AGO: Adaptive Grounding for Open World 3D Occupancy Prediction

Peizheng Li[1,2] Shuxiao Ding[1,4] You Zhou[1] Qingwen Zhang[5] Onat Inak[1,6] Larissa Triess[1], Niklas Hanselmann[1,2,3] Marius Cordts[1] Andreas Zell[2]
[1]Mercedes-Benz AG, Sindelfingen [2]University of Tübingen [3]Tübingen AI Center [4]University of Bonn [5]RPL, KTH Royal Institute of Technology [6]TU Berlin

ICCV HONOLULU HAWAII OCT 19-23, 2025

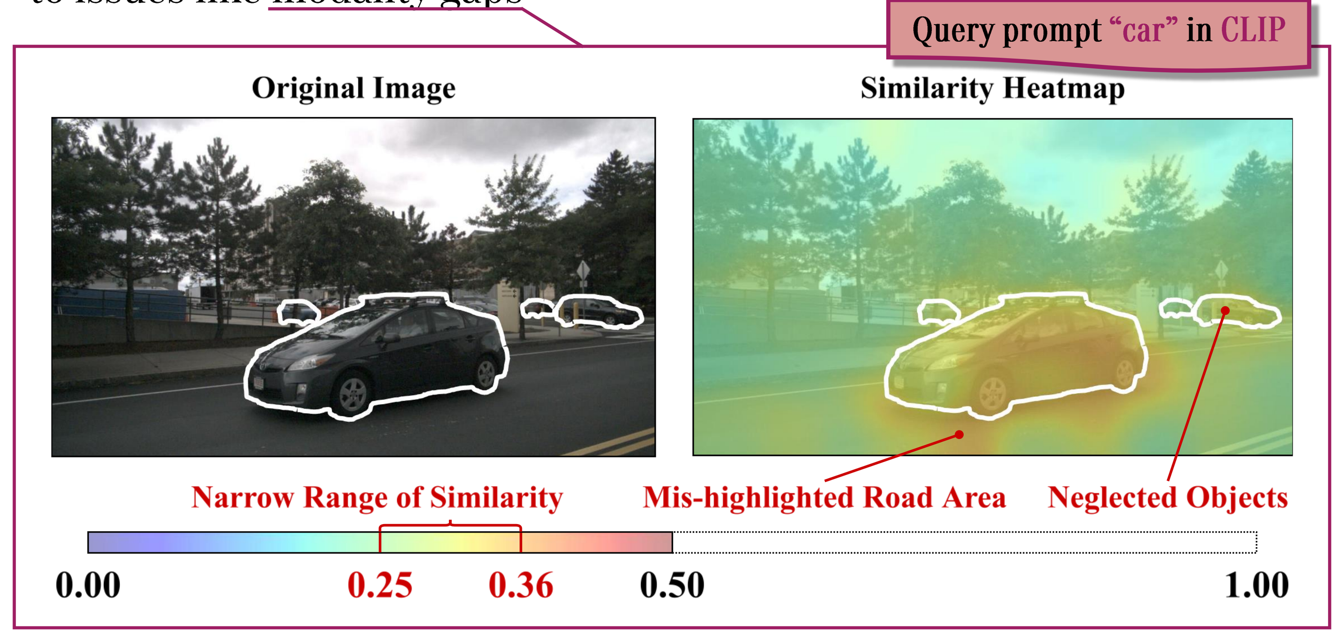
Motivation

3D semantic occupancy prediction is central to scene understanding for autonomous driving, yet it:

- heavily relies on extensive manual 3D annotations
- is constrained by predefined closed semantic spaces

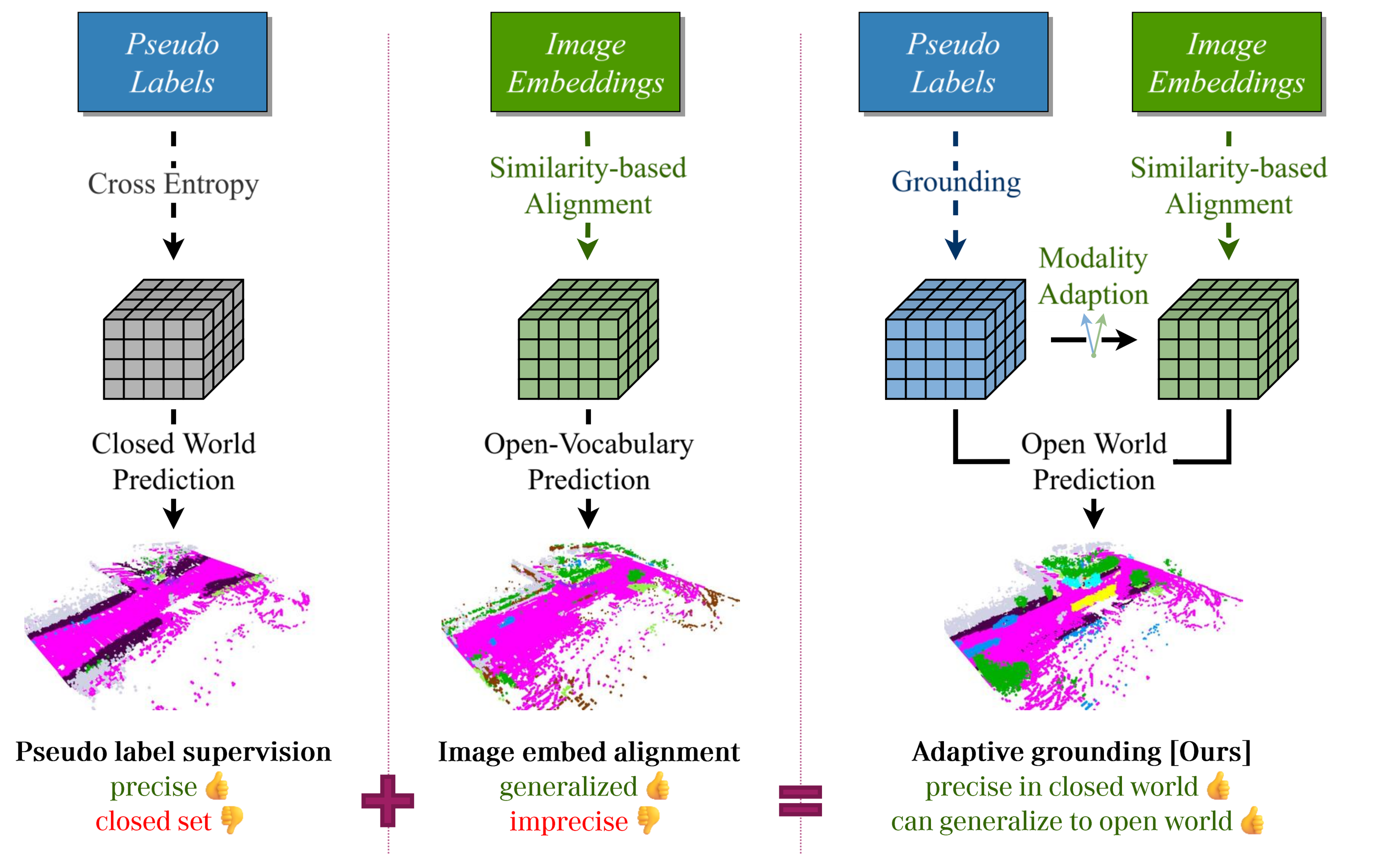
Existing VLM-based methods:

- rely on fixed-class pseudo-labels → struggles to predict novel classes
- base on image-text alignment → suffers from severe mismatches due to issues like modality gaps



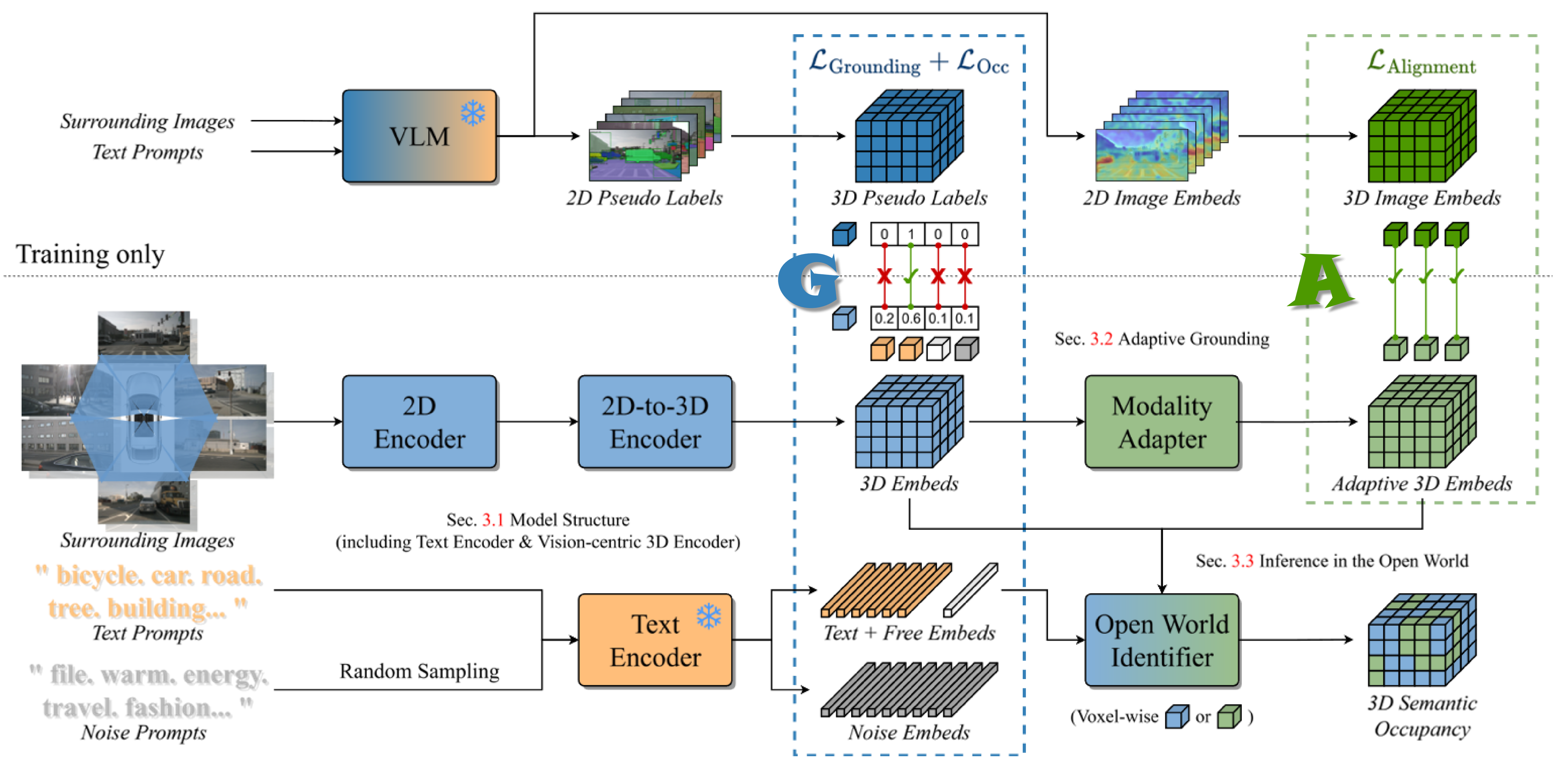
Goal: Enable open-world 3D semantic occupancy prediction with flexible adaptation to unknowns.

Insights



- AGO combines the advantages of existing methods based on pseudo-label supervision (Grounding instead of traditional CE to achieve open-vocabulary capability) or feature alignment.
- Modal adapters prevent feature space conflicts while promoting convergence.
- Entropy-based criteria enable adaptive selection of suitable features outputs.

Method



Benchmark Results

Closed World

Method	Image Backbone	oth.	bar.	bic.	bus	car	c. v.	mot.	ped.	t. c.	tra.	tru.	d. s.	o. f.	sid.	ter.	man.	veg.	mIoU*	mIoU
SimpleOcc [14]	ResNet-101	0.00	0.67	1.18	3.21	7.63	1.02	0.26	1.80	0.26	1.07	2.81	40.44	0.00	18.30	17.01	13.42	10.84	7.99	7.05
POP-3D† [45]	ResNet-101	0.06	0.02	0.46	1.83	4.87	0.00	0.00	1.29	0.00	0.65	2.62	**55.90**	1.60	9.99	25.17	15.75	21.11	9.42	8.31
SelfOcc [19]	ResNet-50	0.00	0.15	0.66	5.46	12.54	0.00	0.80	2.10	0.00	0.00	8.25	55.49	0.00	**26.30**	26.54	14.22	5.60	10.54	9.30
OccNeRF [51]	ResNet-101	0.00	0.83	0.82	5.13	12.49	3.50	0.23	3.10	1.84	0.52	3.90	52.62	0.00	20.81	24.75	18.45	13.19	10.81	9.53
GaussianOcc [15]	Swin	0.00	1.79	5.82	**14.58**	13.55	1.30	2.82	**7.95**	**9.76**	0.56	9.61	44.59	0.00	20.10	17.58	8.61	10.29	11.26	9.94
GaussTR [20]	VFMs	0.00	2.09	5.22	14.07	**20.34**	**5.70**	7.08	5.12	3.93	0.92	**13.36**	39.44	0.00	15.68	22.89	21.17	21.87	13.26	11.70
LangOcc [3]	ResNet-50	0.00	3.10	**9.00**	6.30	14.20	0.40	**10.80**	6.20	**9.00**	3.80	10.70	43.70	**2.23**	9.50	26.40	19.60	**26.40**	13.27	11.84
VEON [57]	ViT-L	**0.90**	**10.40**	6.20	**17.70**	12.70	**8.50**	7.60	6.50	5.50	**8.20**	11.80	54.50	0.40	25.50	**30.20**	**25.40**	25.40	**17.07**	**15.14**
AGO (ours)	ResNet-101	**1.53**	**6.75**	**6.43**	14.00	**22.82**	5.57	**16.66**	**13.20**	6.80	**10.53**	**15.89**	**71.48**	**4.48**	**34.48**	**41.37**	**29.33**	**25.66**	**21.39**	**19.23**

- In closed-world scenarios, AGO demonstrates substantial improvements across both static and dynamic categories.

Open World

Training Stages	Method	ped.	d. s.	sid.	veh.	cyc.	k. mIoU	car	bus	c. v.	tra.	tru.	bic.	mot.	bar.	t. c.	ter.	man.	veg.	u. mIoU	mIoU
Pretraining	POP-3D† [45]	0.00	58.77	13.80	8.27	1.10	16.39	-	-	-	-	-	-	-	0.00	0.00	3.95	**0.13**	0.60	0.94	8.66
	SelfOcc† [19]	0.98	60.29	14.68	7.11	0.00	16.61	-	-	-	-	-	-	-	0.00	0.00	0.00	0.00	0.00	0.00	8.31
	GaussTR† [20]	6.11	60.06	18.02	6.77	2.25	18.64	-	-	-	-	-	-	-	0.00	0.00	4.95	0.07	8.05	2.61	10.63
	AGO (ours)	**7.82**	**63.09**	**25.53**	**9.19**	**5.03**	**22.13**	-	-	-	-	-	-	-	0.00	0.00	**7.04**	0.03	**10.88**	**3.59**	**12.86**
Zero-shot Evaluation	POP-3D† [45]	0.00	58.77	13.80	-	-	24.19	6.72	0.00	0.00	0.59	4.34	1.17	**1.20**	0.00	0.00	3.95	0.13	0.60	1.56	6.08
	SelfOcc† [19]	0.98	60.29	14.68	-	-	25.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.06
	GaussTR† [20]	6.11	60.06	18.02	-	-	28.06	5.07	**1.65**	0.00	0.04	1.84	2.58	0.27	0.00	0.00	4.95	0.07	8.05	2.04	7.25
	AGO (ours)	**7.82**	**63.09**	**25.53**	-	-	**32.15**	**7.67**	0.00	0.00	**1.33**	**6.50**	**4.50**	0.00	0.00	0.00	**7.04**	0.03	**10.88**	**3.16**	**8.96**
Few-shot Finetuning	POP-3D† [45]	0.00	44.90	12.79	-	-	19.23	5.59	0.03	0.00	0.29	2.05	1.26	1.03	0.00	0.00	5.72	0.21	6.75	1.91	5.37
	SelfOcc† [19]	7.85	65.65	25.29	-	-	32.93	1.41	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	3.63	6.04	10.96	1.84	8.06
	GaussTR† [20]	7.84	66.36	25.55	-	-	33.25	10.85	1.58	0.00	0.00	1.32	1.42	0.00	0.00	0.00	12.74	9.12	8.16	3.77	9.66
	AGO (ours)	**13.00**	**71.54**	**29.91**	-	-	**38.15**	**18.73**	**5.49**	0.00	**0.41**	**2.16**	**3.72**	**2.22**	**0.43**	0.00	**29.63**	**21.43**	**17.73**	**8.50**	**14.43**

- In open-world scenes, AGO exhibits superior zero-shot performance while rapidly adapting to novel categories with only a few shots.

Experiments & Analysis

Training Paradigm

Training Paradigm	Self.	O.W. Pre.	O.W. Z.S.	O.W. F.S.
	mIoU			
Align	10.28	15.4 / 0.8 / 8.1	23.5 / 1.2 / 5.6	24.4 / 4.0 / 8.1
Gro.	19.08	20.6 / 0.0 / 10.3	**33.0** / 0.3 / 6.8	**38.4** / 3.0 / 10.1
Gro. + Align	18.89	18.3 / 2.2 / 10.2	29.3 / 1.4 / 7.0	37.3 / 5.7 / 12.0
AGO	**19.23**	**22.1 / 3.6 / 12.9**	32.2 / **3.2** / **9.0**	38.2 / **8.5** / **14.4**

- Grounding-based methods and alignment-based methods expertise in closed- and open-world scenarios, respectively.
- Adaptive feature space projection is superior to the simple superposition of grounding and alignment supervision within the same feature space.

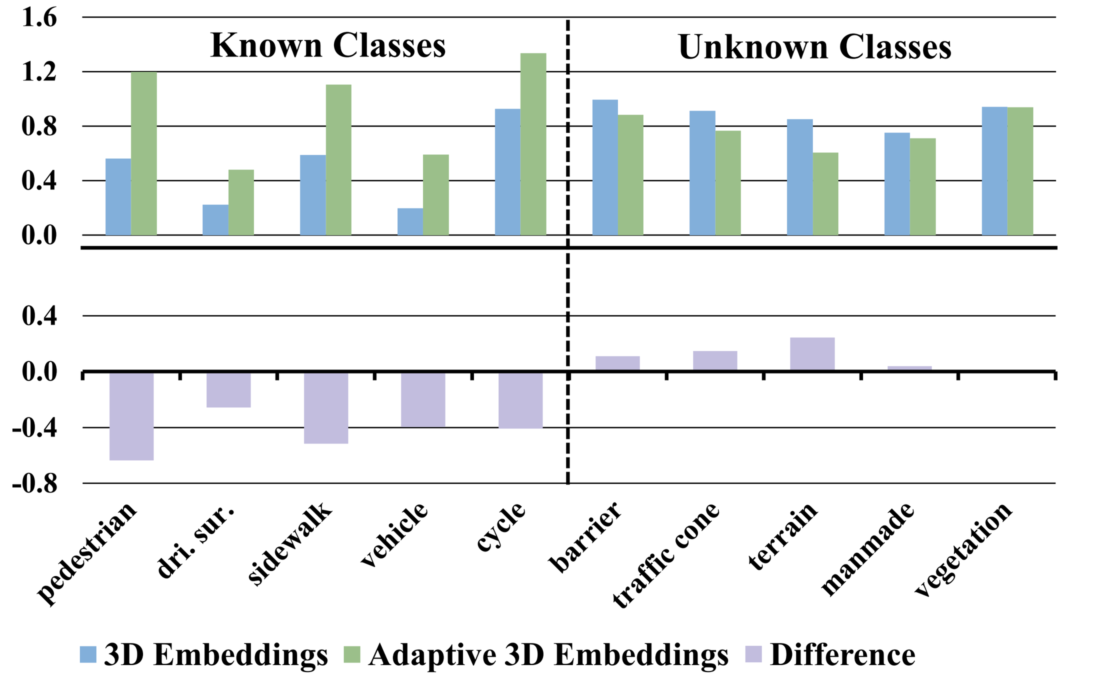
Architecture Components

Method	Noise Prompts	$\mathcal{L}_{\text{Occ}}$	Self-supervised IoU	Self-supervised mIoU
AGO			53.10	18.01
AGO		✓	55.18 (+2.08)	18.97 (+0.96)
AGO	✓	✓	55.45 (+0.27)	19.23 (+0.26)

- Noise prompts and separate $\mathcal{L}_{\text{Occ}}$ can systematically enhance the model's occupancy prediction capability.

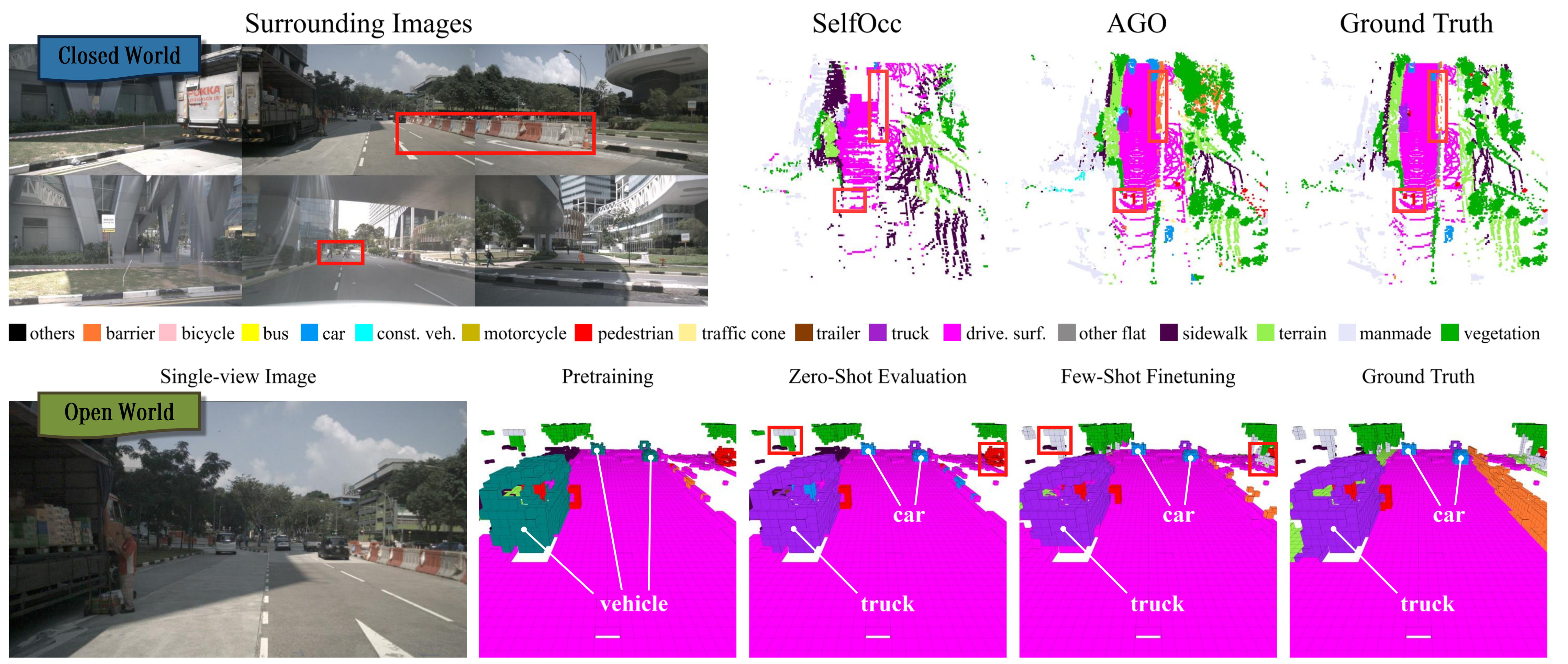
OW Inference Strategy

Open World Identifier	Discrimination Criteria	Open World Pretraining mIoU	Open World Zero-shot mIoU
✗	None	22.2 / 1.1 / 11.7	**32.8** / 1.6 / 7.8
✓	Max Confidence	**22.4** / 3.1 / 12.8	32.6 / 2.8 / 8.7
✓	Min Entropy	22.1 / **3.6** / **12.9**	32.2 / **3.2** / **9.0**



- Entropy-based criteria enable adaptive selection of more certain predictions for objects in the open world.

Visualization



Conclusion

- We introduce AGO, a novel open-world occupancy prediction framework that adaptively distills knowledge from pretrained VLMs into 3D perception for autonomous driving.
- Our method achieves state-of-the-art performance on the closed-world self-supervised Occ3D-nuScenes benchmark and significantly outperforms existing methods in open-world occupancy prediction.

arXiv
GitHub