— Supplementary Material —
InstanceGaussian: Appearance-Semantic Joint Gaussian Representation for 3D Instance-Level Perception

Haijie Li[1] Yanmin Wu[1] Jiarui Meng[1] Qiankun Gao[1] Zhiyao Zhang[3]
Ronggang Wang[1,2] Jian Zhang[1,2*]

[1]School of Electronic and Computer Engineering, Peking University, China
[2]Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology, Peking University Shenzhen Graduate School, China
[3]College of Information Science and Engineering, Northeastern University, China

In the supplementary materials, we provide additional experimental results and implementation as follows.

Contents

1. Implementation Details

1.1. Training Strategy.

For ScanNet [3] dataset, we freeze the point cloud coordinates and disable 3DGS [5] densification. For the LeRF [6] dataset, we optimize the point cloud coordinates and enable 3DGS densification. We stop 3DGS densification in 10k steps.

1.2. Training Time

We train each scene on a single 24G 3090 GPU (with actual memory usage around 5 to 10 GB). For the LeRF dataset, each scene takes around 200 images and trains for approximately 70 minutes. For the ScanNet dataset, each scene takes around 100-300 images and trains for approximately 30 minutes. Tab. 1 shows the comparison with the baseline method in training time. We selected level 3 to extract the SAM [7] mask.

Method	Langslpat	LEGaussian	OpenGaussian	Ours
Time (s)	1153	1011	1378	1847

Table 1. Comparation of training time in Scannet.

1.3. ScanNet Dataset Setting

We randomly selected 10 scenes from ScanNet for evaluation, specifically: scene0000_00, scene0062_00, scene0070_00, scene0097_00, scene0140_00, scene0200_00, scene0347_00, scene0400_00, scene0590_00, scene0645_00. The 19 categories (defined by ScanNet) used for text query are respectively: wall, floor, cabinet, bed, chair, sofa, table, door, window, bookshelf, picture, counter, desk, curtain, refrigerator, shower curtain, toilet, sink, bathtub; 15 categories are without picture, refrigerator, showercurtain, bathtub; 10 categories are further without cabinet, counter, desk, curtain, sink. We downsampled the training images by a factor of 2 and selected SAM level 3 to extract supervision signals.

For the fixed point clouds in the ScanNet dataset, suboptimal processing often leads to degraded visual quality. To address this, we identified well-optimized point clouds based on their contributions during rendering and utilized them to train a lightweight MLP. The MLP takes position and color as inputs and predicts segmentation labels as outputs. Subsequently, the trained MLP is employed to infer segmentation results for the point clouds. This approach yields smoother segmentation outcomes, significantly enhancing visual quality.

2. Comparison with more Methods

We conducted comparisons with additional baselines in Tab. 2, demonstrating the superior performance of our method.

3. Ablation Study

3.1. More Detailed Ablation Metrics

We conducted a detailed ablation study with mA50 and mA25 metric in Tab. 3 and Tab. 4. We chose joint representation and joint training as the best strategy for the segmentation topic, as mIoU is generally more reasonable than mAcc.

3.2. Hyperparameters

We conducted ablation experiments of hyperparameters in Tab. 5, demonstrating the robustness of our instantiated algorithm. The hyperparameters reported in our paper were not carefully tuned, and we discovered even better performance when adjusting them.

Method	Semantic seg.		Instance seg.	
	mIoU	mAcc.	mIoU	mAcc.
GAGA[8]	*-unsupported-*		16.76	31.49
SAGA[2]	9.44	16.23	34.16	70.14
Ours	**40.66**	**54.01**	**50.27**	**80.22**

Table 2. Comparison with more methods.

Joint		Semantic seg		Instance Seg		
Rep.	Train.	mIoU	mAcc	mIoU	mA50	mA25
✓		30.71	44.22	47.4	44.51	**84.42**
	✓	33.15	44.45	49.57	**52.82**	82.17
✓	✓	**40.66**	**54.01**	**50.27**	52.57	80.22

Table 3. Detailed ablation of joint representation and training.

Condition		Semantic seg		Instance Seg		
Feat.	Vox.	mIoU	mAcc	mIoU	mA50	mA25
	✓	21.75	33.39	27.50	16.16	49.33
✓		28.98	40.58	43.41	40.63	72.50
✓	✓	**40.66**	**54.01**	**50.27**	**52.57**	**80.22**

Table 4. Detailed ablation of aggregation condition.

hyperparameters	sample number 200∼1000	voxel size 0.05∼0.5	connectivity thredshold 0.06∼0.18
semantic seg. mIoU	37.87±2.79	38.96±1.70	38.48±2.17
instance seg. mIoU	49.60±1.31	49.56±0.71	49.90±1.29

Table 5. Abliation of hyperparameters.

3.3. Randomness and Runtime of FPS

We randomly sampled different starting points for FPS in Tab. 6, and the results demonstrated the robustness of our method. We also evaluated runtime across different sample number (K) in Tab. 7. While the merging algorithm has a complexity of O(K^2), K shows limited impact on overall runtime in practice (up to 3.43%).

Sample times	Semantic Seg. mIoU	Instance Seg. mIoU
5	39.44±1.38	48.57±0.98

Table 6. Ablation experiment of randomness in FPS.

Variation	Merging time(s)	Total time(s)	Percentage
200∼1000	5∼63	1772∼1847	0.28%∼3.43%

Table 7. Ablation experiment of different K.

3.4. Aggregation Algorithm

We compare our aggregation algorithm against clustering features by HDBSCAN[1] in Tab. 8, demonstrating our superiority.

Method	Semantic Seg. mIoU	Instance Seg. mIoU
HDBSCAN	31.84	44.76
Ours	**40.66**	**50.27**

Table 8. Ablation of aggregation algorithm.

4. More Visual Results

To demonstrate the effectiveness of our approach, we conducted additional experiments on ScanNet[3] to prove our performance on category-agnostic 3D instance segmentation (Fig. 1) and open-vocabulary query point cloud understanding (Fig. 2). Additionally, we provide more 2D instance segmentation results on LeRF[6] (Fig. 3). We also performed experiments on GraspNet [4] dataset (Fig. 4), and the results indicate the generalization capabilities of our method.

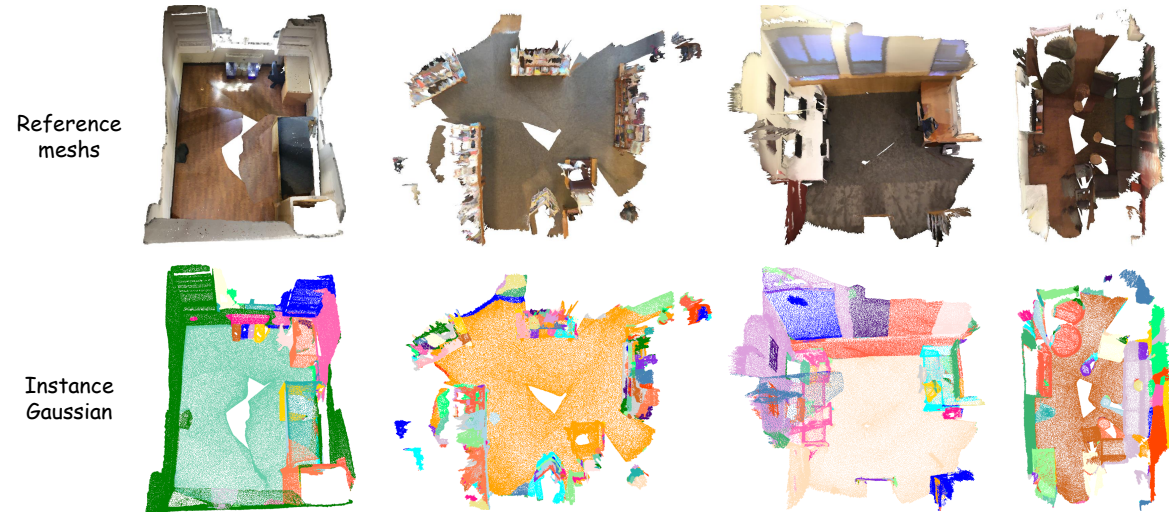

Figure 1. Top row: Reference mesh of scenes. Bottom row: The visualization result of category-agnostic 3D instance segmentation in ScanNet dataset.

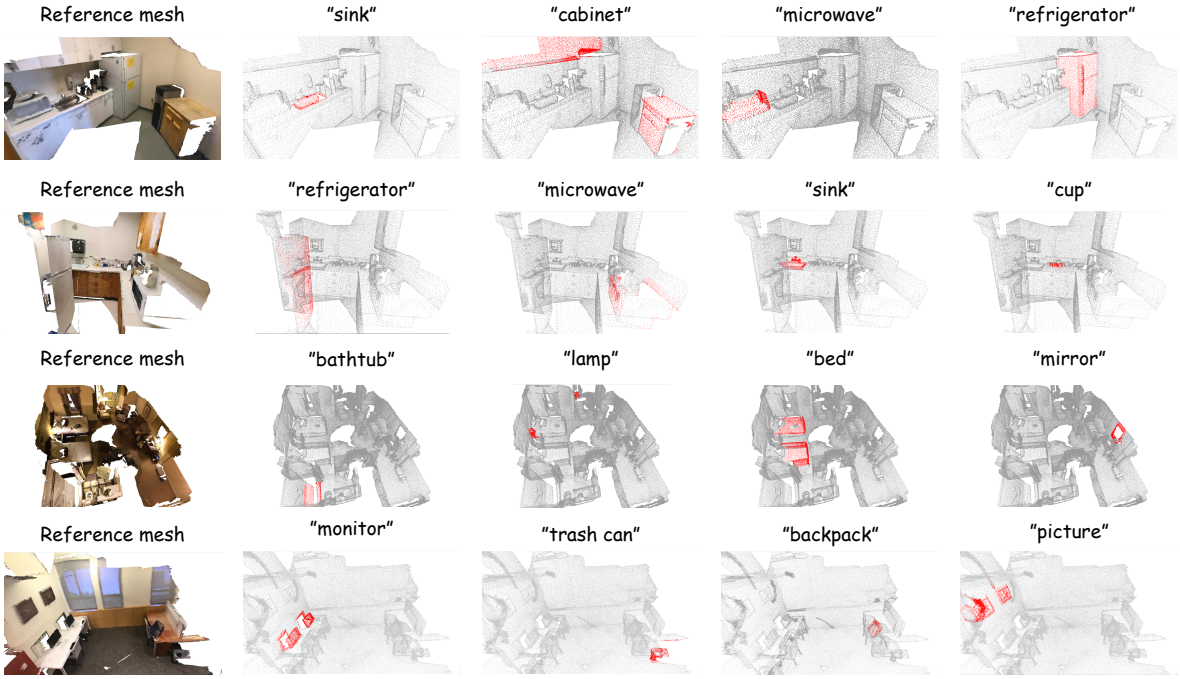

Figure 2. The visualization result of open-vocabulary query point cloud understanding in ScanNet [3] dataset.

Additionally, we visualized our feature maps (Fig. 5) to validate the effectiveness of appearance-semantic joint Gaussian representation.

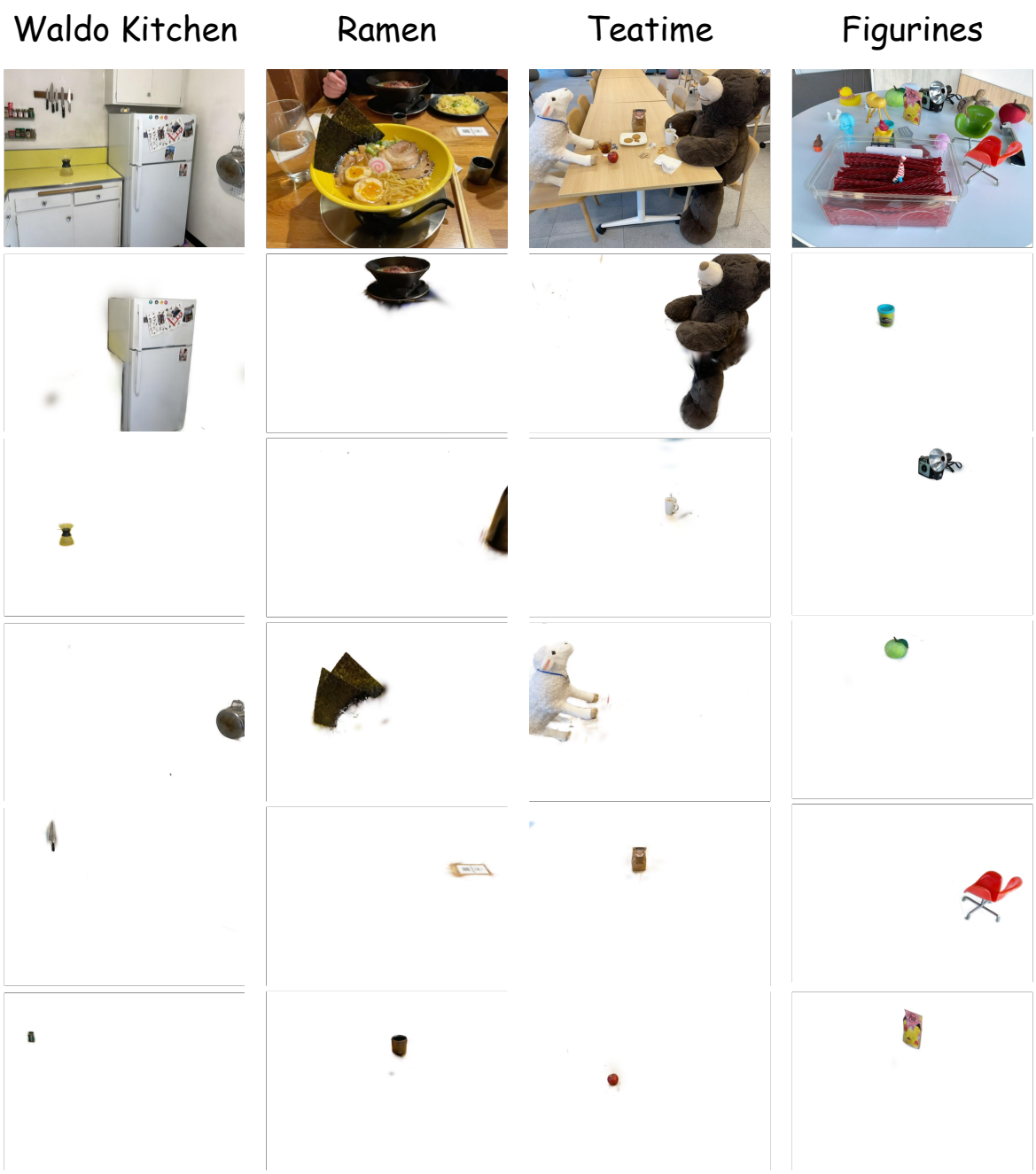

Figure 3. The 2D visualization result of 3D instance segmentation in LeRF dataset.

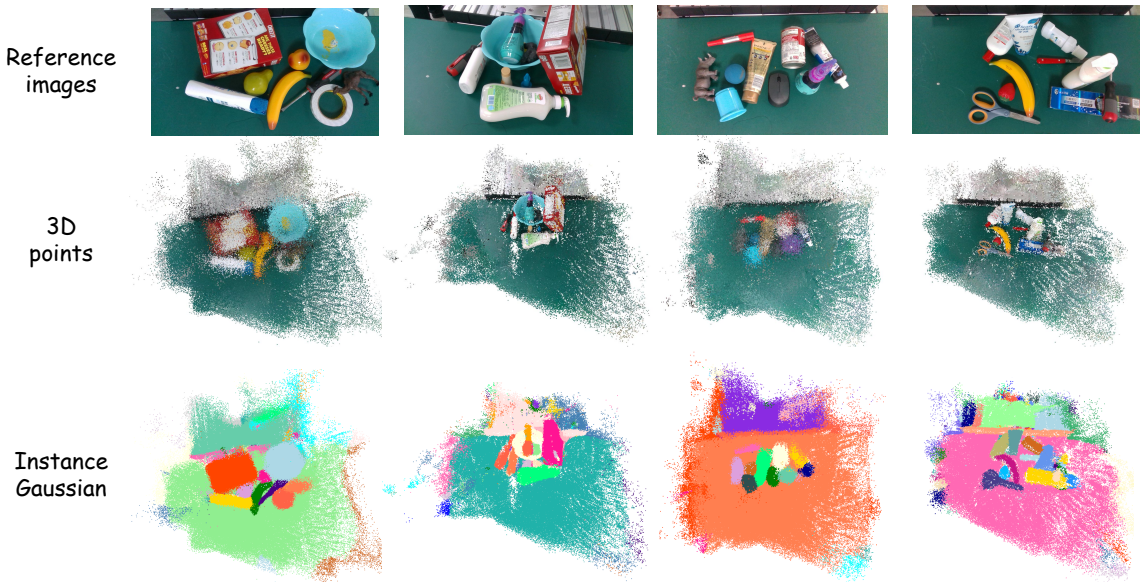

Figure 4. Top: Reference image of scenes. Middle: Constructed 3D Gaussians/points. Bottom: The visualization result of category-agnostic 3D instance segmentation in GraspNet dataset.

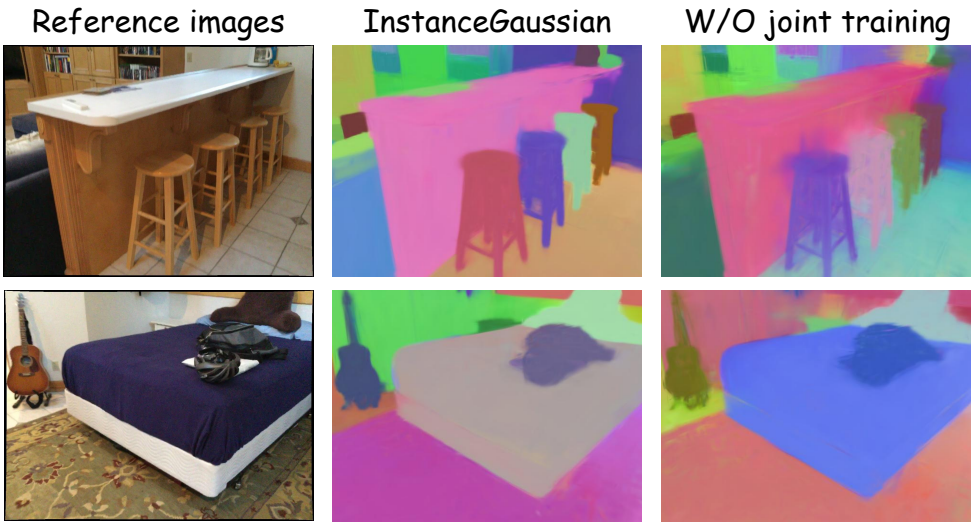

Figure 5. The feature map of InstanceGaussian. Left: Reference images of the scene. Middle: The visualization result of feature maps of InstanceGaussian. Right: The visualization result of feature maps without joint training.

Figure 6. The failure case of ramen in LeRF dataset. Frequent mask failures will undermine the ability to distinguish different food in the bowl.

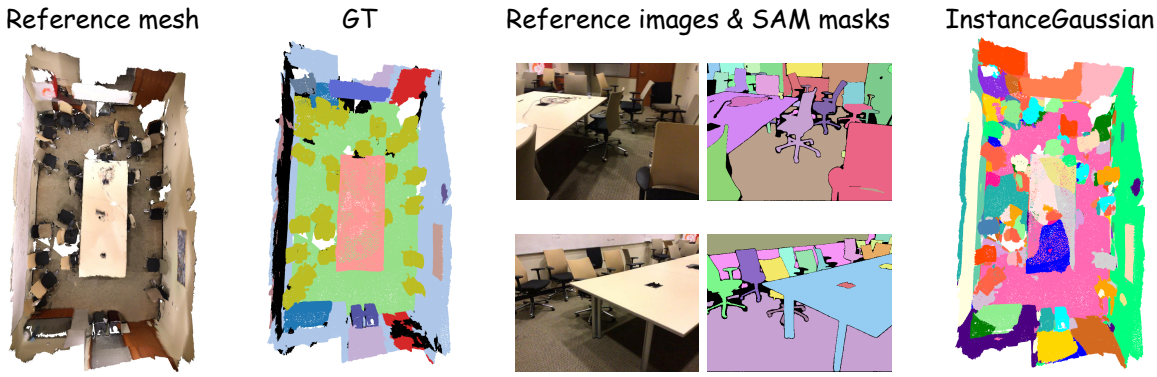

Figure 7. The failure case of aggregate the large meeting table in scene0140_00. Due to the table's large size and the lack of fully captured views, our method struggles to learn consistent features and achieve accurate aggregation, even when the SAM mask is correct.

5. Analysis of Failure Cases

We analyze two main issues affecting the model's performance.

Incorrect SAM mask. Our method demonstrates robustness against sparse segmentation errors when SAM masks are predominantly accurate. However, frequent mask failures prevent effective learning of region-specific object features, ultimately compromising segmentation accuracy (Fig. 6).

Failure in learning consistent feature for large object. Our method demonstrates superior performance in aggregating small objects (e.g., doors, chairs, televisions) but encounters challenges with large-scale objects (e.g., floors, meeting tables). Due to the limited coverage of individual photographs, large objects are often only partially captured, with few complete observations across views. Under such conditions, $\mathcal{L}_s$ (Eq. 2) fails to learn consistent features for complete objects and $\mathcal{L}_c$ (Eq. 3) amplifies the divergence between parts (Fig. 7). In contrast, small objects benefit from more complete observations, enabling $\mathcal{L}_s$ (Eq. 2) to learn coherent features.

References

[1] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172, 2013. 3

[2] Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians. *arXiv preprint arXiv:2312.00860*, 2023. 2

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 4

[4] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 2023. 2

[6] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[8] Weijie Lyu, Xueting Li, Abhijit Kundu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Gaga: Group any gaussians via 3d-aware memory bank, 2024. 2