Supplementary Material for "LangSplat: 3D Language Gaussian Splatting"

Minghan Qin[1,*], Wanhua Li[2,*,✉], Jiawei Zhou[1,*], Haoqian Wang[1,✉], Hanspeter Pfister[2]
[1]Tsinghua University [2]Harvard University
qmh21@mails.tsinghua.edu.cn, wanhua@seas.harvard.edu, zhoujw22@mails.tsinghua.edu.cn
wanghaoqian@tsinghua.edu.cn, pfister@seas.harvard.edu

1. Video Demo

In Figure 1 of our main paper, we have visualized the language features learned by LERF and our method. For a fair comparison, we perform PCA on the decoded feature $\Psi(\boldsymbol{F_t^l}) \in \mathbb{R}^{D\times H\times W}$ for our method. However, one benefit of our method is that we are able to directly visualize the learned language features in the encoded 3-dimensional latent space, which can ensure color consistency between frames.[1] Specifically, we normalize the encoded 3-dimensional latent features $\boldsymbol{H}_t^l(v) \in \mathbb{R}^{3\times H\times W}$ and visualize them by treating the 3-dimensional features as RGB channels.

We strongly recommend readers refer to our video demo at https://www.youtube.com/watch?v=XMlyjsei-Es to observe the learned 3D language fields in the scene-specific latent space. The video demonstrates that our method has acquired a 3D language representation that is both 3D-consistent and distinctly shaped, which significantly distinguishes it from existing methods that often only learn 3D language representations with blurred boundaries. Meanwhile, our approach achieves a speedup of 119 × compared to LERF at a resolution of 988 × 731 and further improves to 199 × faster at a resolution of 1440×1080.

2. More Implementation Details

For each text query, we can obtain three relevancy maps with our trained 3D language Gaussians, each representing one semantic level defined by SAM. Then we use different strategies to choose the best semantic level and obtain the predictions for different tasks.

3D Object Localization on LERF. To mitigate the impact of outliers, we first employ a mean convolution filter with a size of 20 to smooth the values of three relevancy maps. For the smoothed relevancy maps, we select the one with the highest smoothed relevancy score and take the corresponding position as the final prediction.

3D Semantic Segmentation on LERF. Similarly, to mitigate the influence of outliers, we apply a mean filter with a size of 20 to smooth the three relevancy maps. Subsequently, we select the relevancy map with the maximum smoothed relevancy score for binary mask prediction. For the selected relevancy map, we first normalize its relevancy scores and then use a threshold to obtain a binary image as the final prediction mask.

3D Semantic Segmentation on 3D-OVS. For each class query, we obtain three relevancy maps. We apply a threshold of 0.4 to these relevancy maps, setting relevancy scores below 0.4 to 0 and relevancy scores above 0.4 to 1, resulting in three binary maps. We calculate the average relevancy scores within the mask region for each relevancy map and select the relevancy map with the highest average response as the final predicted binary map.

3. More Quantitative Results

In addition to the mIoU metric, the Accuracy metric is also employed on the 3D-OVS dataset in [5].[2] Therefore, we also compare our method with other state-of-the-art methods on the 3D-OVS dataset using the Accuracy metric. The results are shown in Table 1. We observe that our method consistently outperforms other methods, which further illustrates the superiority of our method.

4. More Ablation Study

To reduce the memory cost of our 3D language Gaussians, we proposed the scene-specific autoencoder to learn a latent feature. We show the ablation results of different latent dimensions d on the bench scene of the 3D-OVS dataset in Table 2. We observed that as d increases, the mIoU performance improves, with only a slight increase in the time cost. We chose $d = 3$ because it allows us to directly visualize the learned 3D language field in the latent space by treating

[*] Equal contribution. [✉]Corresponding authors.

[1] The consistency of color in PCA visualizations across different frames is not ensured.

[2] After checking with the authors of 3D-OVS, we confirmed that the mAP results reported in [5] are, in fact, the Accuracy results.

Method	*bed*	*bench*	*room*	*sofa*	*lawn*	overall
LSeg [3]	87.6	42.7	46.1	16.5	77.5	54.1
ODISE [6]	86.5	39.0	59.7	35.4	82.5	60.6
OV-Seg [4]	40.4	89.2	49.1	69.6	92.1	68.1
FFD [2]	86.9	42.8	51.4	9.5	82.6	54.6
LERF [1]	86.9	79.7	79.8	43.8	93.5	76.7
3D-OVS [5]	96.7	96.3	98.9	91.6	97.3	96.2
LangSplat	**99.2**	**98.6**	**99.3**	**97.9**	**99.4**	**98.9**

Table 1. Quantitative comparisons of 3D semantic segmentation on the 3D-OVS dataset. We report the accuracy scores (%).

d	1	2	3	8
mIoU (%)	6.46	91.93	94.19	95.20
Speed (s/q)	0.2770	0.2779	0.2788	0.2807

Table 2. The ablations of latent dimension d for our scene-specific autoencoder. The results are obtained on the bench scene of the 3D-OVS dataset. The image resolution is 1440×1080.

the 3-dimensional features as the RGB channels. We also strongly encourage readers to refer to our video demo to observe how our learned language field accurately captures the precise 3D shape of objects in the scene-specific latent space.

5. More Visualization Results

3D Object Localization on LERF. We visualize more examples on the LERF dataset for open-vocabulary 3D object localization in Figure 1. We found that for text queries such as “red apple” and “plate”, LERF failed to correctly locate the 3D positions, whereas our method succeeded. For text queries like “waldo” and “chopsticks”, although LERF could identify the correct location, its activation values were more dispersed, whereas our method was able to focus more precisely on the queried object.

3D Semantic Segmentation on LERF. We demonstrate more examples on the LERF dataset for open-vocabulary 3D semantic segmentation in Figure 2. We observed that the results produced by LERF were unable to provide the precise shape of the queried object and exhibited a significant amount of noise, whereas our method could accurately depict the object’s shape. These results show the effectiveness of our proposed LangSplat.

3D Semantic Segmentation on 3D-OVS. We show more scenes on the 3D-OVS dataset for open-vocabulary 3D semantic segmentation in Figures 3, 4, 5, and 6, respectively. Compared to the previous state-of-the-art method 3D-OVS, our approach provides more precise object boundaries and exhibits reduced noise, which illustrates that our LangSplat learns a more accurate 3D language field.

References

[1] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *ICCV*, pages 19729–19739, 2023. 2

[2] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *NeurIPS*, 35:23311–23330, 2022. 2

[3] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 2

[4] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, pages 7061–7070, 2023. 2

[5] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly supervised 3d open-vocabulary segmentation. In *NeurIPS*, 2023. 1, 2

[6] 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 *CVPR*, pages 2955–2966, 2023. 2

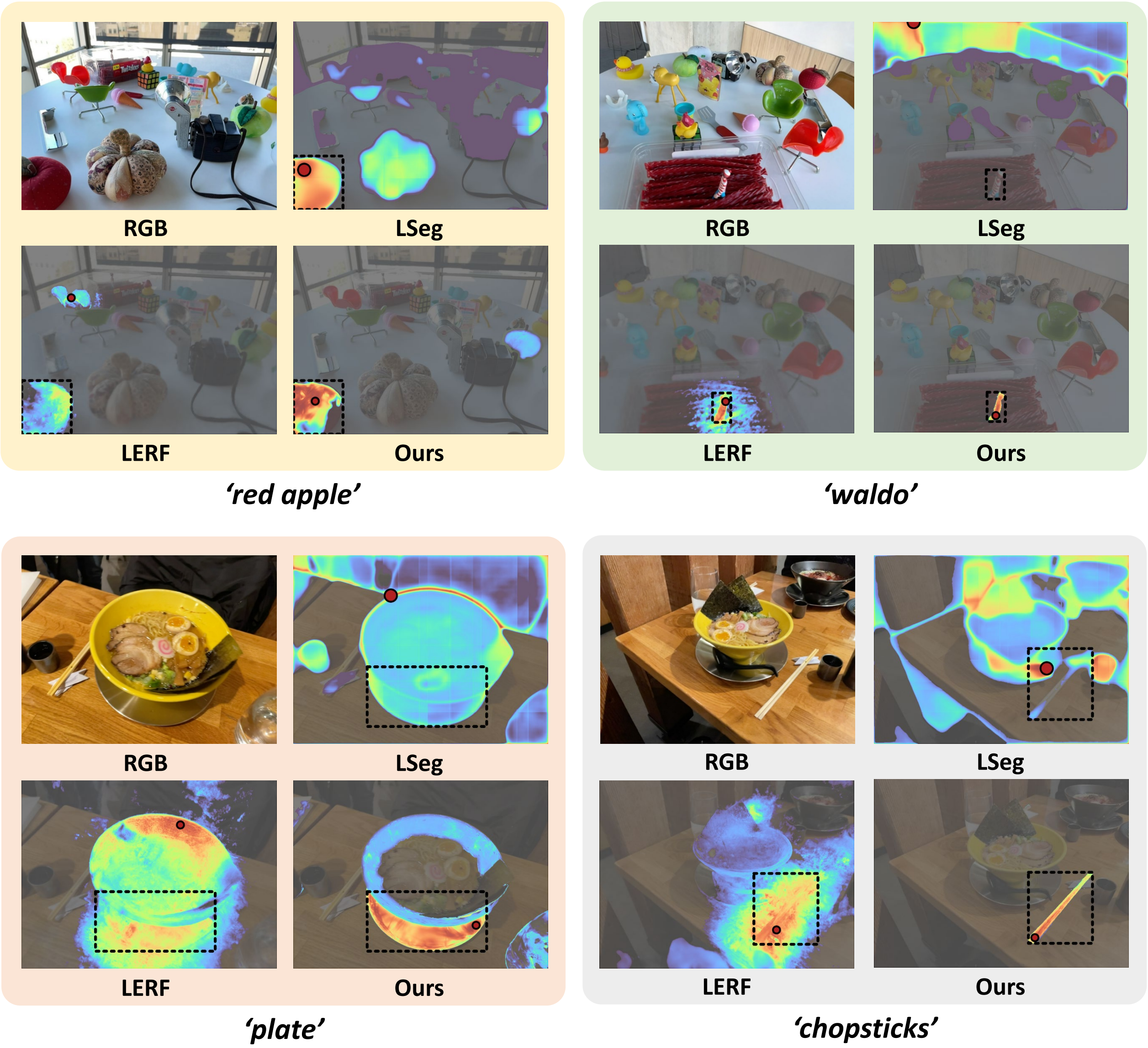

Figure 1. More qualitative comparisons of open-vocabulary 3D object localization on the LERF dataset. The red points are the model predictions and the black dashed bounding boxes denote the annotations.

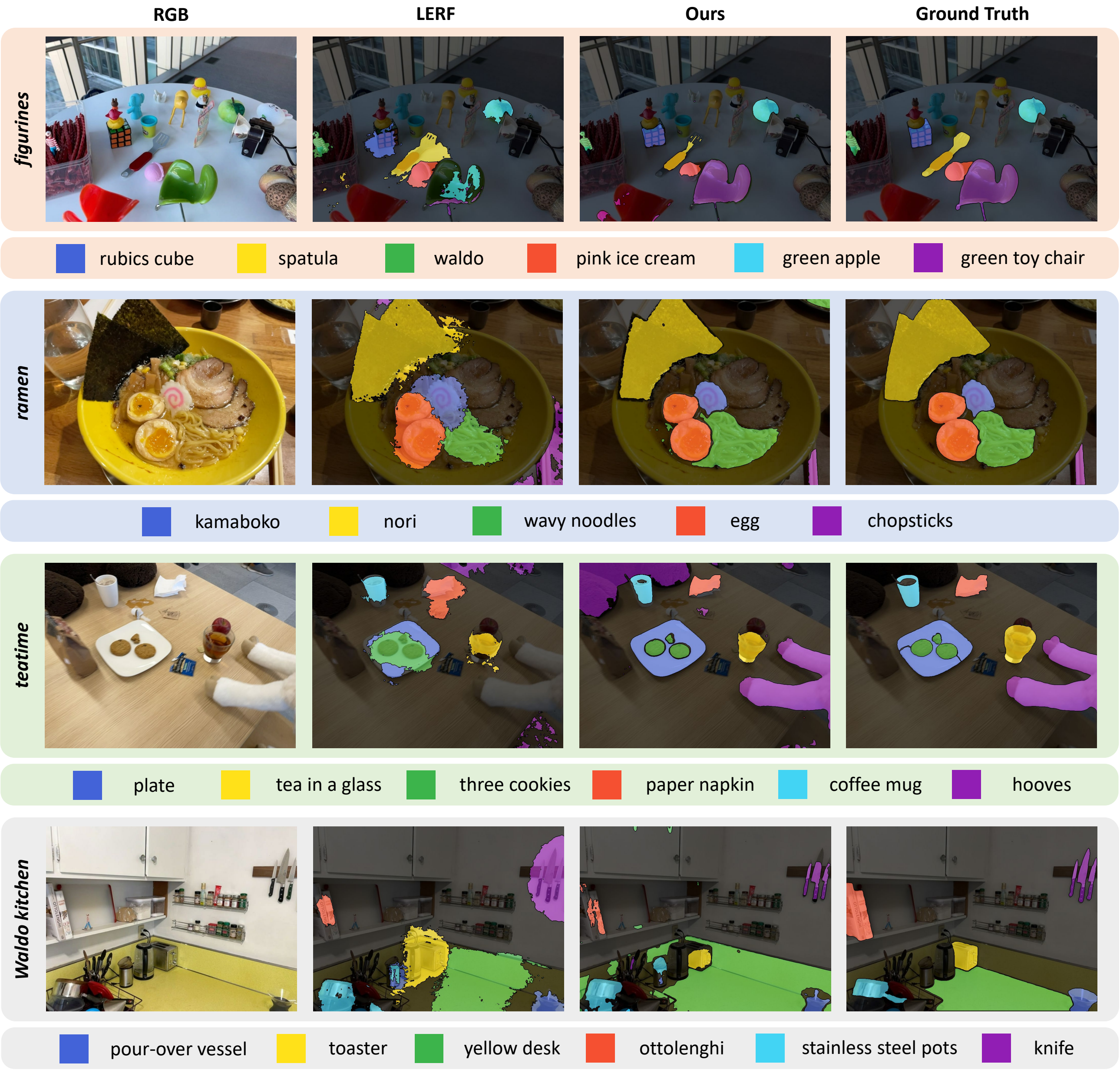

Figure 2. More qualitative comparisons of open-vocabulary 3D semantic segmentation on the LERF dataset.

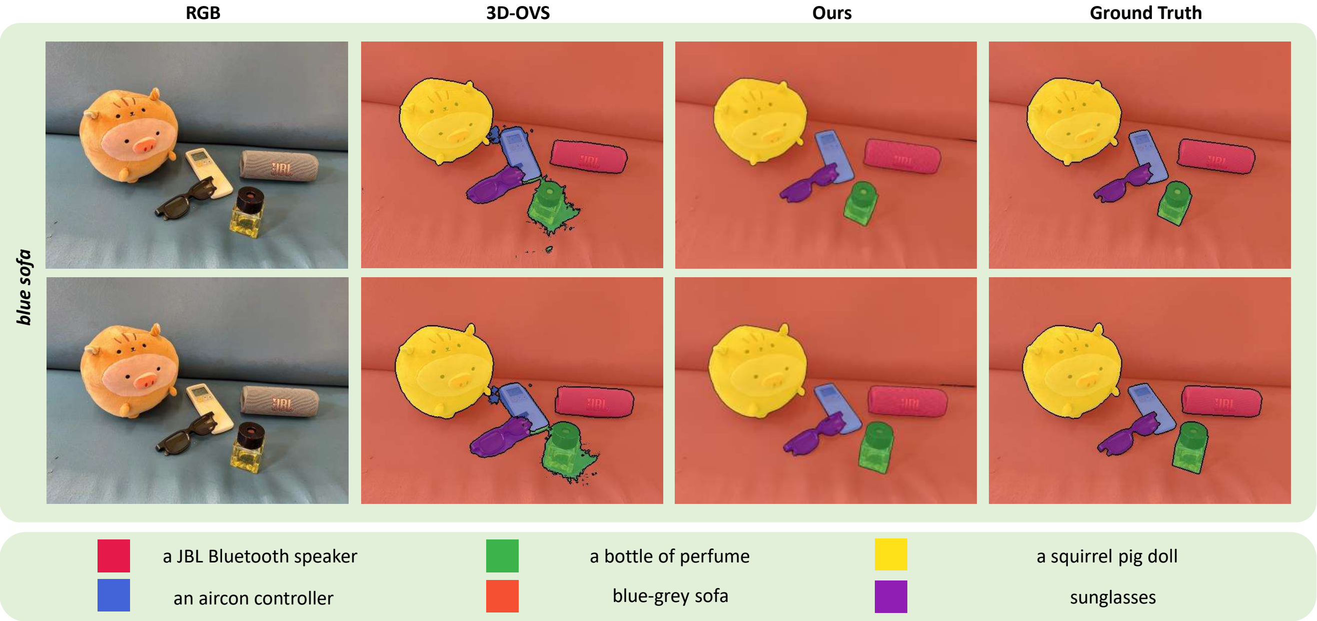

Figure 3. Qualitative comparisons on the blue sofa scene of the 3D-OVS dataset.

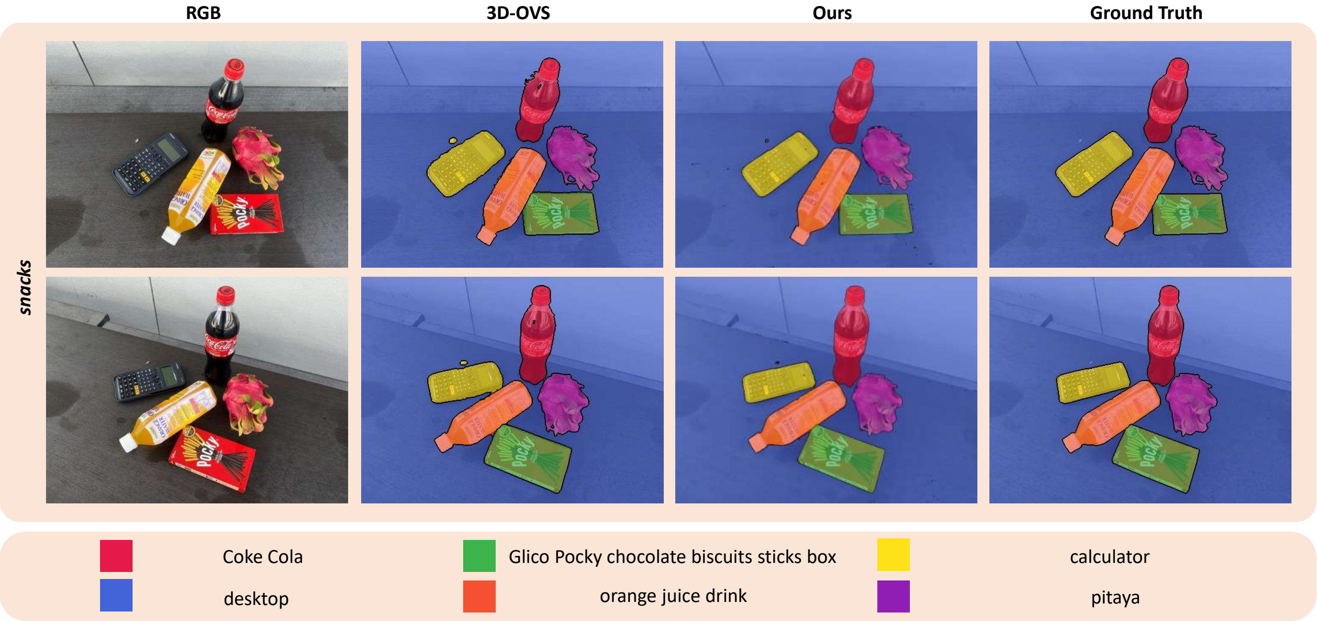

Figure 4. Qualitative comparisons on the snacks scene of the 3D-OVS dataset.

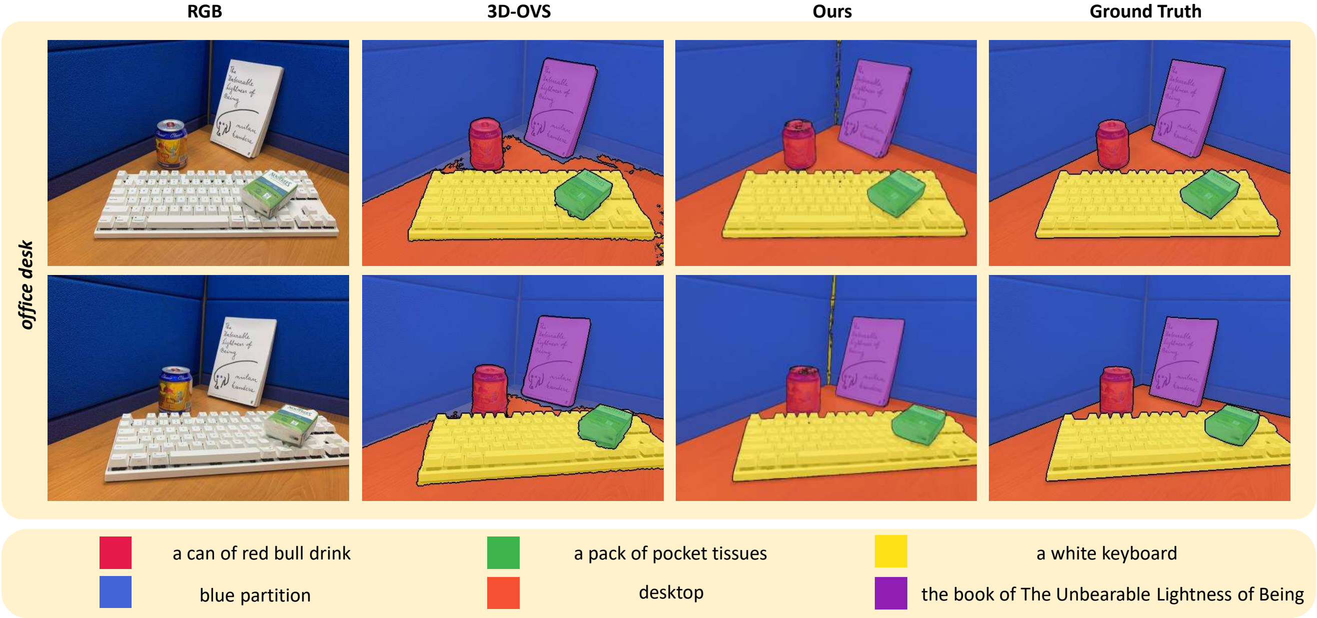
RGB
3D-OVS
Ours
Ground Truth
office desk
🟥 a can of red bull drink
🟩 a pack of pocket tissues
🟨 a white keyboard
🟦 blue partition
🟧 desktop
🟪 the book of The Unbearable Lightness of Being

Figure 5. Qualitative comparisons on the office desk scene of the 3D-OVS dataset.

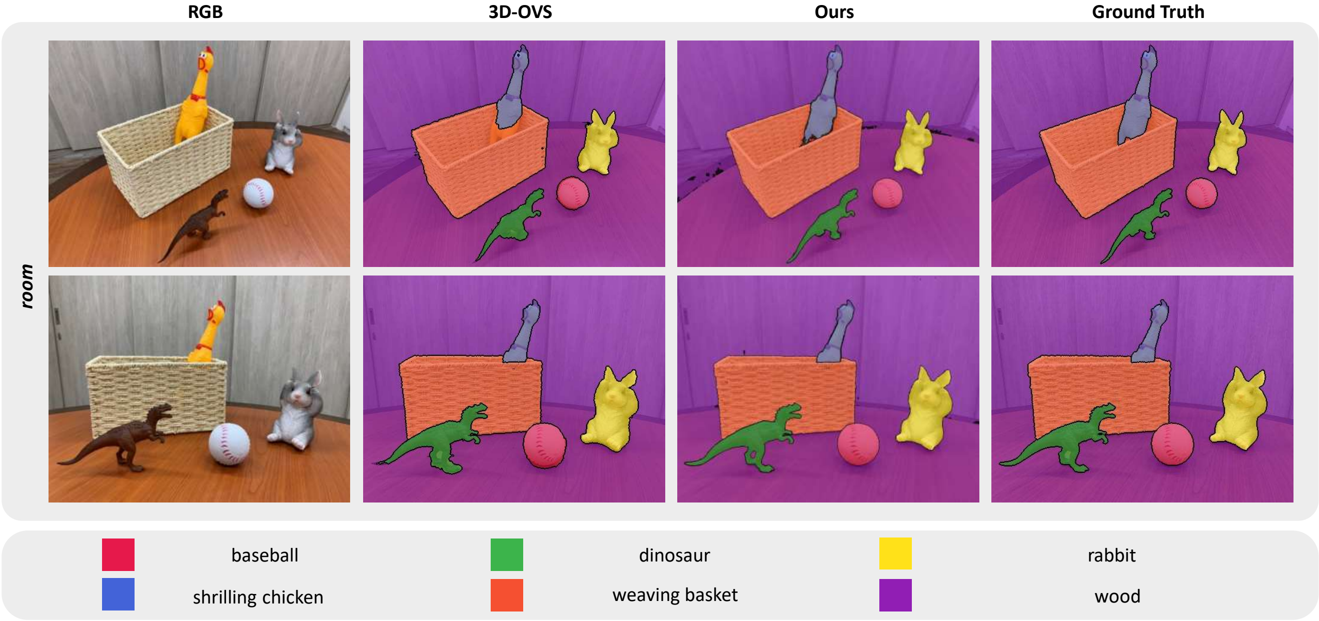

Figure 6. Qualitative comparisons on the room scene of the 3D-OVS dataset..