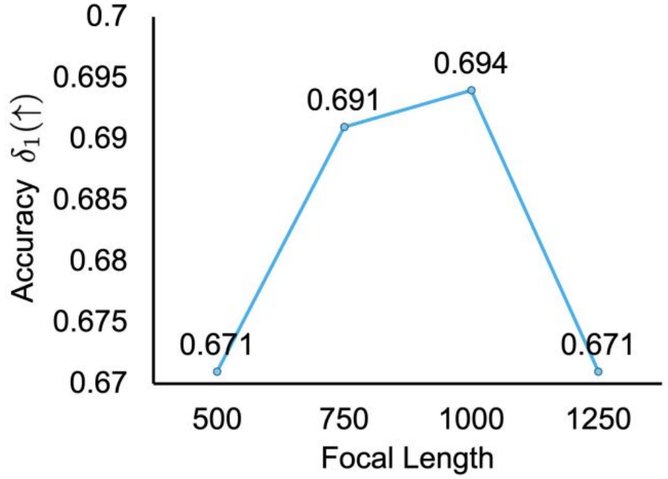
focal=26 mm, depth=2 m | focal=52 mm, depth=2 m | focal=26 mm, depth=1 m

Yin et al. 2023

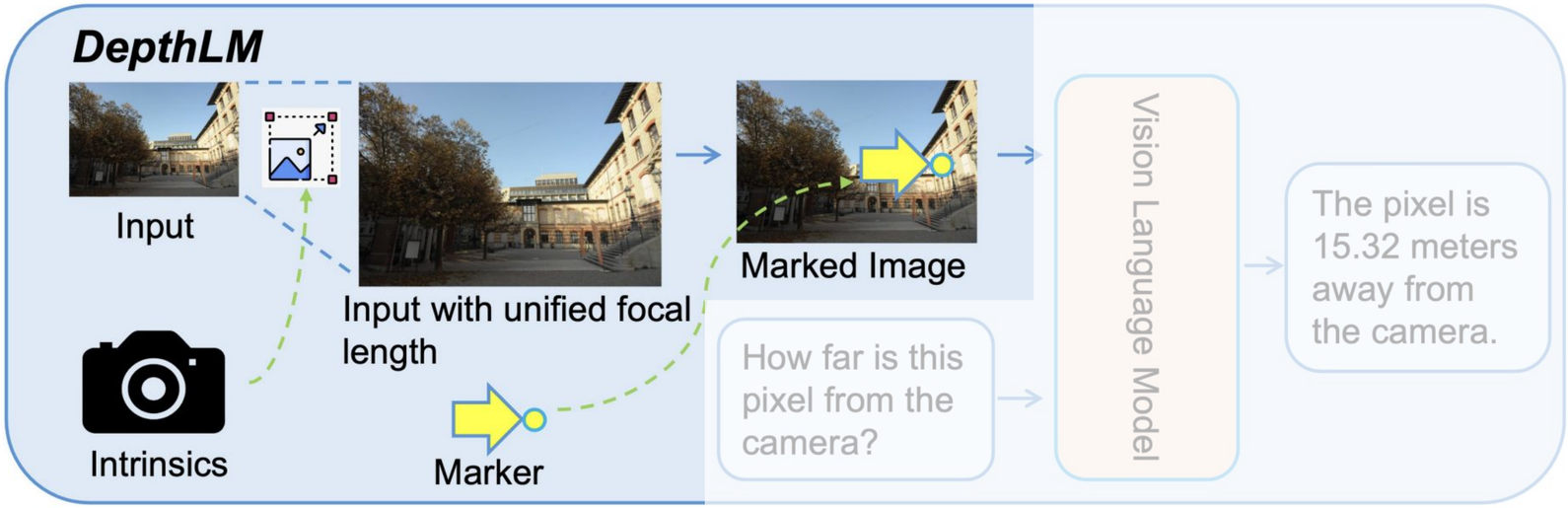
Bottleneck 1: camera ambiguity



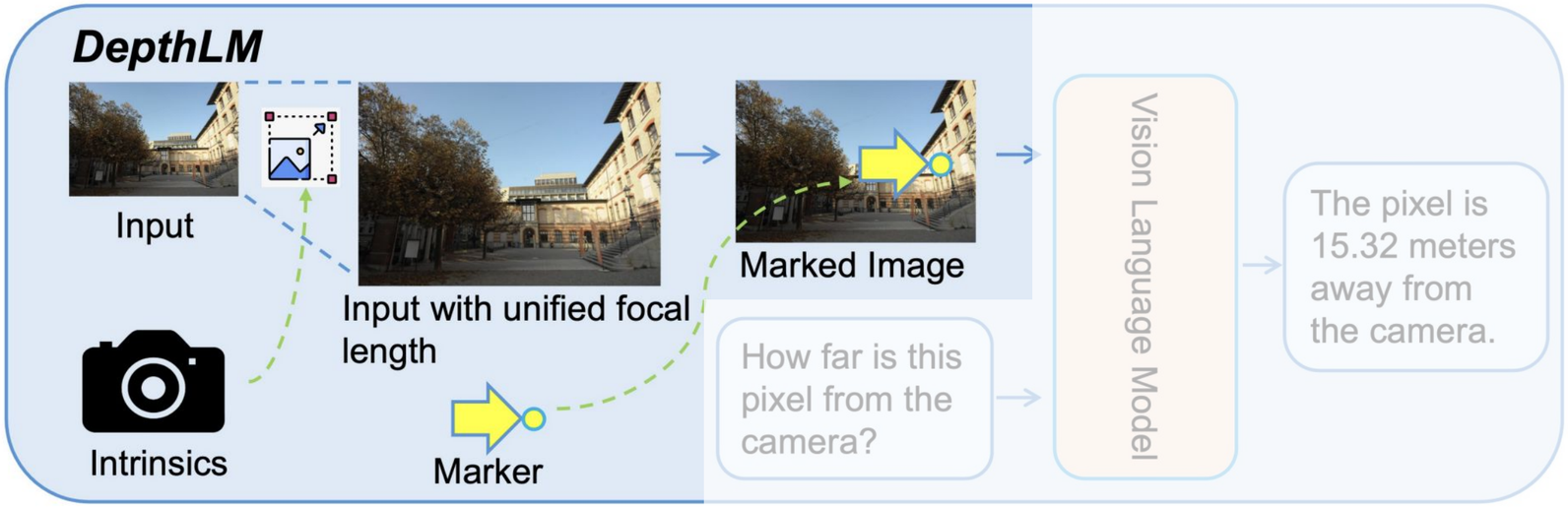


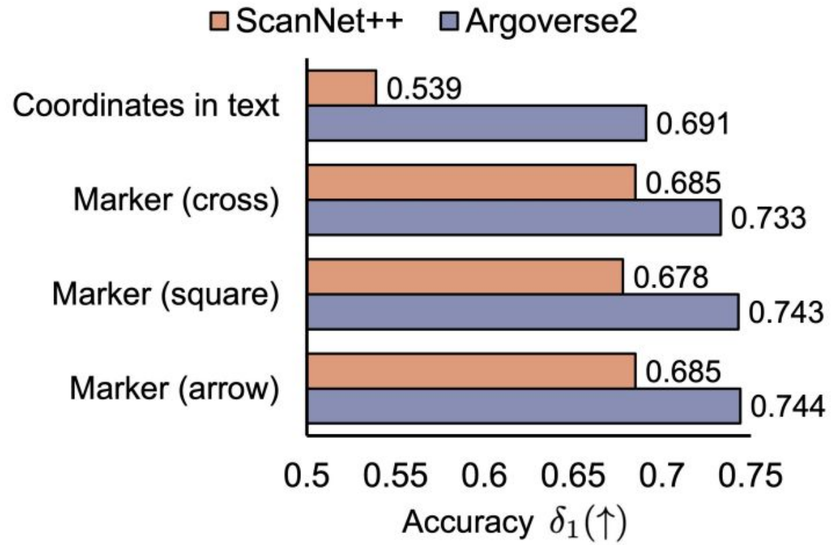


Bottleneck 2: pixel reference

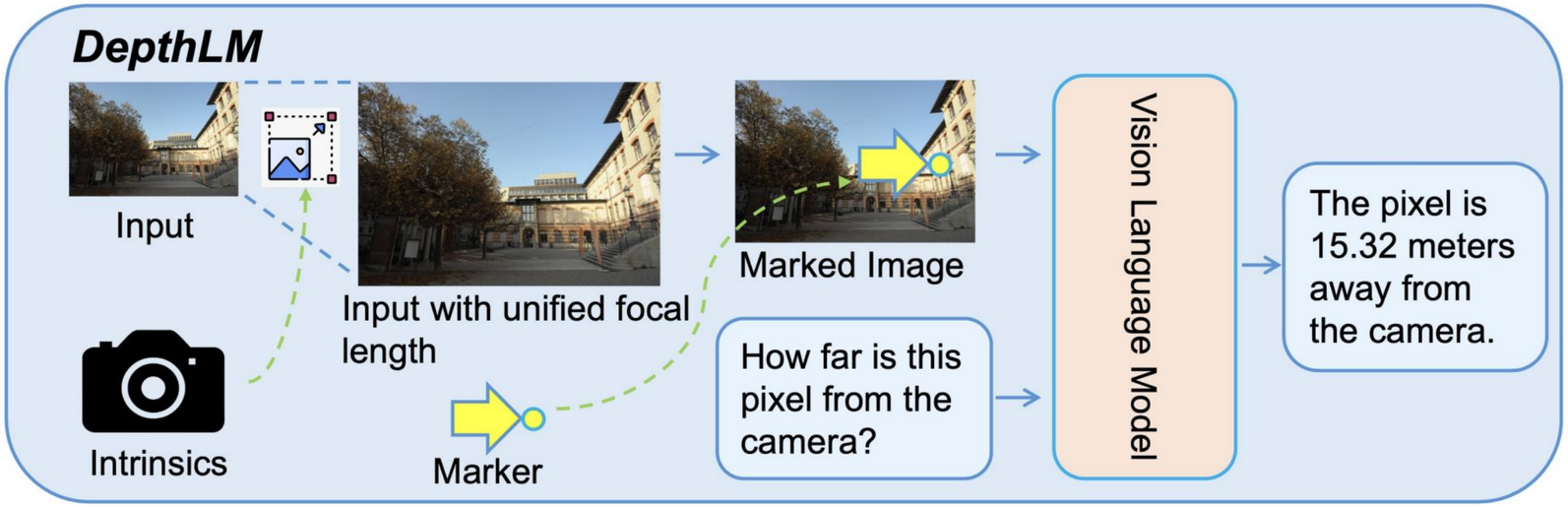


Render Markers

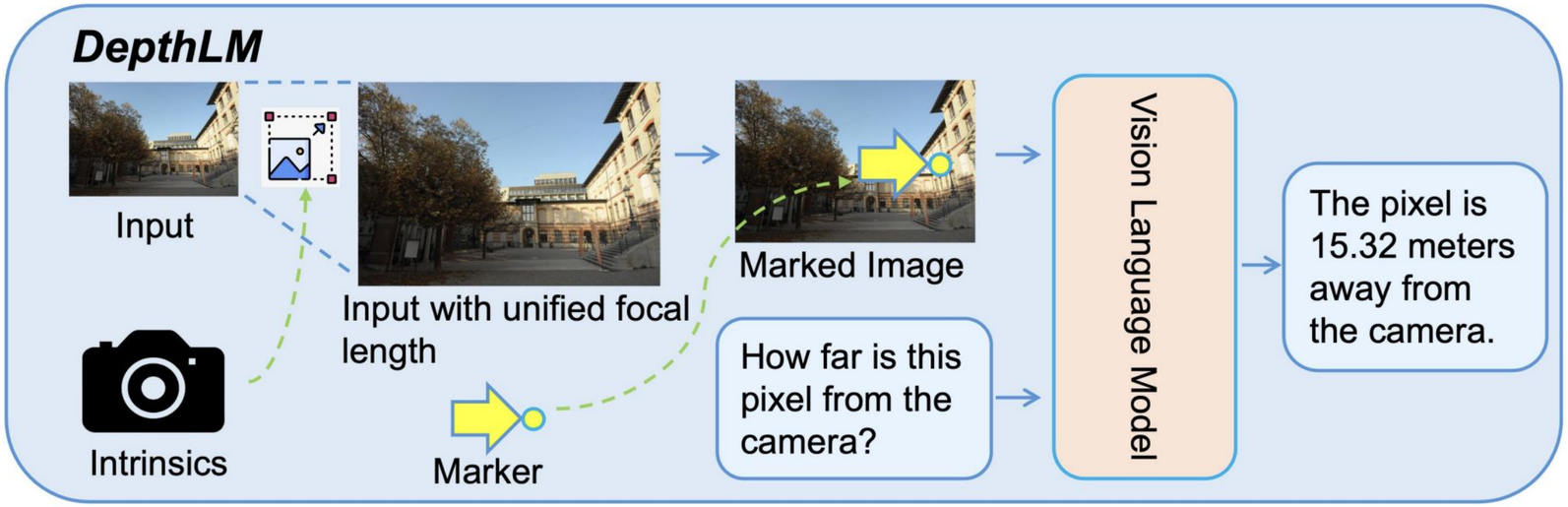


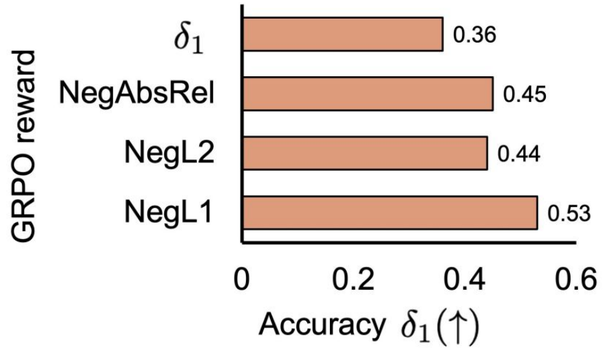


Text Supervision is All You Need!

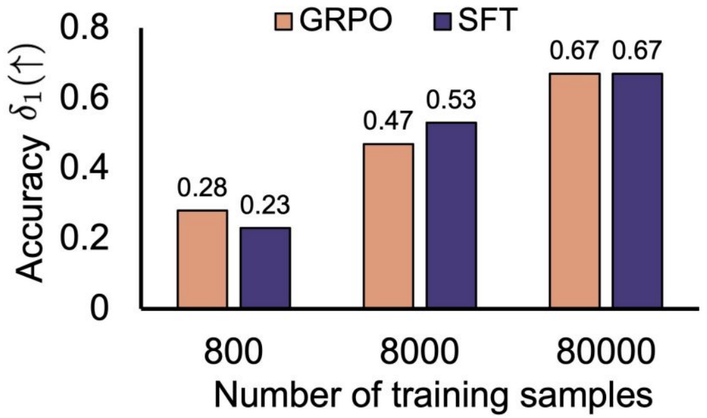


SFT is More Efficient Than GRPO



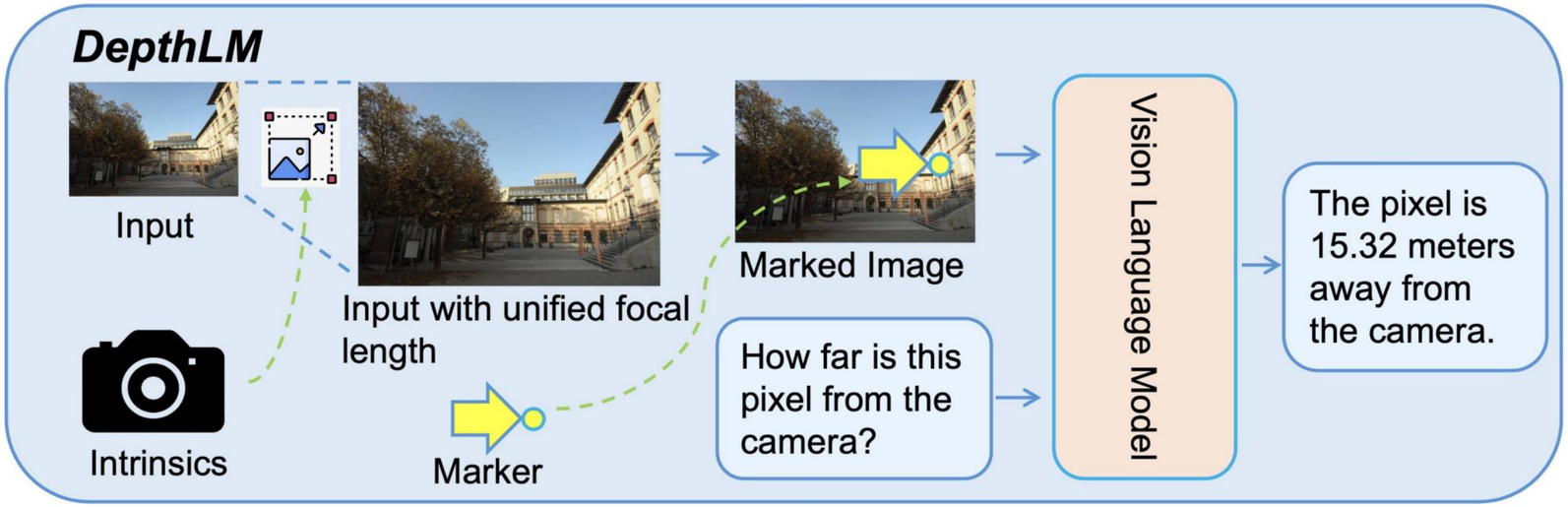


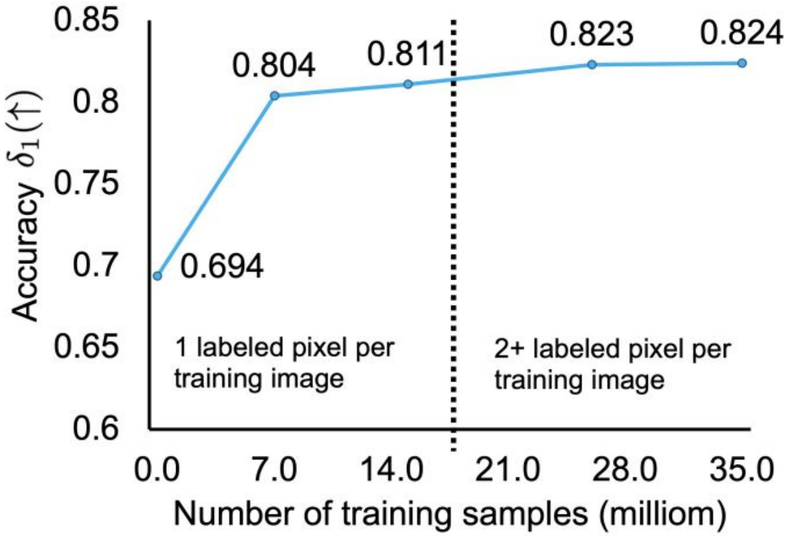
(a) Effect of GRPO rewards (8K training samples).



(b) Accuracy vs training data size.

1 Labeled Pixel Per Image Works!





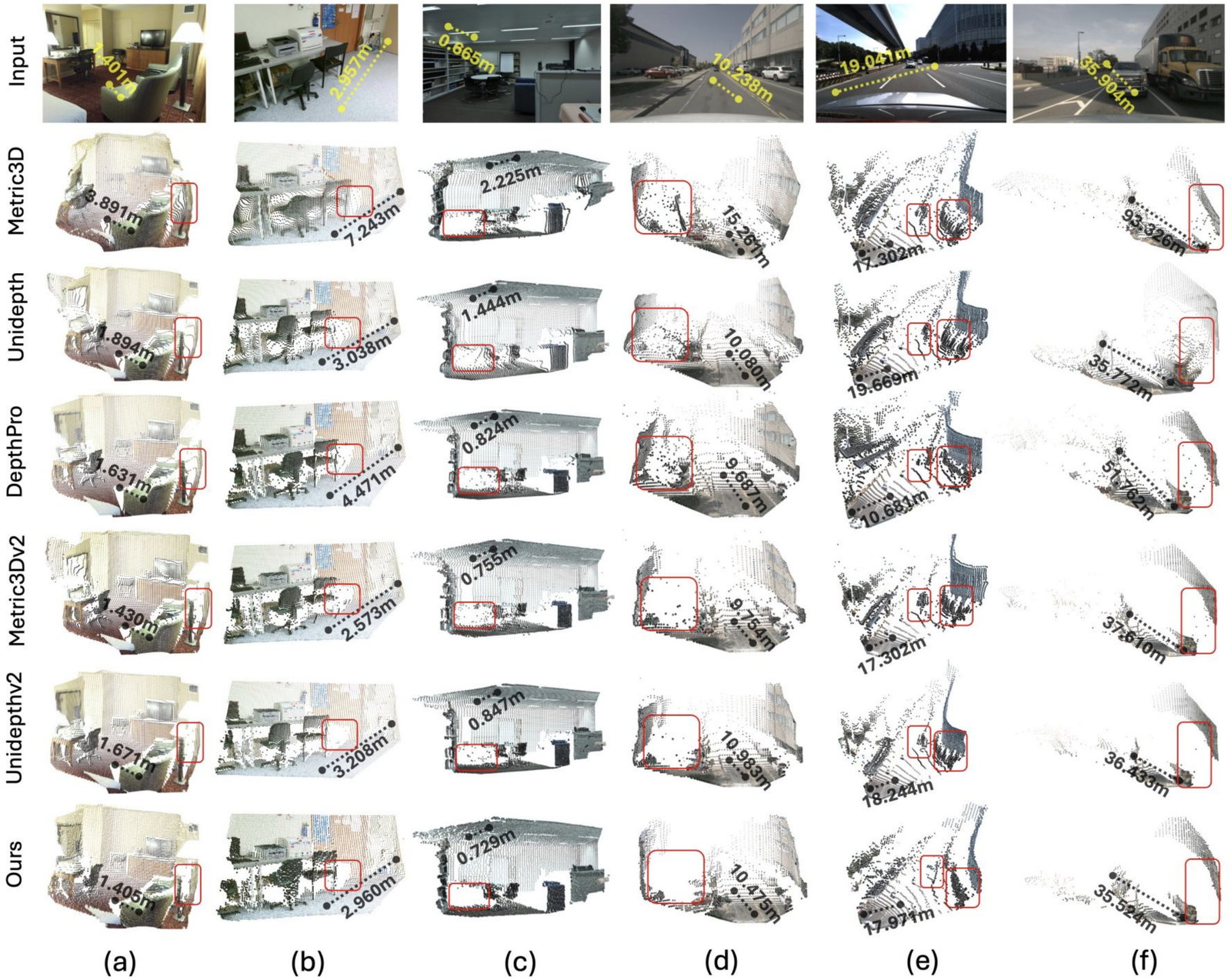
Result

$\delta_1(\uparrow)$ of different methods	*Out*			*Out+In*	*In*				Average
	Argoverse2	DDAD	NuScenes	ETH3D	ScanNet++	sunRGBD	iBims1	NYUv2	
Naive Prediction with Constant Answers									
Always Output 2.0m	0.006	0.010	0.010	0.106	0.305	0.384	0.280	0.383	0.186
VLMs									
Qwen2.5-VL (3b)	0.133	0.083	0.090	0.087	0.120	0.134	0.080	0.128	0.106
Qwen2.5-VL (7b)	0.077	0.120	0.070	0.126	0.135	0.089	0.160	0.168	0.118
Qwen2.5-VL (72b)	0.119	0.140	0.186	0.220	0.272	0.276	0.212	0.324	0.219
Molmo (7b-D)	0.200	0.132	0.200	0.126	0.244	0.299	0.200	0.225	0.203
Pixtral (12b)	0.157	0.132	0.118	0.141	0.318	0.308	0.270	0.145	0.199
Gemini-2.5-Pro	0.280	0.252	0.365	0.328	0.380	0.270	0.466	0.394	0.342
GPT-o3	0.208	0.283	0.309	0.305	0.375	0.426	0.375	0.470	0.344
GPT-5	0.218	0.302	0.382	0.313	0.428	0.471	0.307	0.540	0.370
Spatial VLMs									
SpaceLLaVA (13B)	0.100	0.067	0.083	0.090	0.269	0.233	0.208	0.178	0.154
SpatialRGPT (8B)	0.055	0.046	0.100	0.220	0.346	0.369	0.240	0.265	0.205
VLMs Trained on Metric Depth Estimation									
Seed1.5-VL (official setup)	0.009	0.012	0.013	0.219	0.495	0.321	0.459	0.412	0.243
Seed1.5-VL (our prompt)	0.040	0.074	0.028	0.309	0.593	0.689	0.627	0.841	0.400
Ours (3b)	0.808	0.724	**0.870**	**0.745**	0.838	0.850	0.890	0.868	0.824
Ours (7b)	**0.833**	**0.747**	0.865	0.718	**0.850**	**0.859**	**0.920**	**0.915**	**0.838**
Ours - Pixtral (12b)	0.734	0.670	0.819	0.653	0.834	0.786	0.870	0.799	0.771

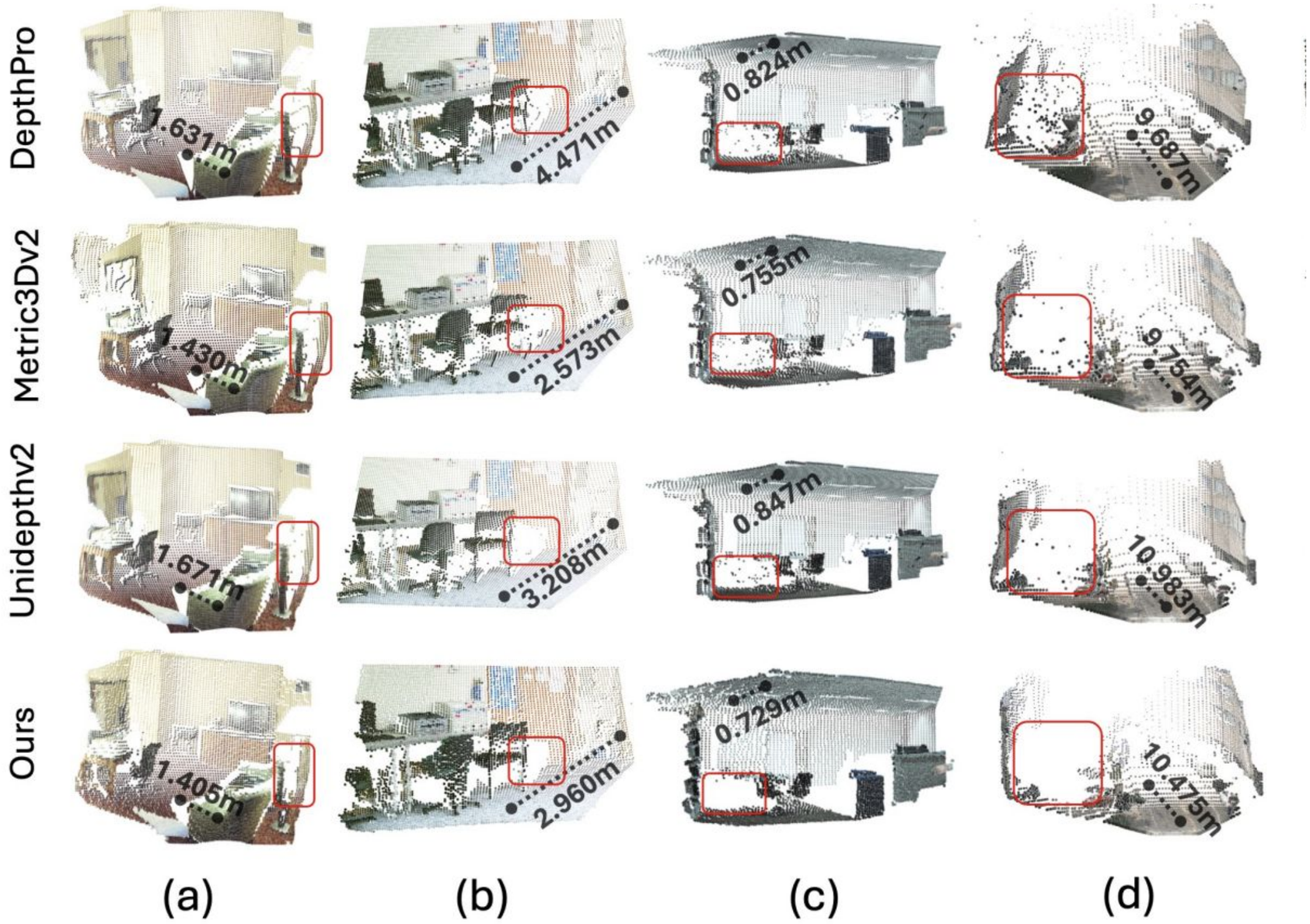
Result

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

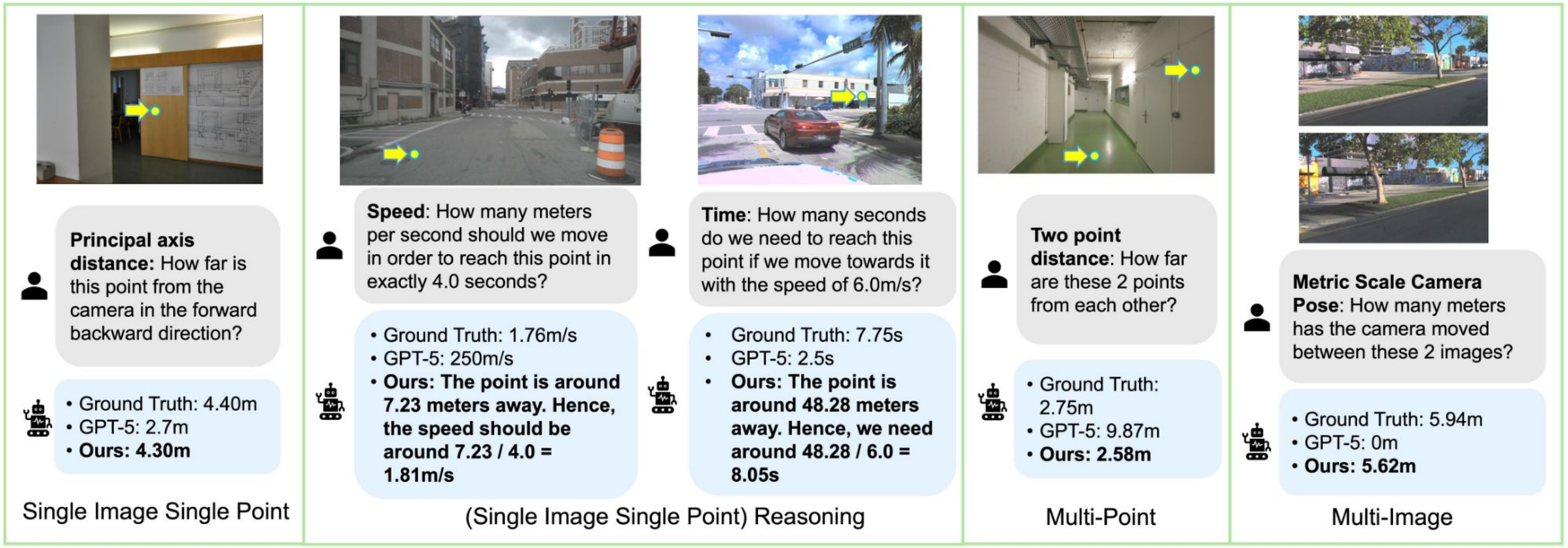
Visualizations



Visualizations



Scale Up to Diverse Tasks!

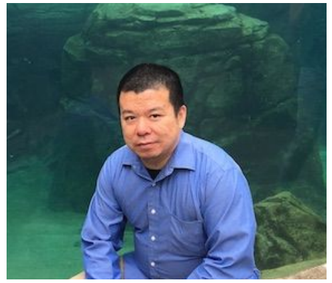


$\delta_1(\uparrow)$ on different tasks	*Single image single point*		*Reasoning*		*Multi-point*	*Multi-image*	
	Distance	Principal axis distance	Speed	Time	Two point distance	Pose	Average
ALWAYS OUTPUT 2.0	0.186	0.172	0.087	0.094	0.119	0.189	0.141
QWEN2.5-VL (7B)	0.118	0.085	0.136	0.087	0.066	0.048	0.09
SPACELLAVA (13B)	0.154	0.163	0.116	0.122	0.157	0.047	0.127
SPATIALRGPT (8B)	0.205	0.132	0.167	0.122	0.143	0.195	0.161
SEED1.5-VL (our prompt)	0.400	0.174	0.223	0.119	0.101	0.000	0.170
Gemini-2.5-Pro	0.342	0.209	0.213	0.209	0.140	0.025	0.189
GPT-5	0.370	0.241	0.199	0.181	0.150	0.120	0.210
OURS (7B)	**0.828**	**0.831**	**0.817**	**0.816**	**0.657**	**0.876**	**0.804**

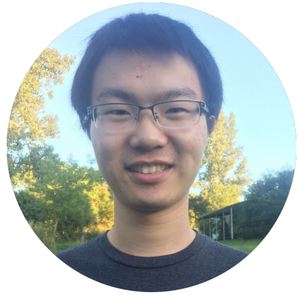
DepthLM Team

Zhipeng Cai

Ching-Feng Yeh



Hu Xu



Zhuang Liu

Gregory Meyer

Xinjie Lei

Changsheng Zhao

Shangwen Li

Vikas Chandra

Yangyang Shi

Code & Model

Meta

PRINCETON UNIVERSITY