⚡LightSplat: Fast and Memory-Efficient Open-Vocabulary 3D Scene Understanding in Five Seconds Supplementary Materials

Jaehun Bang[1] Jinhyeok Kim[2*] Minji Kim[1] Seungheon Jeong[1] Kyungdon Joo[1†]
[1] UNIST [2] POSTECH
{devappendcbangj, mzkim, sypsss, kyungdon}@unist.ac.kr jh4011@postech.ac.kr

Overview

In these supplementary materials, we provide additional details on the following topics:

- Sec. A introduces the DL3DV-OVS dataset, a large-scale open-vocabulary benchmark covering four indoor-outdoor scenes with diverse object queries.
- Sec. B outlines our experimental setup, including 3DGS, feature extraction, and hyperparameters.
- Sec. C summarizes feature distillation time and reports millisecond-level inference from lightweight clustering.
- Sec. D presents additional qualitative comparisons that demonstrate robustness in diverse 3D scenes.
- Sec. E showcases text-driven 3D editing enabled by cluster-level semantic control.
- Sec. F discusses limitations regarding object feature selection and the precision-speed trade-off.

A. DL3DV-OVS Dataset

We propose DL3DV-OVS, an open-vocabulary extension of DL3DV for evaluating model robustness in large-scale and challenging scenes. The dataset spans indoor and outdoor environments with broad spatial layouts and diverse scene structures. It provides 22 text queries for objects such as cars, chairs, monitors, and streetlights. These items vary significantly in size and distance, often appearing multiple times within a single view. DL3DV-OVS consists of four scenes: park, road, office, and shop. For each scene, we provide 960×540 multi-view images, text queries, and the corresponding ground-truth 2D masks. Table S1 lists the text queries for each scene.

Table S1. **Scene-wise text queries in the DL3DV-OVS dataset.** Each scene provides a set of text queries for evaluating open-vocabulary 3D object selection.

Scene	Text Queries		
Park	bench	trash bin	car
	parking signpost	information board	
Shop	black drawer	framed artwork	lamp
	potted plant	white sofa	bed
	fire extinguisher	round wooden coffee table	
Road	bench	car	fire hydrant
	planter	roof	trash bin
	streetlight		
Office	chair	keyboard	monitor
	whiteboard	mouse	

B. More Implementation Details

B.1. Experimental Setup

Training Setup. For consistent comparison of speed and accuracy, all models are evaluated on a single RTX 4090 GPU. LUDVIG is the only exception, running on an A6000 because its high-dimensional language feature operations exceed the 24GB memory limit.

3D Gaussian Splatting. To ensure a fair comparison, we use a pre-trained 3DGS trained for 30,000 iterations, identical to LangSplat and OpenGaussian. We do not modify any Gaussian parameters and simply allocate an additional 2-byte space per Gaussian to store an index. A full language feature requires 2028 bytes per Gaussian, so storing it for 100,000 Gaussians would require about 200 MB, which is much larger than the 23.6 MB needed for all standard Gaussian parameters. In our method, we no longer store this 2024-byte feature for each Gaussian. Instead, each Gaussian stores only a small index, with the total index size being just 0.2 MB, which is less than 0.1% of the 200 MB required for per-Gaussian language features.

Object Feature Extraction. LangSplat uses all three hierarchical mask levels produced by SAM and learns separate language features for each level. To ensure both fair comparison and computational efficiency, we follow OpenGaussian and adopt a large-mask strategy. We extract mask-level

CLIP features by cropping each SAM mask, resizing it to 224×224 with zero padding, and encoding it using OpenCLIP [2]. With this setup, our method integrates semantics and geometry without relying on such hierarchical structures, achieving SOTA performance with $50\times$ faster speed.

B.2. Hyperparameters

To ensure fair evaluation, we use the same hyperparameters for all scenes within each dataset. Denser views or more compact scenes produce higher mask overlap and thus require stricter IoU thresholds. In contrast, datasets with more diverse text queries require slightly relaxed feature similarity thresholds. Accordingly, for LERF-OVS, ScanNet, and DL3DV-OVS, we set (contrib, noise, IoU, feat) to (0.04, 200, 0.6, 0.75), (0.04, 500, 0.35, 0.8), and (0.09, 450, 0.5, 0.8), respectively. Unlike prior methods that tune hyperparameters per scene (e.g., teatime) [7], our method uses fixed settings and still performs consistently well across diverse scenes, including the challenging ramen scene. Across all three datasets, our method achieves a 400-720$\times$ speedup in feature distillation and improves mIoU over OpenGaussian by 5.43, 7.68, and 19.65 points, respectively. These results demonstrate strong robustness across diverse environments.

B.3. Evaluation Setup

Model Comparison. Since Dr.Splat provides only training code, we adopt the reported inference metrics from its paper and measure other results ourselves. Dr.Splat is expected to have higher inference time, as it processes 128-dimensional features through a codebook and compares them against all Gaussians. In contrast, our method performs semantic comparison at the 3D cluster level instead of evaluating every individual Gaussian. By attaching semantics to clusters through index-feature mapping without relying on a codebook, the comparison set remains compact and efficient. As shown in Table 3, this enables our method to achieve a 2ms average inference time across ScanNet scenes.

3D Object Selection. This section provides a detailed explanation of the 3D Object Selection task. Previous works, such as LangSplat and LEGaussians, struggle to accurately distinguish objects in 3D space. This is because these methods distill features after rendering 3D Gaussians into 2D, where overlapping Gaussians blur the features and result in indirect, less effective supervision. As a result, high-quality features extracted in 2D often do not transfer well to 3D, and 2D-based evaluation provides only a partial view of the model's 3D understanding. To address this, we adopt the 3D Object Selection task, which directly selects objects in 3D from text queries. We evaluate on LERF-OVS by first selecting objects directly in 3D, then rendering them back to 2D for comparison with ground-truth masks, following OpenGaussian and Dr.Splat.

3D Semantic Segmentation. We evaluate 3D semantic segmentation on the ScanNet benchmark, following the evaluation protocol used in OpenGaussian. The protocol fits 3D Gaussians directly to the ScanNet GT points, matching their locations and counts. This establishes a one-to-one correspondence between Gaussians and GT points, enabling direct comparison of predicted and GT labels. For fair comparison, all baselines follow the same setting.

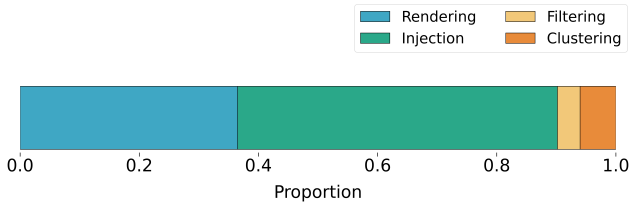

Figure S1. **Module-wise feature distillation time.** This figure shows the time proportion of each module during feature distillation. Injection, filtering, and clustering refer to indexed feature injection, 3D-aware mask filtering, and context-aware 3D clustering, respectively.

Table S2. **Inference time comparison on LERF-OVS and DL3DV-OVS.** We report inference in seconds. While prior methods rely on slow 2D or per-Gaussian processing, `LightSplat`'s cluster-level inference achieves millisecond runtime.

Methods	LERF-OVS	DL3DV-OVS
LangSplat	1.046	0.175
LEGaussians	0.383	0.770
LUDVIG	5.923	1.865
OpenGaussian	0.003	0.005
Ours	0.001	0.003

C. More Quantitative Results

Module FD Time. Fig. S1 shows the average time consumed by each module in our feature distillation pipeline across the LERF-OVS scenes. Most of the computational cost arises from rendering and indexed feature injection, since both steps require computation for each view. In contrast, the subsequent 3D-aware mask filtering and context-aware clustering operate in a single step, adding almost no additional cost or delay. This design keeps the overall distillation process extremely fast, and the ablation study shows that these components improve performance without slowing the pipeline. Across scenes, rendering, injection, filtering, and clustering account for 36.5%, 53.7%, 3.8%, and 6.0% of the total time, respectively.

Inference Time. Table S2 presents inference-time results on LERF-OVS and DL3DV-OVS, providing evaluations beyond the ScanNet results in the main paper. Rendering-based methods such as LangSplat and LEGaussians incur significant computational overhead because their inference

pipeline requires additional decoding of semantic features. LangSplat performs per-Gaussian latent decoding through its autoencoder, while LEGaussians applies per-pixel semantic decoding using its MLP, both further increasing latency. OpenGaussian is also slow because it computes high-dimensional features for every Gaussian. LUDVIG introduces additional overhead by diffusing 2048-dimensional features over a graph. In contrast, `LightSplat` performs cluster-level inference through an index-feature mapping, avoiding per-Gaussian comparisons and diffusion. As a result, `LightSplat` maintains millisecond-level inference even in large indoor-outdoor scenes.

D. More Qualitative Results

We present additional qualitative results omitted from the main paper due to space constraints. These results highlight the robustness of our model across a wide range of scenes.

Clustering Process. Fig. S2 illustrates how `LightSplat` converts 2D masks into stable 3D object clusters. We use the same hyperparameters for the two DL3DV-OVS scenes shown above and the two additional indoor-outdoor scenes below, demonstrating robustness without scene-specific tuning. The process begins with SAM masks extracted from the input images. These masks are injected into the 3D scene based on per-pixel contribution, producing a mask ID field where each Gaussian receives the index of its most influential 2D mask. This intermediate field may appear noisy due to view-dependent inconsistencies. We address this by filtering out unstable masks in the 3D-aware mask filtering stage using 2D-3D correspondences. Finally, context-aware 3D clustering groups semantically and spatially consistent masks into clean object-level clusters. The resulting cluster ID field demonstrates robust and coherent semantics across diverse indoor and outdoor environments.

3D Object Selection. Fig. S3 and Fig. S4 present additional qualitative results on the LERF-OVS and DL3DV-OVS datasets. Unlike models where iterative aggregation blurs semantics in 3D [1, 3–6], our method transfers mask-level semantics directly to 3D. It then merges related masks into object-level clusters, preserving clearer object structure than Gaussian-level aggregation. As a result, our method achieves high accuracy in challenging real-world cases:

- small objects attached to others (kamaboko, wavy noodle)
- multiple instances of the same object (knives, cars)
- semantically subtle categories (yellow pouf)
- thin or complex structures (jake with thin legs, streetlight)
- single outdoor objects (trash bin, information board)

3D Semantic Segmentation. Fig. S5 presents more qualitative results on the ScanNet dataset. These results show that our approach captures a broader range of semantics in indoor scenes, while also producing more complete object coverage with well-aligned boundaries.

These additional qualitative results further highlight the robustness of our model to diverse and challenging queries, from small objects to complex structures, while outperforming other models in both accuracy and speed on LERF-OVS, ScanNet, and DL3DV-OVS.

E. 3D Scene Editing

Fig. S6 shows how our method naturally extends to 3D scene editing. By managing semantics at the cluster level, our method enables fast and accurate segmentation of text-specified objects in 3D space. Building on this segmentation capability, users can directly remove, recolor, or reposition the identified objects, enabling high-fidelity scene edits. We demonstrate this capability through recoloring and enlarging as representative examples. Enlargement is performed by adjusting Gaussian distances and scales, while recoloring is achieved by modifying SH coefficients. Notably, these editing operations add minimal overhead, keeping inference time nearly unchanged in interactive scenarios. Together, these results highlight the speed, accuracy, and flexibility of our approach for interactive 3D manipulation in immersive content creation, AR/VR, and robotics.

F. Limitations

Object Feature Selection. SAM provides high-quality object segmentation masks, and CLIP enables effective image-text interaction as a foundation model trained on large-scale datasets in a shared embedding space. However, SAM masks do not always produce perfectly accurate boundaries. In addition, CLIP can struggle to distinguish semantically similar objects (e.g., red apple vs. green apple) due to their high feature similarity. These limitations could be further mitigated with extra modules or iterative training, but this would disrupt the speed–accuracy balance of our approach.

Hyperparameters. Our method performs semantic injection and 3D clustering in a single step using only four main hyperparameters, making it lightweight and intuitive. Compared to training-based methods, which involve numerous hyperparameters for model initialization and optimization, our approach requires far fewer parameters, avoiding the high training cost and memory usage. Despite its simplicity, our method can still be influenced by the choice of hyperparameters and can be further refined for improvement.

References

[1] Yash Bhalgat, Iro Laina, João F Henriques, Andrew Zisserman, and Andrea Vedaldi. N2f2: Hierarchical scene understanding with nested neural feature fields. In *European Conference on Computer Vision*, 2024. 3

[2] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scal-

ing laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[3] Kim Jun-Seong, GeonU Kim, Kim Yu-Ji, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. Dr. splat: Directly referring 3d gaussian splatting via direct language embedding registration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 3

[4] Juliette Marrie, Romain Ménégaux, Michael Arbel, Diane Larlus, and Julien Mairal. Ludvig: Learning-free uplifting of 2d visual features to gaussian splatting scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[5] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[6] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[7] Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, et al. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *Advances in Neural Information Processing Systems*, 37:19114–19138, 2024. 2

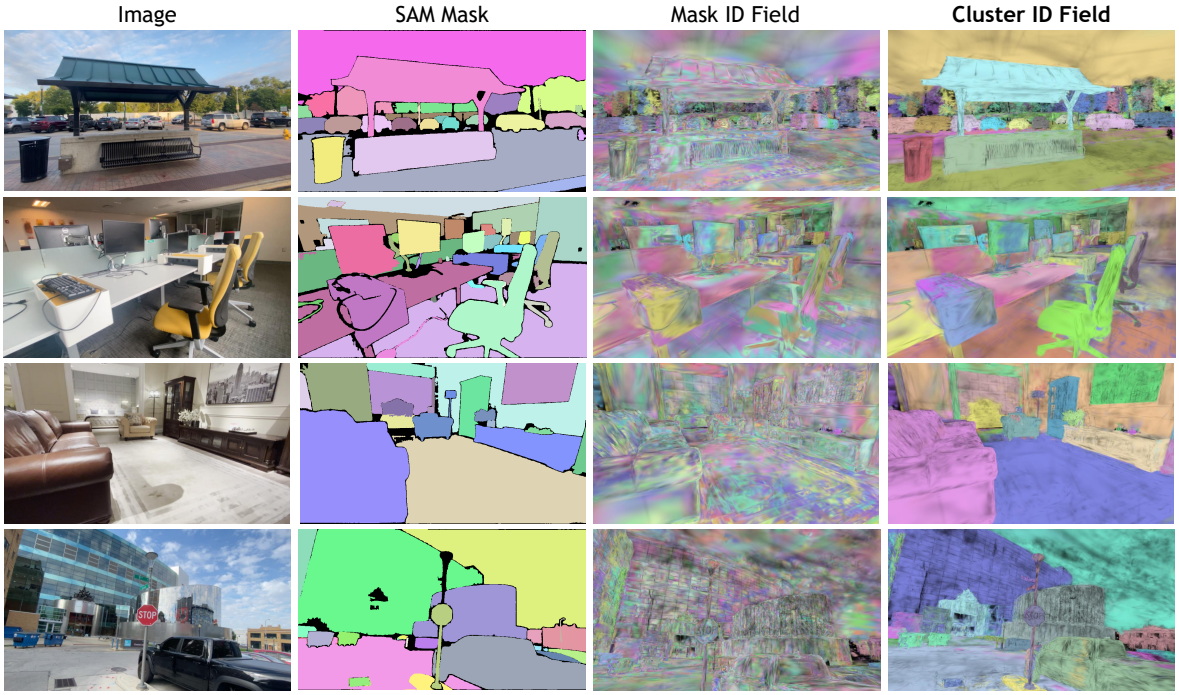

Figure S2. **Clustering Process Visualization.** The figure shows the SAM mask IDs lifted into 3D and refined into coherent object-level clusters, resulting in the final cluster ID field. It also demonstrates that our method produces clean cluster ID fields across diverse indoor and outdoor scenes without scene-specific tuning.

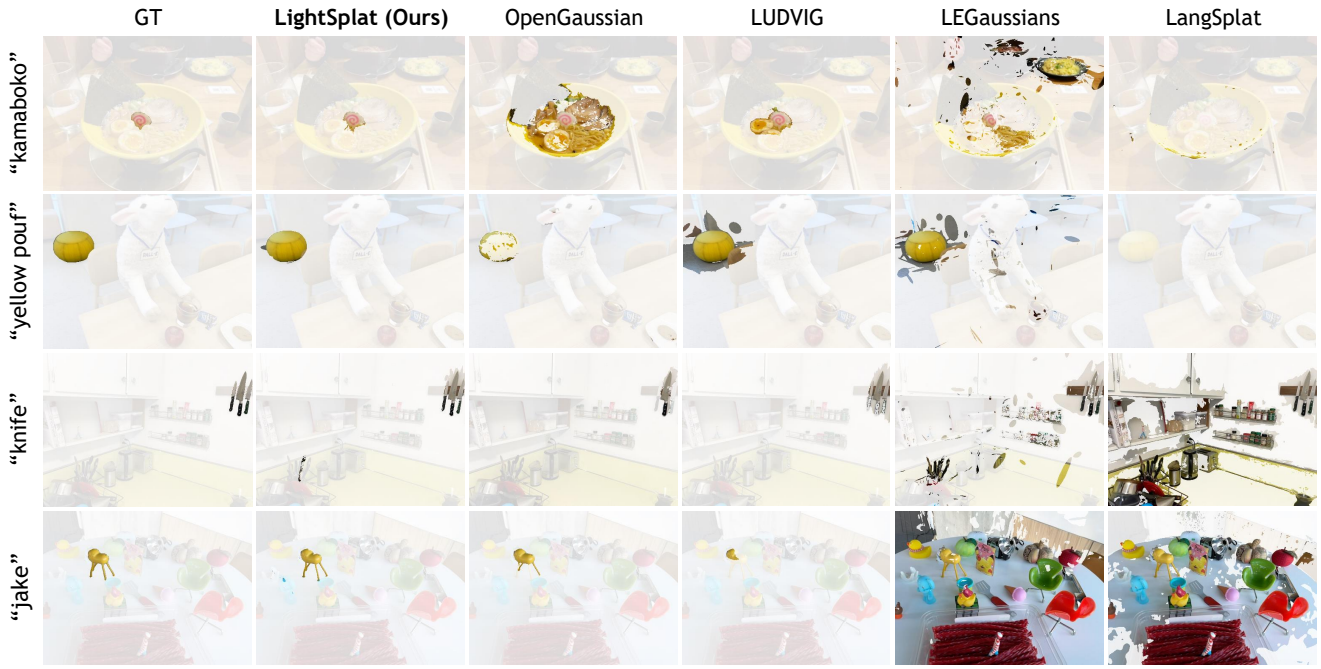

Figure S3. **More qualitative comparison for 3D Object Selection on the LERF-OVS dataset.** We visualize performance across different scenes and text queries in LERF-OVS. With context-aware 3D clustering, our method delivers precise boundaries for challenging queries, including small objects in contact with others, repeated instances, and thin or intricate structures, while achieving the fastest performance.

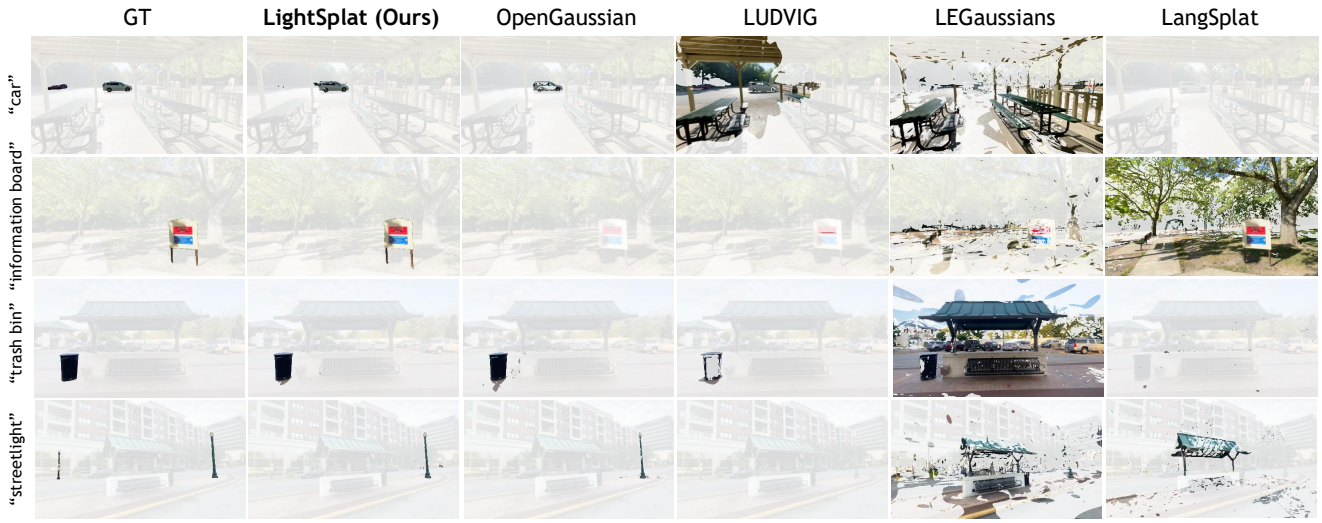

Figure S4. **More qualitative comparison for 3D Object Selection on the DL3DV-OVS dataset.** We visualize model performance across different scenes and text queries in DL3DV-OVS. With context-aware 3D clustering, our method yields accurate boundaries in large indoor-outdoor scenes, handling distant objects, multiple instances, and complex structures while maintaining the fastest performance.

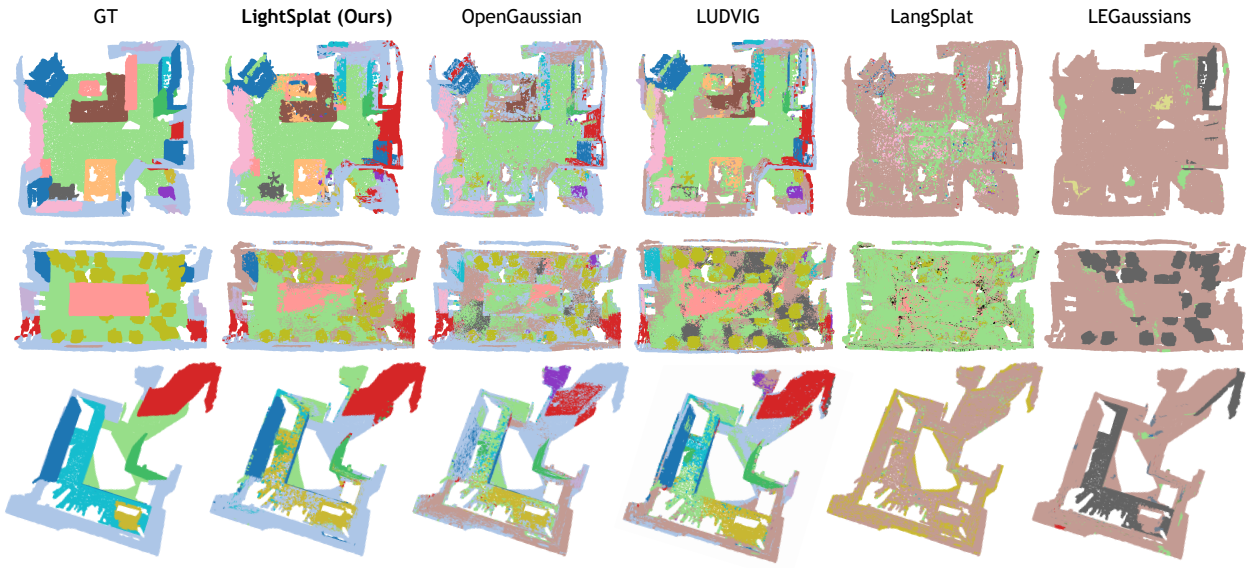

Figure S5. **More qualitative comparison for 3D Semantic Segmentation on the ScanNet dataset.** For intuitive comparison, ground-truth semantics are visualized in distinct colors. Our method captures a broader range of semantics in indoor scenes, providing more complete coverage with precise boundaries. It handles both object-level (e.g., door, cabinet) and large-area semantics (e.g., floor, wall), showing robustness across diverse scenes and outperforming other methods in accuracy and speed.

Figure S6. **Qualitative results of text-driven 3D scene editing on the LERF-OVS dataset.** Recoloring (middle) and enlarging (bottom) are representative examples, enabled by our 3DGS-based approach for fast, accurate, and high-fidelity object manipulation across diverse scenes and conditions.