

Supplementary Materials: GOI: Find 3D Gaussians of Interest with an Optimizable Open-vocabulary Semantic-space Hyperplane

Anonymous Authors

A ADDITIONAL IMPLEMENTATIONAL DETAILS

Our work is based on pretrained vanilla Gaussian scenes. Subsequent to this fundamental step, we embark on a procedure of semantic field optimization, comprising 1500 iterations. Throughout this period, our principal focus is on the optimization of the semantic field, while maintaining the stasis of other parameters. In this stage, we resort to the default values of the unrelated hyperparameters in 3D Gaussian Splatting [2] for anything outside of semantic field optimization.

A.1 Trainable Feature Clustering Codebook

We incorporate a low-dimensional semantic feature with 10 dimensions f within each 3D Gaussian. By default, the Trainable Feature Clustering Codebook (TFCC) is configured with $N = 300$ entries. As a result, the input dimension of MLP decoder \mathcal{D} is set to 10, while the output logits e from \mathcal{D} are a 300-dimensional vector. Importantly, the decoder \mathcal{D} is simplified to contain solely a lone fully-connected layer, deemed sufficient for efficacious feature decoding.

In order to augment the efficiency of reconstruction, k -means clustering is employed for initializing the TFCC. Between 30 to 50 feature maps are sampled from densely observed viewpoints. Subsequently, for each pixel-wise feature, we adopt the k -means clustering based on the cosine similarity amid features.

The resultant loss in the course of the TFCC and low-dimensional feature f optimization is

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_T + \lambda_{joint} \mathcal{L}_{joint} + \lambda_{e2e} \mathcal{L}_{e2e} \\ &= \lambda_{ent} \mathcal{L}_{ent} + \lambda_{max} \mathcal{L}_{max} + \lambda_{joint} \mathcal{L}_{joint} + \lambda_{e2e} \mathcal{L}_{e2e},\end{aligned}\quad (1)$$

We allocate a weightage of $\lambda_{ent} = 0.3$ for \mathcal{L}_{ent} , whilst the remainder are set as 1. The annealing temperature τ derived from \mathcal{L}_{ent} begins at 1, escalating to 2 post 1000 iterations.

A.2 Optimizable Semantic-space Hyperplane

We use the Grounded-SAM [4] model as our Referring Expression Segmentation (RES) model. The text query t and the RGB image are processed by the RES model to generate a binary mask \hat{m} of the target area as the pseudo-label. This mask is subsequently used with m in logistic regression to optimize W and b . We fine-tune the OSH with the objective:

$$\mathcal{L}_{OSH} = -\frac{1}{P} \sum_{i=1}^P [w \cdot \hat{m}_i \log(\sigma(m_i)) + (1 - \hat{m}_i) \log(1 - \sigma(m_i))], \quad (2)$$

where P denotes all samples, $\sigma(\cdot)$ denotes Sigmoid function, w is a hyperparameter. Considering that regions of interest tend to be significantly smaller than non-interest regions, we set $w = \frac{1}{10}$ to increase the penalty weight for misclassifying target areas, thereby accelerating convergence.

B EXPERIMENTAL DETAILS

B.1 Expanding the Mip-NeRF360 Dataset

Within each of the four selected scenes (Room, Bonsai, Garden, and Kitchen) from the Mip-NeRF360 dataset [1], we’ve identified four notably distinctive objects. For every individual object, we’ve established ten distinct viewpoints in the scenario, and employed the SAM [3] ViT-H model to generate object masks for these pre-selected perspectives. Moreover, we present textual descriptions founded on either the appearance of the chosen objects (e.g., “sofa in dark green”), or their spatial relationship with other objects (e.g., “table under the bowl”). Consequently, our expanded evaluation set for Mip-NeRF360 includes tuples encapsulating the viewpoint image, ground truth mask, and a concise text description.

We have listed the textual descriptions of each individual object selected within the scenes in Table 1. Additionally, in Figure 2, we exhibit the ground truth segmentation masks pertinent to select objects in our expanded Mip-NeRF360 evaluation dataset.

Scene	Text Description
Room	bowl on the table, brown slipper, sofa in dark green, table under the bowl
Bonsai	black chair, flowerpot on the table, orange bottle, purple table
Garden	brown table, flowerpot on the table, green football, green grass
Kitchen	chair, red gloves, table mat, wooden table

Table 1: Text description for select objects of each scene in our extended version of the Mip-NeRF360 evaluation dataset.

B.2 More Results

B.2.1 Qualitative Results. Figure 1 serves as a visual representation of our comprehensive query results derived from the Mip-NeRF360 dataset. The effect of executing queries on an identical object, but from varying viewpoints, is lucidly demonstrated. The takeaway is that our outcomes have effectively demarcated the object boundaries and simultaneously exhibited consistency when observed from multiple viewpoints.

B.2.2 Quantitative Results. We base our evaluation on metrics such as mean Intersection over Union (mIoU), mean Pixel Accuracy (mPA), and mean Precision (mP), akin to the LEGaussian [5] method. The efficiency and efficacy of our approach have previously been demonstrated. Furthermore, Tables 2 and 3 provide a detailed exposition of our scene-level metrics derived from the Mip-NeRF360 [1] and Replica [6] datasets. Notably, our proposed methodology consistently outperforms, irrespective of the scene encompassing

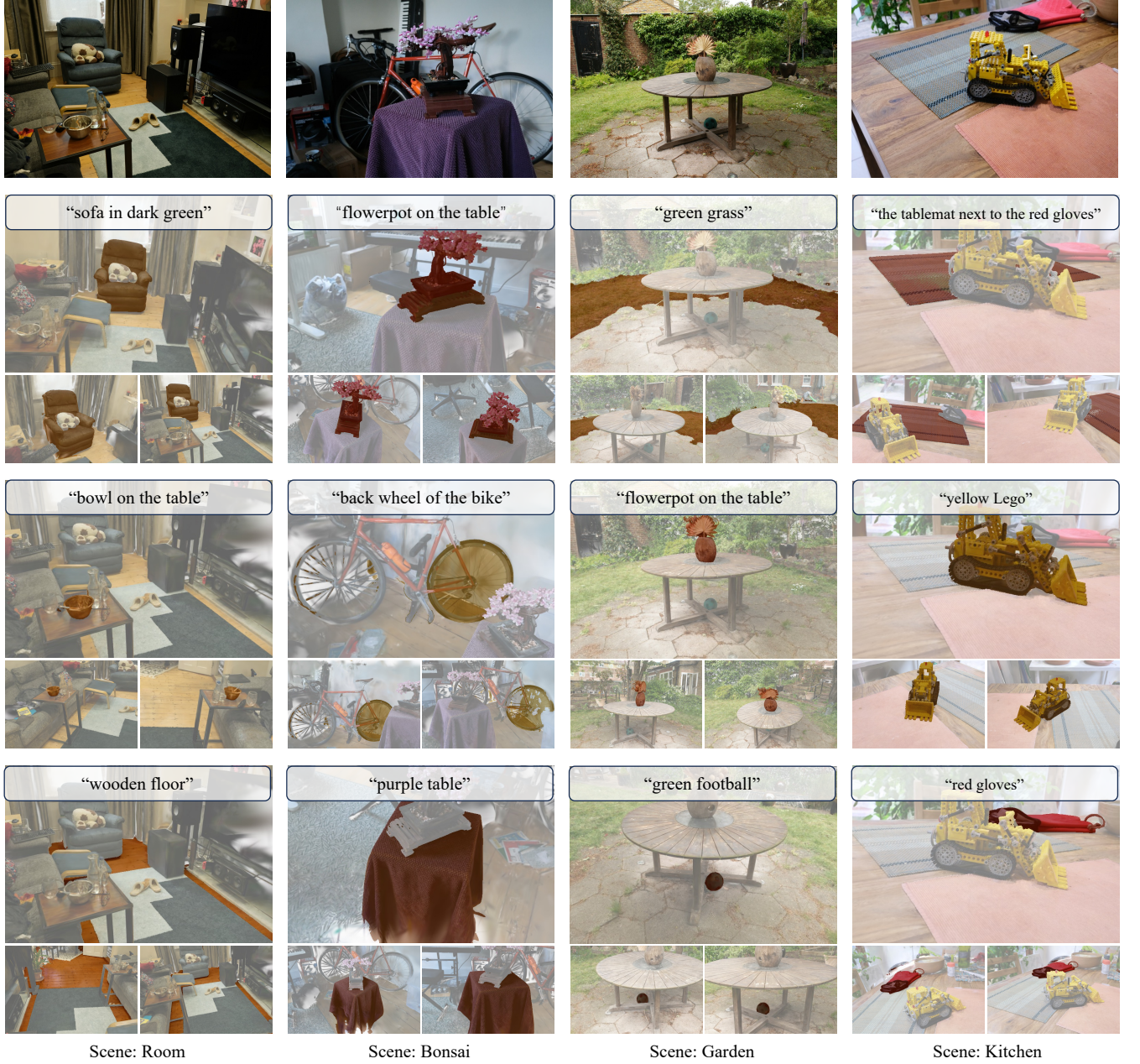


Figure 1: Extensive query visualization on the Mip-NeRF360 dataset. In each column, the images delineated on the top row and the descriptions in the bottom line typify the scene under examination. Within each depicted scene, we have identified three distinct objects to constitute our query. Three distinctive viewpoints from the same scene are exhibited for every given prompt.

the datasets. Additionally, we provide a video that juxtapose our methodology with others, facilitating a more effective elucidation of our superior performance.

B.3 3D Manipulations

As addressed in Sec. 3.5, the low-dimensional feature f in 3D Gaussians and the rendered 2D pixel-wise feature \hat{f} are fundamentally

equivalent. We can also retrieve the high-dimensional semantic feature v for the feature f , as depicted in the following equation.

$$v = \mathcal{T} \left[\underset{j=1,2,\dots,N}{\operatorname{argmax}} (e_j) \right], \text{ where } e = \mathcal{D}(f) \quad (3)$$

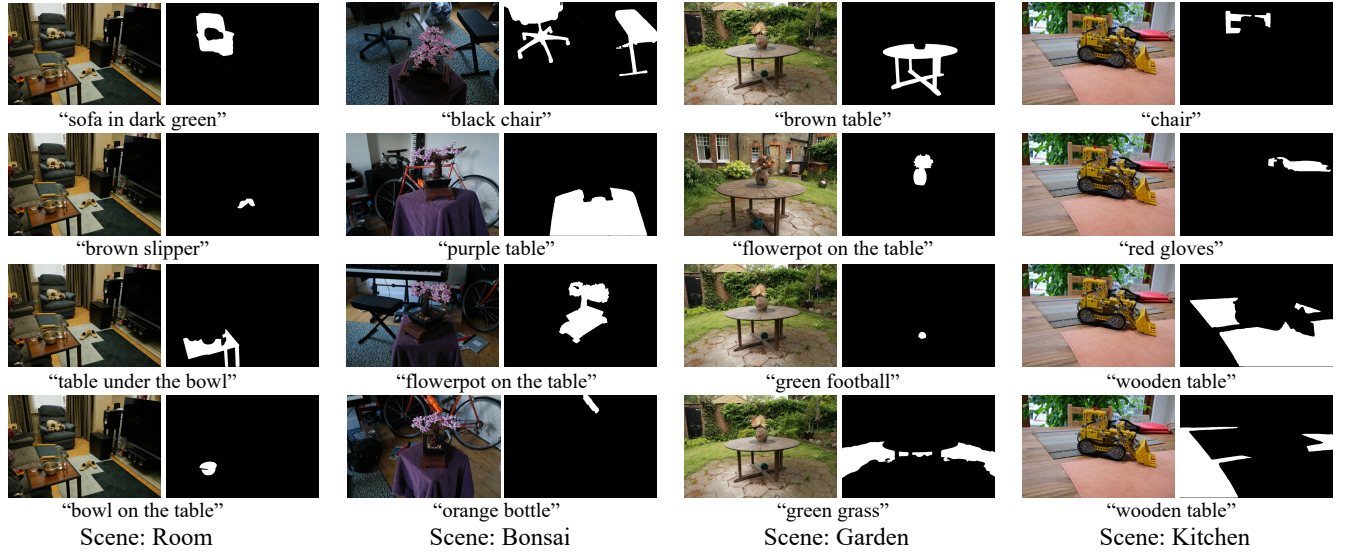


Figure 2: Ground truth segmentation masks for select objects in our extended version of the Mip-NeRF360 evaluation dataset.

Scene	Metric	LERF	Feat. 3DGS	Works GS Grouping	LangSplat	Ours
Room	mIoU	0.0806	0.1748	0.4909	0.6263	0.8504
	mPA	0.8458	0.8246	0.8190	0.9104	0.9718
	mP	0.5400	0.5919	0.7663	0.8442	0.9485
Bonsai	mIoU	0.3214	0.4623	0.4305	0.5914	0.9147
	mPA	0.8852	0.8027	0.8244	0.8083	0.9630
	mP	0.6603	0.7793	0.7926	0.9338	0.9129
Garden	mIoU	0.2986	0.4507	0.4203	0.5006	0.8499
	mPA	0.8586	0.8863	0.6825	0.7579	0.9577
	mP	0.6504	0.7774	0.7302	0.8227	0.9312
Kitchen	mIoU	0.3788	0.4678	0.4222	0.4995	0.8434
	mPA	0.6837	0.7981	0.7085	0.7517	0.9351
	mP	0.7708	0.6853	0.7152	0.8392	0.9520
Average	mIoU	0.2698	0.3889	0.4410	0.5545	0.8646
	mPA	0.8183	0.8279	0.7586	0.8071	0.9569
	mP	0.6553	0.7085	0.7511	0.8600	0.9362

Table 2: Per-scene and average performance on the Mip-NeRF360 dataset

wherein \mathcal{T} and \mathcal{D} are the TFCC and the MLP decoder, and the subscript j iterates over the elements of the logits e , ascending from 1 up to its length N .

Through this process, we are able to comprehend the 3D Gaussian-level semantic feature. Subsequently, via the Optimizable Semantic-space Hyperplane, we can effectively extract the Gaussians of interest. Consequently, our GOI approach can be harnessed for downstream tasks, enabling efficient 3D manipulations such as deletion, localization, and inpainting.

REFERENCES

- [1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022*. IEEE, 5460–5469. <https://doi.org/10.1109/CVPR52688.2022.00539>
- [2] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* 42, 4 (2023), 1–14.
- [3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.
- [4] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling

Scene	Metric	Works				
		LERF	Feat. 3DGS	GS Grouping	LangSplat	Ours
Room 0	mIoU	0.3095	0.4980	0.5937	0.4843	0.6589
	mPA	0.7761	0.8499	0.8872	0.8134	0.9039
	mP	0.6622	0.7484	0.8241	0.7734	0.8301
Room 1	mIoU	0.3573	0.4244	0.4525	0.5819	0.8020
	mPA	0.7974	0.7826	0.7480	0.8205	0.9383
	mP	0.6810	0.7260	0.7667	0.8694	0.9314
Office 0	mIoU	0.2962	0.5513	0.3388	0.4471	0.5042
	mPA	0.6736	0.8415	0.6664	0.7700	0.7597
	mP	0.7004	0.7786	0.7135	0.7395	0.7384
Office 1	mIoU	0.1630	0.3181	0.2829	0.3682	0.5024
	mPA	0.5812	0.6865	0.6460	0.6736	0.7443
	mP	0.5971	0.6710	0.6060	0.6592	0.7353
Average	mIoU	0.2815	0.4480	0.4170	0.4704	0.6169
	mPA	0.7071	0.7901	0.7369	0.7694	0.8365
	mP	0.6602	0.7310	0.7276	0.7604	0.8088

Table 3: Per-scene and average performance on the Replica dataset

open-world models for diverse visual tasks. *ArXiv preprint abs/2401.14159* (2024). <https://arxiv.org/abs/2401.14159>

[5] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. 2023. Language Embedded 3D Gaussians for Open-Vocabulary Scene Understanding. *ArXiv preprint abs/2311.18482* (2023). <https://arxiv.org/abs/2311.18482>

[6] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. 2019. The Replica dataset: A digital replica of indoor spaces. *ArXiv preprint abs/1906.05797* (2019). <https://arxiv.org/abs/1906.05797>