OnlinePG: Online Open-Vocabulary Panoptic Mapping with 3D Gaussian Splatting
— Supplementary Material —

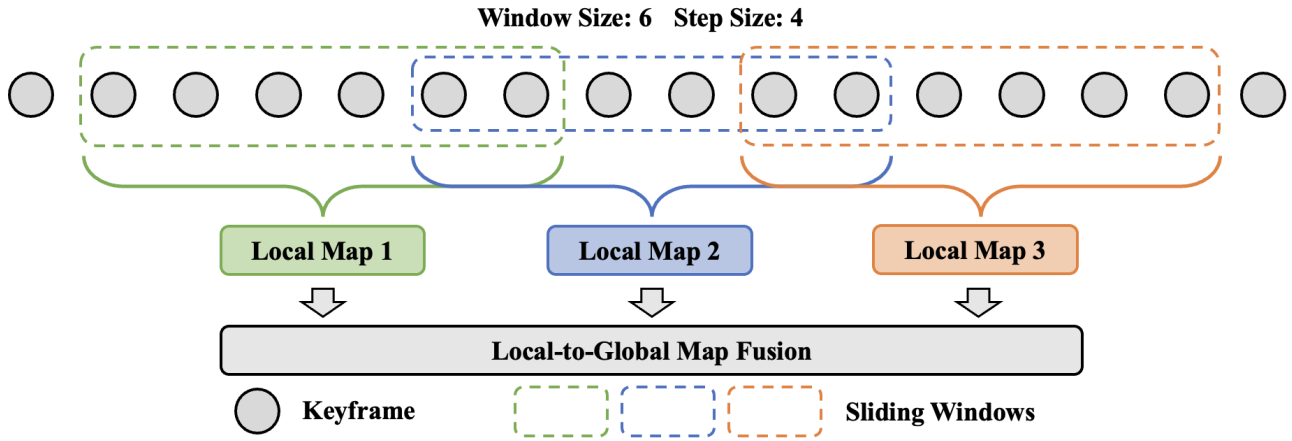

Figure A. **Illustration of Sliding Window**. For a sliding window with a size of 6 and a step size of 4, we perform segment clustering every 4 keyframes and merge the local map into the global map. As shown in the figure, this reduces the reading and updating of the global map, enabling efficient reconstruction.

In this supplementary document, we first provide more implementation details in Sec. A. Next, we supply more visualization results of our methods in Sec. B.

A. Implementation Details

Dataset Setting. For ScanNetV2 [1], we use the following 10 scenes: *scene0000*, *scene0062*, *scene0070*, *scene0097*, *scene0140*, *scene0200*, *scene0347*, *scene0400*, *scene0590*, and *scene0645*. All selected scenes are evaluated on the *00* trajectory. For the evaluation of 3D semantic and panoptic segmentation, we use the 19 classes: *wall*, *floor*, *cabinet*, *bed*, *sofa*, *table*, *door*, *window*, *bookshelf*, *picture*, *counter*, *desk*, *curtain*, *refrigerator*, *shower curtain*, *toilet*, *sink*, and *bathtub*. For Replica dataset [6], the commonly-used 8 scenes $\{$*room0-2*, *office0-4*$\}$ are used for evaluation, and two additional labels, *other furniture* and *ceiling*, are used for evaluation.

Details of Our OnlinePG. Following [4, 9, 10], we adopt CLIP [5] and LSeg [2] as text and image visual-language feature extractors, with feature dimension $D_f = 512$. We use EntitySeg [3] to extract 2D instance segmentation for each keyframe. We sample a keyframe every 20 frames and maintain a sliding window of size 12. Segment clustering and local-to-global map fusion are performed every 7 keyframes. The illustration of sliding window design is shown in Fig. A. For an online image stream, we maintain a fixed-size window and move it in fixed steps. Each time the sliding window moves, it builds a local map for all keyframes in the window and merges it into the global map based on bidirectional matching. The resolutions of the Replica [6] and ScanNetV2 [1] datasets that we used are 640×360 and 640×480, respectively. When performing bidirectional bipartite matching between the local and global maps, we remove matches with scores below the threshold $1/N_{ins}$ to filter out erroneous results, where N_{ins} is the number of candidate matches. For rendering optimization, the learning rates for the 3D Gaussian's location, opacity, scale, color, and rotation are set to 0.00015, 0.05, 0.001, 0.001, 0.01, respectively.

Runtime Analysis. Because the running speed of our method is affected by the number of masks in the 2D image and the number of instances in the 3D scene, we report the running time in 10 selected scenes from ScanNetV2. For the video stream, we select one frame every 20 frames as a keyframe and process it. As shown in Fig. B, we show the runtime of four main parts in our system: (#1) Keyframe

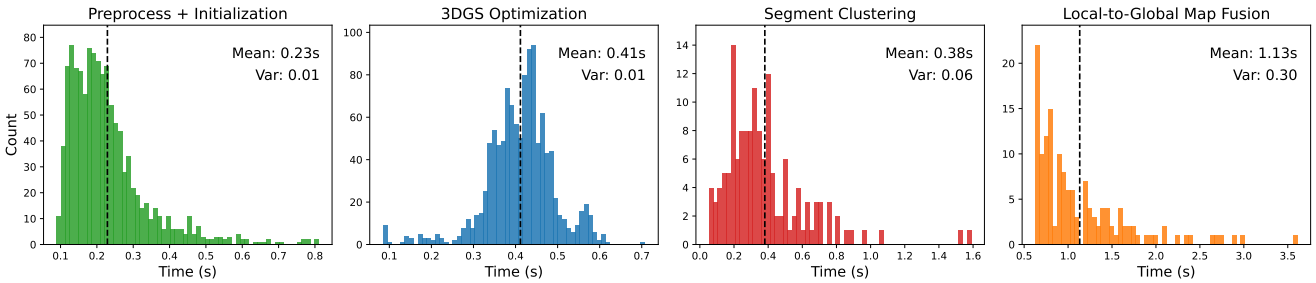

Figure B. **Runtime Performance of OnlinePG**. We divided the system's time consumption into four parts and statistically analyzed the distribution of the execution time of each part across ten scenarios used by ScanNet. The black dashed line represents the average time.

Preprocessing and 3D Segments Initialization, (#2) 3DGS Optimization, (#3) Segment Clustering, and (#4) Local-to-Global Map Fusion. #1 and #2 are called once every time a new keyframe is inserted. #3 and #4 are called only when the sliding window moves a fixed step size.

Besides, the FPS performance of different online methods is also shown in Tab. A and Tab. B. Since OnlineAnySeg [7] needs to obtain the mask and features in advance, the FPS results reported by different methods in the table do not include the time of VLM inference.

B. More Experimental Results

Detailed Performance of Each Scene. The detailed 3D semantic and panoptic segmentation performance of our approach and different online baselines (O2V-Mapping [8] and OnlineAnySeg [7]) are shown in Tab. A and Tab. B, respectively.

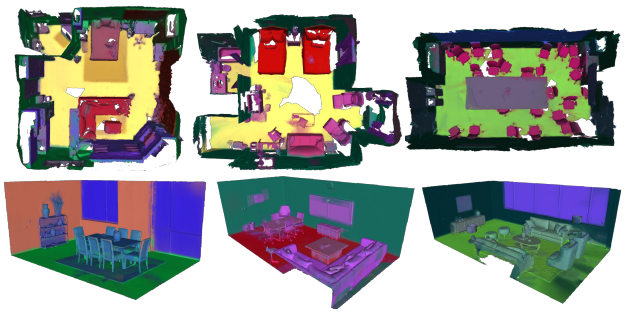

Figure C. **Visualization Results of 3D Language Features**. For better visualization, we perform principal components analysis (PCA) on the high-dimensional language features.

3D Language Feature. In Fig. C, we show the visualization results of the language features reconstructed by our OnlinePG. We used principal component analysis to compress the 512-dimensional features into 3 dimensions for visualization. As can be seen from the figure, the features we obtained can distinguish finer-grained objects, such as carpet and floor, bed and sofa, *etc*. Objects with semantic similarity have similar characteristics, and visualization can show that they have similar colors.

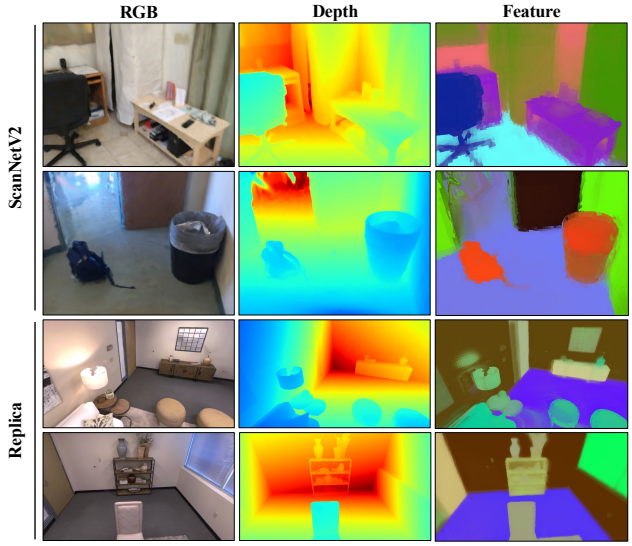

Figure D. **Qualitative Rendering Results**. We show the rendered RGB, Depth, and language feature of different scenes.

Rendering Results. In Fig. D, we present rendering results from four viewpoints on the ScanNetV2 [1] and Replica [6] datasets, including RGB, depth, and language features.

3D Instance Results. In Fig. E, we present 3D instance segmentation results of some scene from the ScanNetV2 [1] and Replica [6] datasets. Different colors represent different 3D instances. Due to the GT mesh of the Replica dataset containing many unobserved areas in the images, this will lead to some noisy instances (on the floor and walls), which will also cause the PRQ (S) of the online method to be worse than that of the offline method.

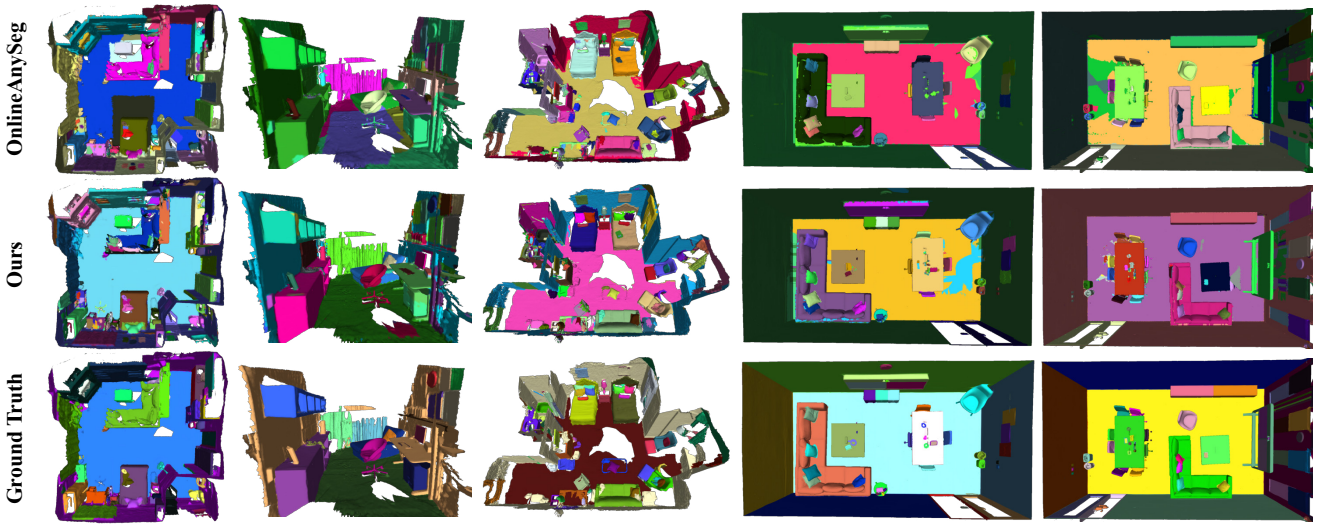

Figure E. **Qualitative 3D Instance Segmentation Comparison**. We show the 3D instance segmentation results from several scenes from ScanNetV2 [1] and Replica [6] datasets.

Table A. **3D Semantic and Panoptic Segmentation Results on Replica Datasets**. We demonstrate the 3D segmentation performance and FPS data of the online method in different scenes of Replica [6].

Method	Metrics	Room 0	Room 1	Room 2	Office 0	Office 1	Office 2	Office 3	Office 4
O2V-Mapping [8]	mIoU	29.25	39.58	32.53	18.78	17.54	20.75	12.59	23.24
	mAcc.	38.00	52.60	43.22	27.56	20.25	30.20	17.95	38.47
	FPS	3.08	3.05	3.08	3.05	3.10	3.13	3.18	3.28
OnlineAnySeg [7]	mIoU	46.76	45.13	51.30	37.54	25.38	32.91	25.45	48.23
	mAcc.	62.41	67.50	66.84	50.47	31.24	46.99	42.46	59.45
	PRQ (T)	30.26	51.65	26.59	45.34	22.10	24.90	16.32	27.17
	PRQ (S)	8.58	15.99	10.97	8.65	2.73	7.15	4.83	11.46
	FPS	9.90	13.06	12.91	15.40	20.94	11.65	10.81	11.45
Ours	mIoU	53.60	67.16	68.31	32.07	10.19	53.53	54.95	43.55
	mAcc.	59.34	80.17	75.86	36.28	16.07	60.06	60.86	50.96
	PRQ (T)	31.03	57.49	51.14	48.45	0.00	48.40	47.46	43.89
	PRQ (S)	11.75	17.22	12.44	7.37	4.47	17.48	9.20	18.05
	FPS	12.07	13.91	14.07	15.85	18.22	14.77	12.42	14.14

Table B. **3D Semantic and Panoptic Segmentation Results on ScanNetV2 Datasets**. We demonstrate the 3D segmentation performance and FPS data of the online method in different scenes of ScanNetV2 [1].

Method	Metrics	0000	0062	0070	0097	0140	0200	0347	0400	0590	0645
O2V-Mapping [8]	mIoU	35.80	46.09	41.21	26.29	24.25	39.08	27.54	39.74	33.18	28.27
	mAcc.	59.92	72.72	55.59	55.73	34.15	63.91	47.12	58.33	49.85	52.22
	FPS	3.38	3.43	3.43	3.53	3.40	3.50	3.43	3.38	3.43	3.38
OnlineAnySeg [7]	mIoU	32.64	31.13	23.26	39.08	12.62	39.32	32.97	33.93	34.66	33.26
	mAcc.	50.61	60.73	31.92	69.62	31.86	52.43	54.51	56.42	55.89	58.02
	PRQ (T)	42.32	33.47	36.29	71.30	53.32	33.76	38.43	25.97	38.95	56.06
	PRQ (S)	18.53	38.20	12.89	41.70	7.94	39.09	20.35	38.26	24.50	21.32
	FPS	15.57	22.91	26.44	22.83	15.72	19.27	20.41	25.22	22.98	16.93
Ours	mIoU	39.75	75.09	46.50	52.59	48.04	47.84	52.76	39.44	37.95	44.92
	mAcc.	61.77	90.36	62.95	70.39	64.54	62.11	72.50	62.18	53.33	60.05
	PRQ (T)	20.21	35.28	49.55	58.63	51.67	46.56	42.08	0.00	33.55	42.16
	PRQ (S)	40.05	63.27	24.34	52.79	39.28	43.94	36.68	56.53	33.36	27.88
	FPS	13.33	18.83	15.68	17.57	14.23	18.48	17.86	16.89	15.26	14.08

References

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2432–2443, 2017. 1, 2, 3

[2] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 1

[3] Qi Lu, Jason Kuen, Shen Tiancheng, Gu Jiuxiang, Guo Weidong, Jia Jiaya, Lin Zhe, and Yang Ming-Hsuan. High-quality entity segmentation. In *IEEE/CVF International Conference on Computer Vision*, 2023. 1

[4] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 1

[6] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1, 2, 3

[7] Yijie Tang, Jiazhao Zhang, Yuqing Lan, Yulan Guo, Dezun Dong, Chenyang Zhu, and Kai Xu. Onlineanyseg: Online zero-shot 3d segmentation by visual foundation model guided 2d mask merging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3676–3685, 2025. 2, 3

[8] Muer Tie, Julong Wei, Ke Wu, Zhengjun Wang, Shanshuai Yuan, Kaizhao Zhang, Jie Jia, Jieru Zhao, Zhongxue Gan, and Wenchao Ding. O2v-mapping: Online open-vocabulary mapping with neural implicit representation. In *European Conference on Computer Vision*, pages 318–333, 2024. 2, 3

[9] Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, and Jian Zhang. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. In *Advances in Neural Information Processing Systems*, 2024. 1

[10] Hongjia Zhai, Hai Li, Zhenzhe Li, Xiaokun Pan, Yijia He, and Guofeng Zhang. Panogs: Gaussian-based panoptic segmentation for 3d open vocabulary scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14114–14124, 2025. 1