Language Driven Occupancy Prediction

Supplementary Material

Table 1. The vocabulary list used in LOcc. The predefined classes of nuScenes [2, 8] are employed as the super classes, while the subclasses are summarized from the nuScenes LiDAR segmentation [4] benchmark. During evaluation, we compute cosine similarity between the estimated 3D language volume features and the text embeddings of these subclasses. Each voxel is assigned the vocabulary with the highest similarity score, then categrized into the corresponding superclass for computing quantitative metrics.

Superclass	List of subclasses
barrier	barrier, traffic barrier
bicycle	bicycle
bus	bus
car	car, vehicle, sedan, SUV
construction vehicle	construction vehicle, crane
motorcycle	motorcycle
pedestrian	pedestrian, person
traffic cone	traffic cone
trailer	trailer
truck	truck
driveable surface	dariveable surface, road
other flat	water, river, lake
sidewalk	sidewalk
terrain	terrain, grass
manmade	building, wall, traffic light, sign, parking meter, hydrant, fence
vegetation	vegetation, tree

1. Subclass Definition

During inference, we define a set of text labels to facilitate semantic assignment. In Occ3D-nuScenes [8], voxels are naively classified into 17 non-free classes. However, these predefined classes are unsuitable for open-vocabulary tasks. For instance, "manmade" and "others" encompass a wide variety of subcategories, making them overly broad and semantically vague. To address this limitation, we subdivide the original superclasses into more specific subclasses, as detailed in Table 1. Each subclass is associated with a language template, "a photo of a {}," where {} is replaced by the subclass name.

This subclassification enables fine-grained open-vocabulary occupancy prediction. For example, voxels identified as building, wall, traffic light, or fence represent distinct semantic categories, even though they all fall under the same superclass "manmade". For quantitative evaluation on the Occ3D-nuScenes [8] benchmark, we map the subclasses back to their corresponding superclasses as outlined in Table 1 and compute the overall mIoU metric.

2. More Analyses for Vocabularies Integrating

During our investigation, we noticed that the LVLM occasionally encounters failures, potentially overlooking significant classes and thus yielding suboptimal segmentation results. To address this issue, we consider multiple frames as a consolidated sequence, merging the vocabularies from each frame into a unified set. This approach allows us to augment frames with incomplete text classes by utilizing information from adjacent frames within the same sequence. Detailed metrics of the pseudo-labeled ground truth generated from both single-frame and unified vocabularies for a particular scene are summarized in Table 2. A comparison between these different sets of vocabularies is provided in Table 3. Additionally, Fig. 1 and Fig. 2 illustrate an example of the segmentation results and their corresponding pseudo-labeled ground truth, respectively. As indicated in Table 2, the absence of the *truck* class in frame *40a982ccf9564c6ea574f1d75f8a7dc0* leads to the segmentation model misclassifying *truck* as *car*, resulting in diminished performance metrics for the *truck* class. In contrast, through leveraging temporal frames, the unified vocabulary approach helps correctly identify and segment the *truck*, thereby overcoming such limitations.

3. Comparison of Predefined Subclasses and Vocabularies Derived from LVLM

We organize the predefined subclasses and vocabularies derived from LVLM for a particular scene, and list them in Table. 4, facilitating a clearer comparison. It is observed that most classes share similar meaning with the predefined ones, while there exists only several additional classes which are challenging to categorize.

4. Ablation Study on the Resolution of the Segmentation Maps

The image feature resolution of SAN [11] is one-eighth that of the input image resolution. To investigate the effect of segmentation resolution, we resize the segmentation maps to 0.75, 0.5, 0.25, and 0.125 times their original resolutions, respectively. These resized maps are then used to generate pseudo-labeled ground truth. The mIoU for each setting, along with the setting of using image features as intermediates, is summarized in Table 5. As shown in the table, a decrease in resolution generally leads to a decrease in mIoU, with performance converging when the scale is one-fourth of the original resolution. These results prove the strengths of our proposed pipeline, compared to using image features

Table 2. Detailed metrics of the pseudo-labeled ground truth with and without vocabulary integration.

Method	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
w/o vocabulary integration	15.96	0.00	6.81	1.35	0.00	19.64	3.51	18.84	10.23	5.67	0.00	10.65	51.70	0.00	15.30	12.10	62.28	53.38
w vocabulary integration	23.58	0.00	10.39	30.79	0.00	32.65	10.59	24.07	12.32	5.76	0.00	42.38	52.50	0.00	21.97	26.10	66.16	65.10

Table 3. Comparison between the single-frame and unified vocabularies.

Frame	Vocabularies
0bcb08a96d264c8ca9f2119c6b0dbeb2	barrier, building, construction vehicle, crane, fence, road, sidewalk, street, terrain, traffic barrier, traffic cone, traffic light, traffic sign, tree, vegetation
40a982ccf9564c6ea574f1d75f8a7dc0	**bike**, building, car, fence, overcast, person, road, sidewalk, sign, street, traffic light, tree
1f1a315571894e99a53928e74e5efcad	barrier, building, car, construction vehicle, crane, fence, pedestrian, person, road, sidewalk terrain, traffic barrier, traffic cone, traffic light, traffic sign, tree, **truck**, vegetation
unified	arrow, barrier, **bike**, building, car, closed, construction vehicle, crane, fence, fire hydrant, gate grass, helmet, man, merge, motorcycle, overcast, palm tree, parking garage, pedestrian, person road, sidewalk, sign, street, street light, street sign, taxi, terrain, traffic barrier, traffic cone traffic light, traffic sign, tree, **truck**, van, vegetation, wall, woman

as intermediates.

5. Ablation Study for Vocabularies Integrating

In this manuscript, we treat the frames within the same scene as a unified sequence and merge their vocabularies into a unified set. To examine the effect of the number of frames on vocabulary unification, we vary the number of frames per sequence and analyze its impact on overall performance. The mIoU metrics for different configurations are presented in Table 6. As the number of frames increases, the mIoU of the pseudo-labeled ground truth also improves, indicating enhanced accuracy and consistency in the merged vocabularies. However, this improvement tends to converge once the number of frames exceed a certain threshold. While incorporating more frames introduces additional vocabulary, which helps recover missing terms from individual frames, it also brings in more irrelevant text. Nevertheless, the results remain stable beyond a certain point. We infer that OV-Seg models, typically trained on high-quality images and regular classes, tend to overlook uncommon classes during inference, thereby mitigating the influence of irrelevant vocabulary.

6. Ablation Study for Autoencoder

We conduct experiments to verify the effectiveness of the autoencoder. Without it, the model is required to generate 3D features with a dimension of 512. In this case, we set the voxel feature dimension after the 2D-to-3D transformation for BEVDet [6] and BEVDet4D [5] to 128. For BEVFormer [7], the BEV queries are kept at a dimension of 256, as using a larger dimension exceeds the GPU memory. As shown in Table 7, introducing autoencoder can effectively reduce the memory requirements with better performance. Here, we hypothesize that this is because CLIP model is trained using 400 million (image, text) pairs, thus its high-dimensional space could be highly compact. Instead, the number of extracted texts is around one thousand in this work, which is significantly smaller than the number of texts used in CLIP training, resulting a sparse space and allowing us to further compress it.

7. Ablation Study for the Conversation Prompt

We conduct ablation studies to evaluate the effectiveness of our proposed chain-of-thought [9] conversation process. In Fig. 3, although the single-step prompt can extract valid classes, the results of our proposed are more complete.

(a) RGB (b) Single-frame vocabularies (c) Unified vocabularies

Figure 1. Comparison of the segmentation maps using different sets of vocabularies. Scene token: 4dd38dcd4e8549d6a37938285f886117.

(a) Single-frame vocabularies (b) Unified vocabularies

Figure 2. Comparison of the pseudo-labeled ground truth using different sets of vocabularies.

Fig. 4 further provides a failure example, where the single-step prompt cannot extract valid texts, while ours can work normally. These results prove the effectiveness of our proposed chain-of-thought [9] prompt.

8. More Ablation Results for the Open-Vocabulary Segmentation Models

In Table 8, we present detailed metrics for the pseudo-labeled ground truth generated from different OV-Seg models.

9. More Qualitative Results

In this supplementary material, we present additional visualization results of the pseudo-labeled 3D language occupancy ground truth generated using various pipelines in Fig. 7 and Fig. 8, including voxel-based model-view projection, image features as intermediates, and our proposed semantic transitive labeling. Furthermore, we provide the results of the OVO models: LOcc-BEVFormer, LOcc-BEVDet, and LOcc-BEVDet4D in Fig. 9 and Fig. 10.

10. Limitations

While LOcc demonstrates strong performance on benchmarks, there remains a gap compared to supervised methods. The overall pipeline's performance depends on the LVLM and OV-Seg models. Due to computational resource constraints, we utilize Qwen-VL-Chat-Int4 [1] to extract texts from images. Replacing it with a stronger model, such as ChatGPT, may help reduce the impact of noisy text extractions. Regarding the segmentation results, we present two segmentation examples in Fig. 5 and Fig. 6. As shown in these images, erroneous segments may lead to incorrect text label assignments for the LiDAR point clouds, which in turn result in incorrect voxel labels. The performance of OV-Seg remains unsatisfactory, especially for images with various corruptions, highlighting the need for further improvements in segmentation accuracy. Despite these drawbacks, LOcc presents a preliminary exploration of occupancy label generation, diminishing the requirements of laborious human annotations and showing better performance than existing methods. We believe LOcc will contribute to advancing the community, and these limitations can be addressed with the advancements in foundational models.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 3

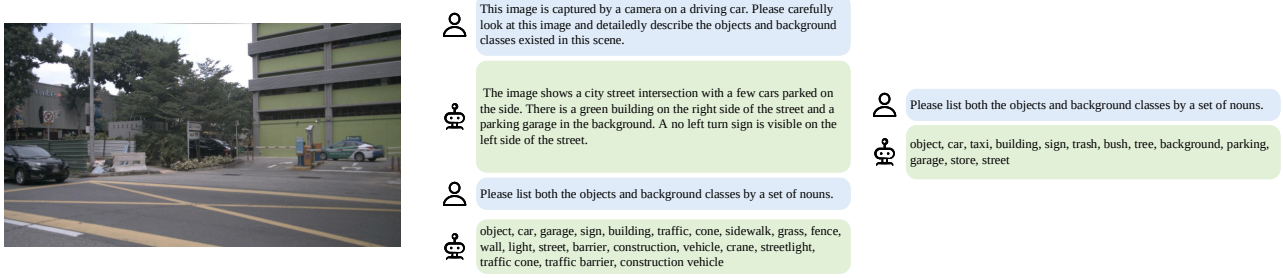
[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 1

[3] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 5

[4] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lub-

Table 4. Comparison between the predefined subclasses and vocabularies derived from LVLM.

Scene	Vocabularies
Predefined classes	barrier, traffic barrier, bicycle, bus, car, vehicle, sedan, SUV, motorcycle, pedestrian, person, traffic cone trailer, truck, dariveable surface, road, water, river, lake, sidewalk, terrain, grass, building, wall traffic light, sign, parking meter, hydrant, fence, vegetation, tree
Scene-0002	arrow,barrier,**bike**,building,car,closed,construction vehicle,crane,fence,fire hydrant,gate grass,helmet,man,merge,motorcycle,overcast,palm tree,parking garage,pedestrian,person road,sidewalk,sign,street,street light,street sign,taxi,terrain,traffic barrier,traffic cone traffic light,traffic sign,tree,**truck**,van,vegetation,wall,woman



(a) RGB

(b) left: chain-of-thought (ours) right: single-step

Figure 3. Ablation on different conversation prompts.

Table 5. Ablation on the resolution of segmentation maps.

	1	0.75	0.5	0.25	0.15	features as intermediates
mIoU	25.53	25.15	25.12	25.11	25.11	22.12

Table 6. Ablation on the number of frames used for vocabulary integrating.

	1	5	10	20	frames within a scene
mIoU	22.66	23.57	24.89	25.57	25.53

Table 7. Ablation study on the language autoencoder.

Method	w/o autoencoder			w/ autoencoder		
	mIoU↑	Memory (G) ↓	Params (M)↓	mIoU ↑	Memory (G) ↓	Params (M)↓
LOcc-BEVFormer	17.36	1.41	61.80	**18.62**	**1.18**	**61.70**
LOcc-BEVDet	20.25	2.22	92.40	**20.29**	**1.11**	**41.60**
LOcc-BEVDet4D	22.62	2.33	100.00	**22.68**	**1.11**	**48.00**

ing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7(2):3795–3802, 2022. 1

[5] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 2

[6] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2

[7] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proceedings of the European Conference on Computer Vision*, pages 1–18, 2022. 2

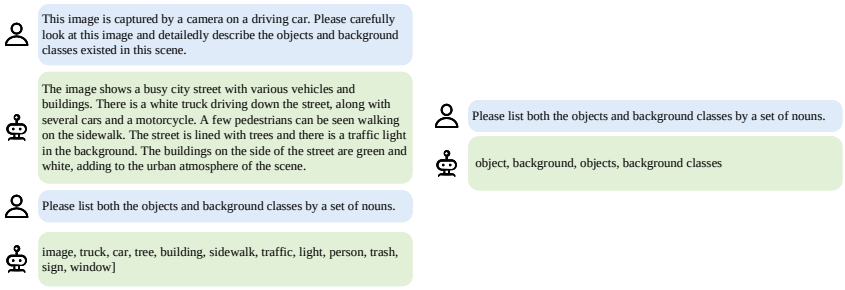
[8] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. In *Advances in Neural Information Processing Systems*, pages 64318–64330, 2023. 1, 9, 10

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 2, 3

[10] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 5

[11] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic

(a) RGB



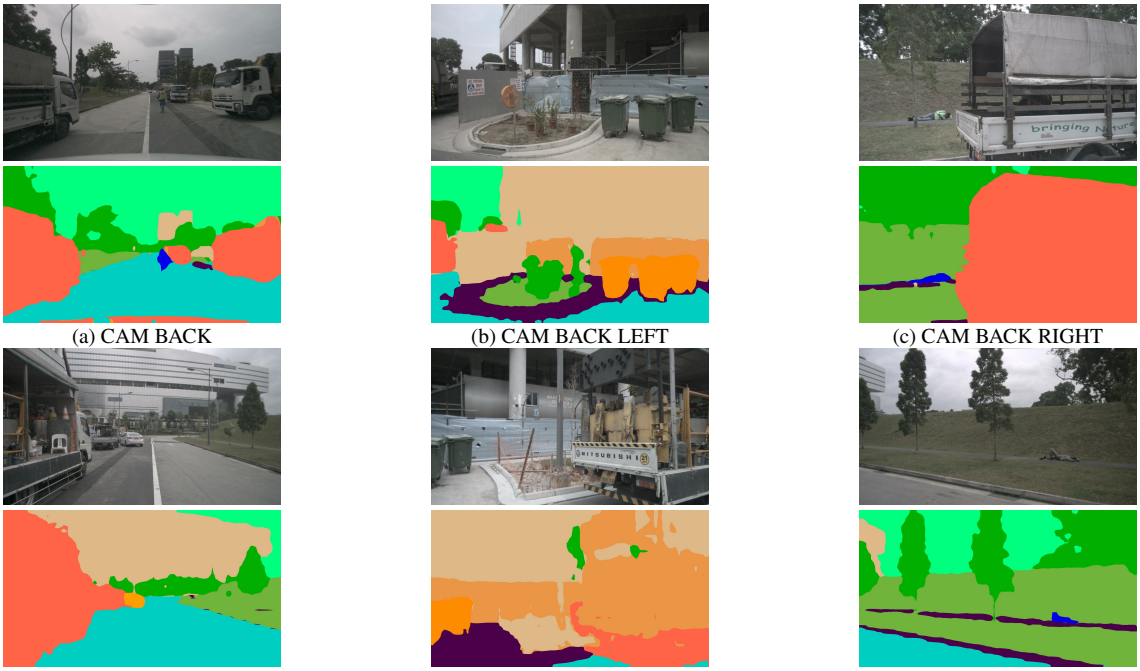
(b) left: chain-of-thought (ours) right: single-step

Figure 4. Ablation on different conversation prompts.

Table 8. Detailed metrics of the pseudo-labeled ground truth from different OV-Seg models on the Occ3D-nuScenes dataset. We report the mean IoU (mIoU) for semantics across different categories. The best results among different methods are in **bold**.

Method	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
SAN [11]	25.53	0.00	9.40	**8.47**	30.16	38.00	11.43	20.20	11.39	**11.54**	**9.55**	25.73	61.86	**0.40**	36.41	33.00	59.42	67.10
ODISE [10]	25.80	0.00	**11.45**	7.83	37.66	**38.10**	6.94	**22.74**	**16.28**	0.52	5.58	24.59	63.09	0.27	37.16	34.12	**62.72**	69.47
CAT-Seg [3]	**26.72**	0.00	3.87	4.56	**41.53**	36.90	**26.65**	17.26	11.28	11.08	3.45	**25.97**	**63.21**	0.34	**37.80**	**36.63**	60.60	**73.03**

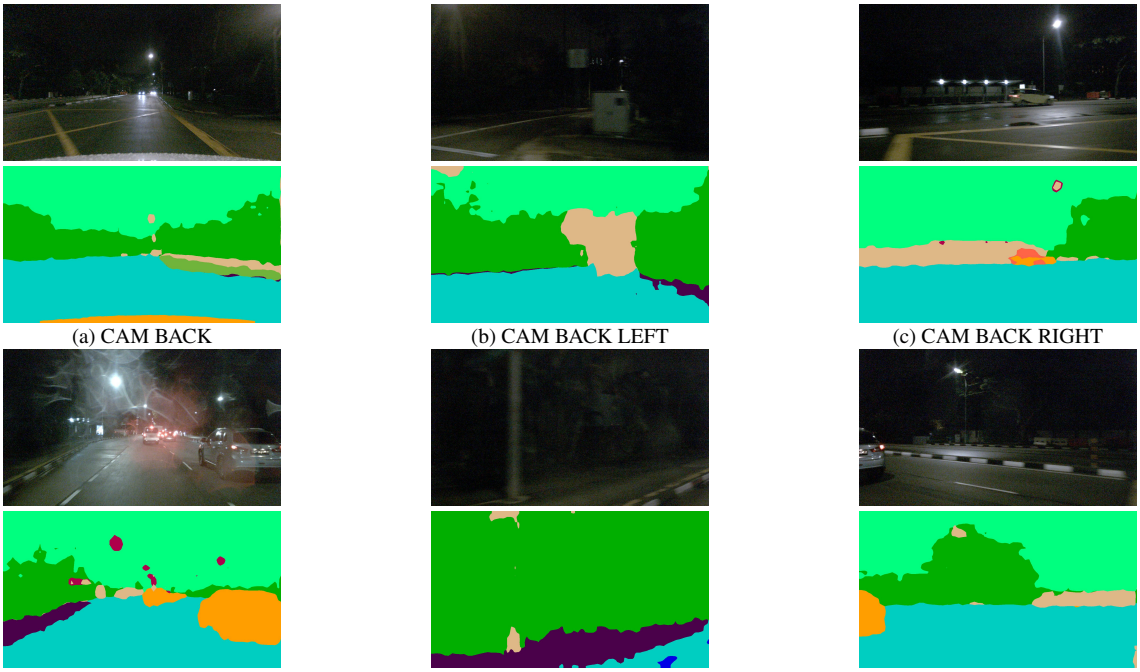
segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 1, 5, 6



(a) CAM BACK (b) CAM BACK LEFT (c) CAM BACK RIGHT

(d) CAM FRONT (e) CAM FRONT LEFT CAM FRONT RIGHT

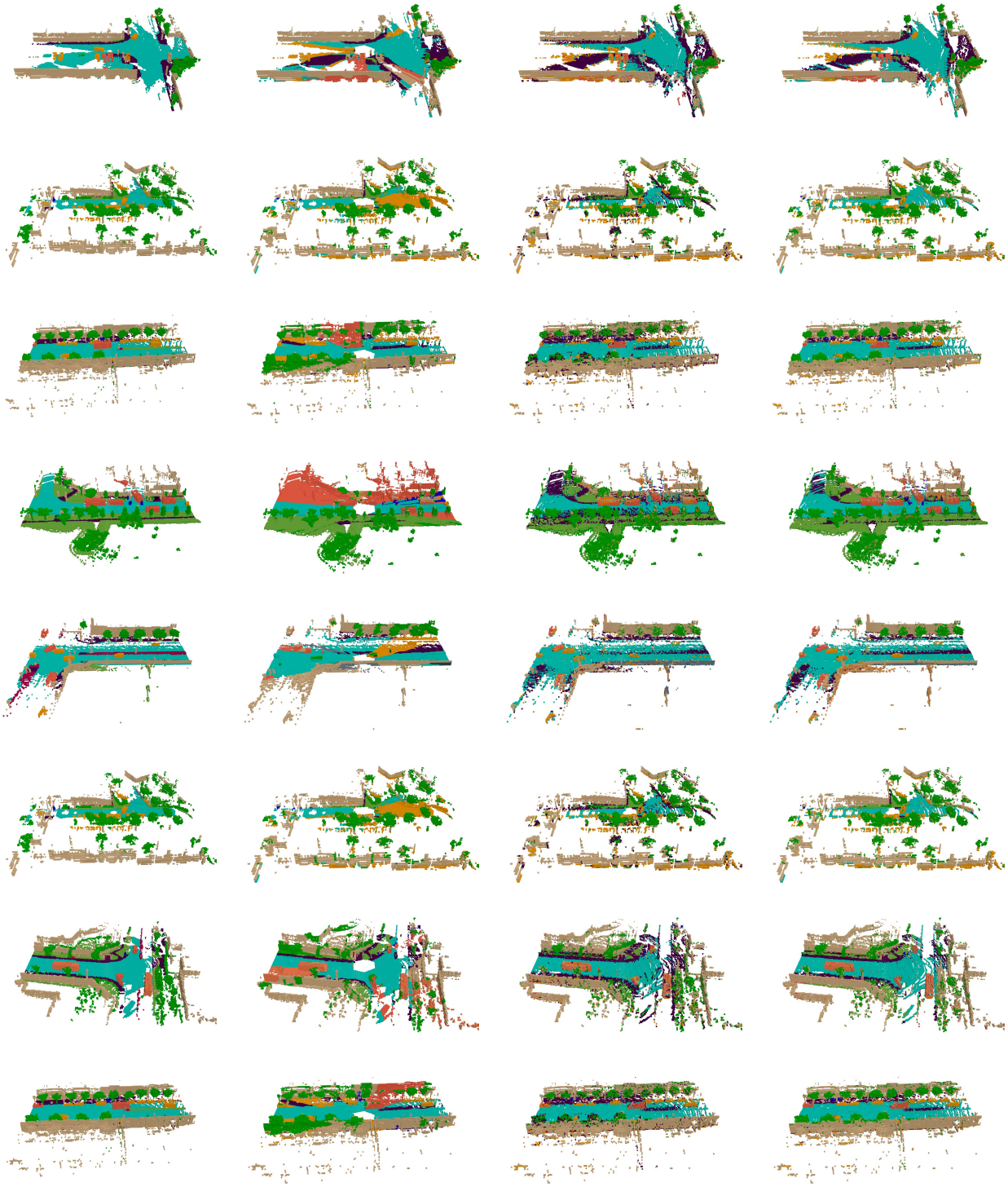
Figure 5. Segmentation results of SAN [11]. Scene token: fff7244095b441d6a053da2951cf2b3b.



(a) CAM BACK (b) CAM BACK LEFT (c) CAM BACK RIGHT

(d) CAM FRONT (e) CAM FRONT LEFT CAM FRONT RIGHT

Figure 6. Segmentation results of SAN [11]. Scene token: 007cbcb1390c440fb48baf3478d1b529.



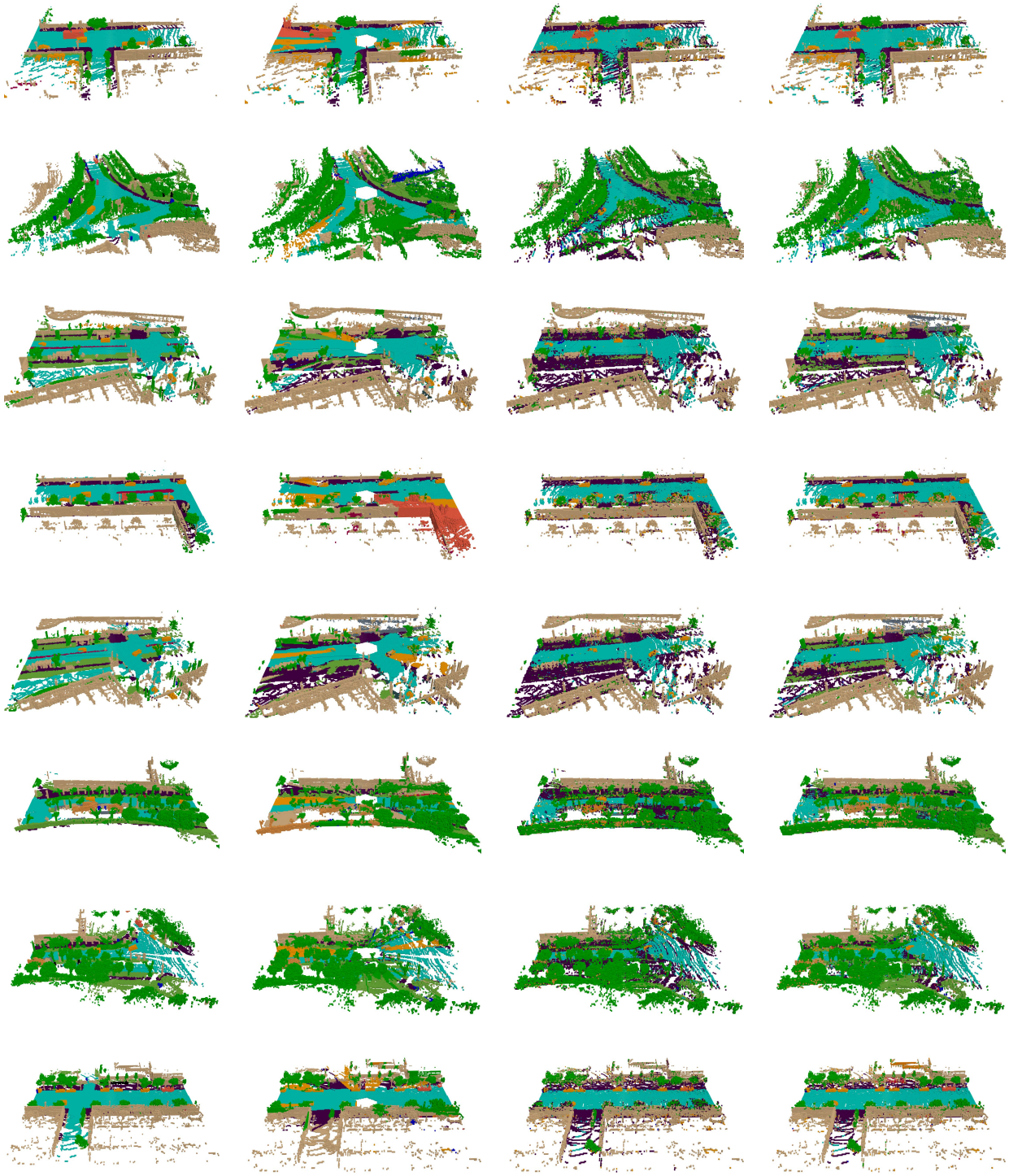
(a) Human-annotated ground truth

(b) Voxel-based model-view projection

(c) Image features as intermediates

(d) Semantic transitive labeling (ours)

Figure 7. Visualization results of pseudo-labeled 3D language occupancy ground truth generated through different pipelines.



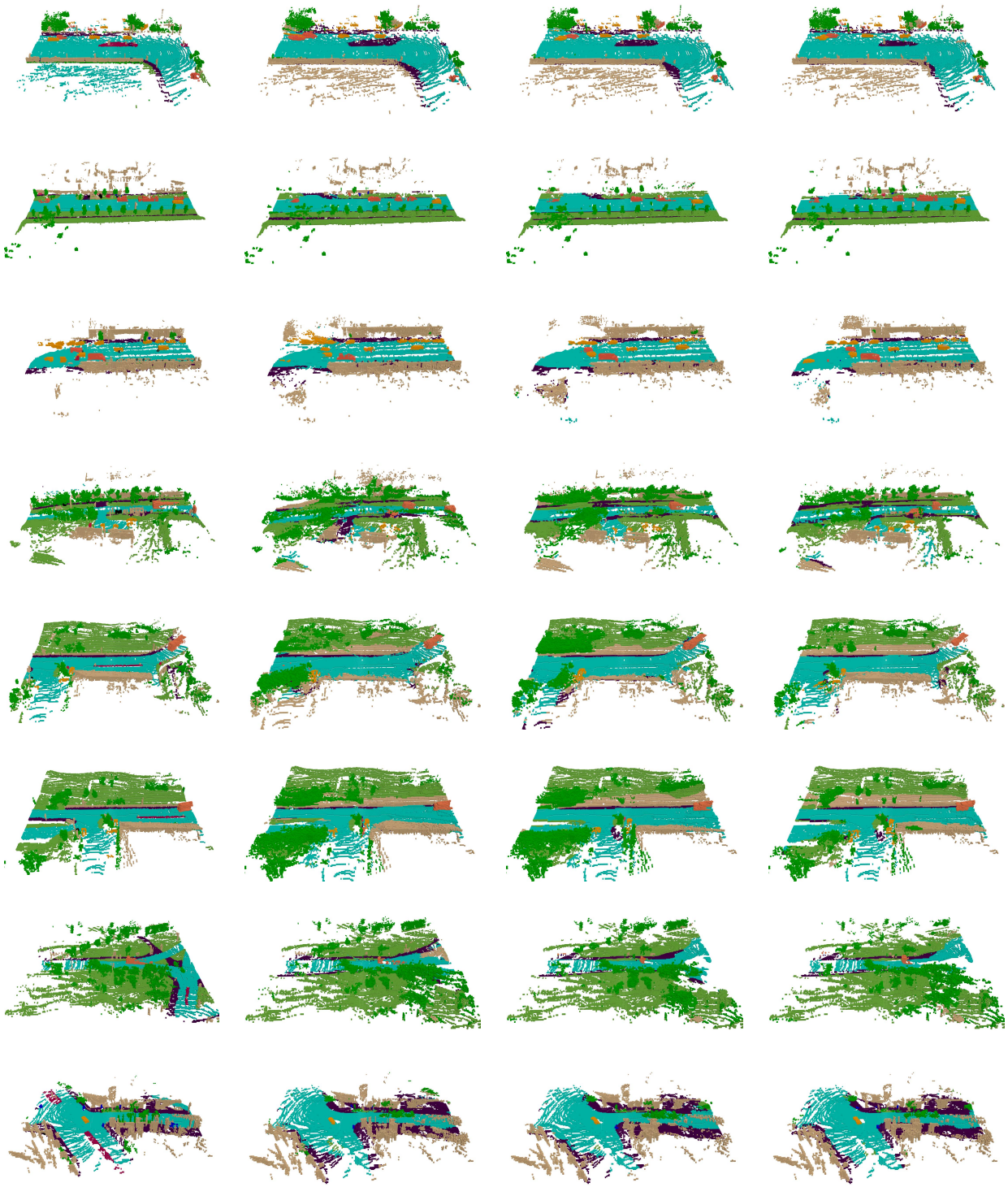
(a) Human-annotated ground truth

(b) Voxel-based model-view projection

(c) Image features as intermediates

(d) Semantic transitive labeling (ours)

Figure 8. Visualization results of pseudo-labeled 3D language occupancy ground truth generated through different pipelines.



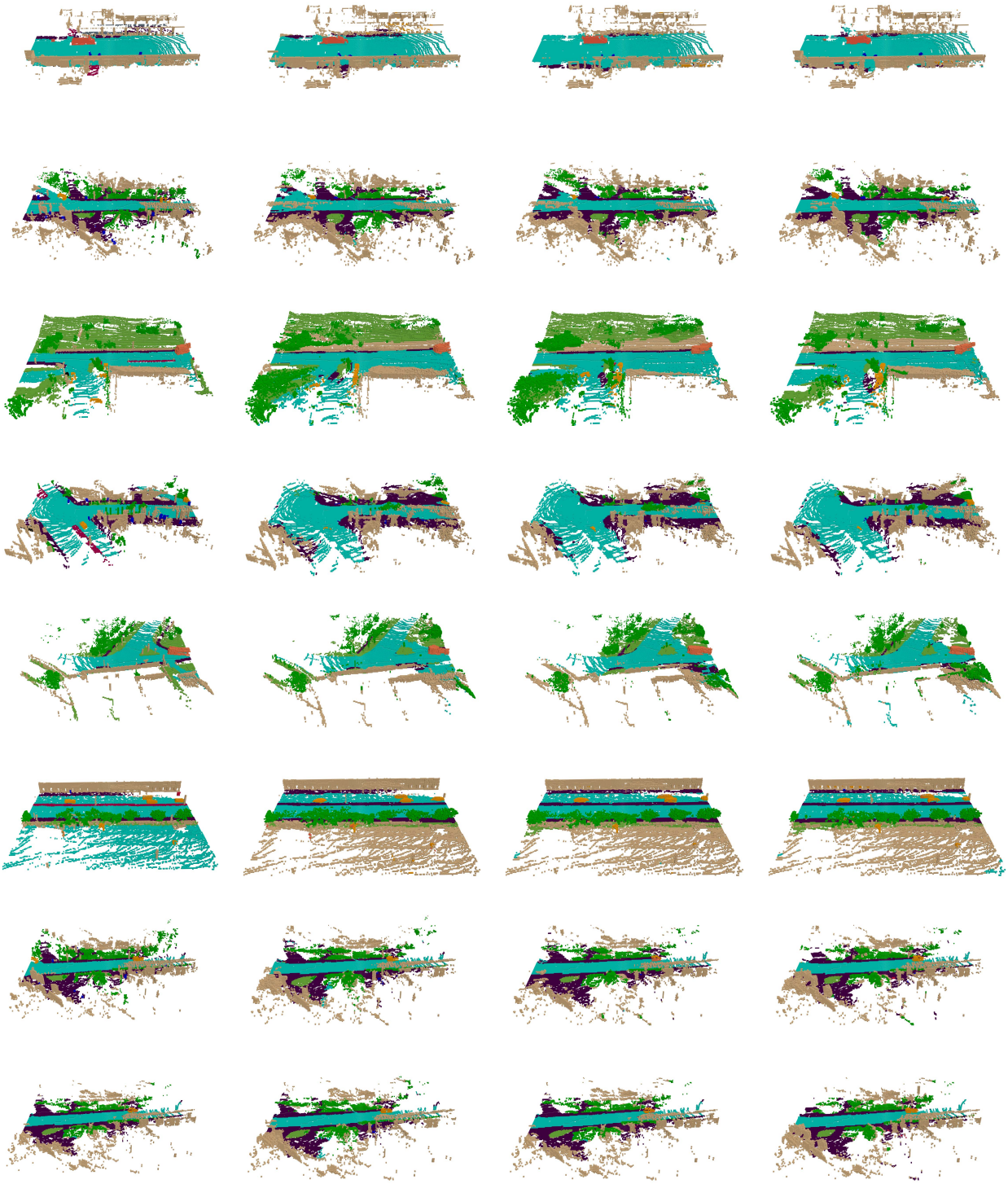
(a) Human-annotated ground truth

(b) LOcc-BEVFormer

(c) LOcc-BEVDet

(d) LOcc-BEVDet4D

Figure 9. Quantitative visualization results on the Occ3D-nuScenes [8]



(a) Human-annotated ground truth

(b) LOcc-BEVFormer

(c) LOcc-BEVDet

(d) LOcc-BEVDet4D

Figure 10. Quantitative visualization results on the Occ3D-nuScenes [8].