vgrep3d: Open-Vocabulary 3D Object Querying in Gaussian Splatting Scenes via SAM-Guided SigLIP 2 Feature Fields

Ryan Abraham
Lassonde School of Engineering, York University
ryanabraham@my.yorku.ca

Supervised by
Prof. James Elder
York University
Elder Labatory

Abstract

This work is submitted as an undergraduate thesis at York University·s Lassonde School of Engineering, supervised by Prof. James Elder (York University, Elder Labatory).

The problem. Robots and autonomous systems need to locate objects in 3D space from natural language descriptions — "find the red car," "where is the stop sign?" This is an open research challenge at the frontier of robotics and machine learning. Existing methods work well on clean, controlled indoor scenes but have not been rigorously tested on real-world outdoor data, where forward-facing cameras, large scene extents, and limited viewpoint coverage create fundamentally harder conditions.

The approach. The dominant paradigm — language feature field distillation — trains a 3D scene representation to embed vision-language features at every point in space, enabling text-based querying at test time. We implement this paradigm using SigLIP 2, a state-of-the-art vision-language model, combined with SAM-guided per-mask feature extraction and Gaussian Splatting as the 3D backbone. A compact per-scene autoencoder (1152→16 dimensions, 72× compression) reduces memory from 3.15 GB to 43.8 MB for 683,644 Gaussians.

What we did. We built and deployed the full pipeline end-to-end on a real-world dashcam driving scene (292 training views, 683,644 Gaussians) — a setting no prior language-field paper has evaluated on. We ran 30 epochs of feature field distillation (5h 15m 36s on A10G GPU) and queried the trained field with six semantic prompts.

What we found. The pipeline correctly encodes scene-level semantic structure: "sky" activates only 7.6% of Gaussians with a geometrically distinct centroid, while "road" activates 28.2% consistent with ground-plane geometry. However, tight per-object 3D bounding boxes are unachievable — "red car" activates 63.9% of all Gaussians. We provide evidence-based diagnoses of both root causes: limited view diversity from forward-facing dashcam capture, and SigLIP 2 crop-level embedding flatness in driving scenes. All code and trained artifacts are open-sourced at https://github.com/ryamabra/vgrep3d.

1 Introduction

1.1 The Problem

Imagine a delivery robot navigating a busy street. It needs to answer: "where is the red car I must avoid?" or "which door is the entrance?" — not by recognizing objects in a single 2D camera frame, but by localizing them in the 3D world it operates in. This capability, *open-vocabulary 3D object grounding*, sits at the frontier of robotics and machine learning research, and is actively studied at groups including Berkeley BAIR, NVIDIA Research, CMU Robotics, and the Vector Institute.

The challenge has two layers. First, 3D scene representations must be rich enough to support spatial reasoning about arbitrary objects. Second, language must be grounded into that 3D space — not as bounding-box labels on 2D images, but as a continuous field of meaning that a robot can query from any viewpoint and at any scale.

1.2 How the Field Approaches It

The key insight, pioneered by Kerr et al. at UC Berkeley BAIR as **LERF** [Kerr et al., 2023], is elegant: train a 3D scene representation — in their case a NeRF — to embed CLIP language features at every 3D point alongside color. After training, any text query produces a 3D relevancy heatmap: "wooden spoon," "waldo," "electricity," "yellow." LERF's critical contribution was *multi-scale supervision* — the same 3D point receives CLIP features from image crops of different sizes, teaching the field both coarse ("utensils") and fine-grained ("wooden spoon") concepts simultaneously. Justin Kerr and Chung Min Kim (PhD students supervised by Ken Goldberg and Angjoo Kanazawa at UC Berkeley) built LERF and GARField; our pipeline is a direct descendant of their work.

LangSplat [Qin et al., 2024] transplanted this paradigm from NeRF into 3D Gaussian Splatting [Kerbl et al., 2023], gaining real-time rendering and a simpler optimization landscape. OpenGaussian [Wu et al., 2024] added 3D-consistent instance grouping. CAGS [Sun et al., 2025] and OpenInsGaussian [Huang et al., 2025] address structural failure modes in the SAM per-mask embedding step. However, every published system benchmarks exclusively on clean, object-centric indoor captures with >180° viewpoint coverage — controlled conditions that do not reflect robotic deployment.

1.3 How We Tackled It

This undergraduate thesis builds and evaluates vgrep3d, an end-to-end open-vocabulary 3D querying pipeline, on a real-world dashcam driving scene — a setting no prior language-field paper has evaluated on. Our approach:

(1) SigLIP 2 as the backbone. We substitute the standard CLIP backbone with SigLIP 2 (ViT-SO400M-14-SigLIP-384, 1152-

dim) [Tschannen et al., 2025], which achieves stronger dense visual understanding via a sigmoid-based training objective. To our knowledge this is the first evaluation of SigLIP 2 in a Gaussian feature-field context.

(2) Full end-to-end deployment on outdoor dashcam data. We run COLMAP → gsplat → SAM feature extraction → autoencoder → 30-epoch field distillation → 6-prompt query evaluation, all on a single real-world forward-facing driving scene (292 views, 683,644 Gaussians, 900×1600 resolution).

(3) Honest failure analysis from first principles. Rather than reporting only successes, we inspect the actual pipeline artifacts — cached `.npz` feature files, per-epoch loss logs, query centroid coordinates — to diagnose precisely where and why the method falls short.

1.4 What We Found

The pipeline successfully encodes scene-level semantic geometry: sky and road are geometrically separable from object-like queries, and centroid positions land in the correct spatial hemisphere. The 72× autoencoder compression (3.15 GB → 43.8 MB) works well in practice.

However, tight per-object 3D bounding boxes are not achievable for this scene, and we identify two root causes with direct evidence: (1) *view diversity* — forward-facing dashcam capture provides insufficient angular variation for per-Gaussian separation; and (2) *embedding flatness* — SigLIP 2's image tower produces compact-clustered embeddings for driving-scene crop fragments, even with SAM per-mask extraction fully implemented (verified: 92 distinct masks per frame). Both root causes are recognized open problems in the research community, and our results provide a concrete empirical data point for the field.

All code, trained artifacts, and figures are open-sourced at `https://github.com/ryamabra/vgrep3d`.

2 Related Work

Language-augmented Gaussian fields. LangSplat [Qin et al., 2024] established the dominant recipe: per-scene autoencoder compression, per-Gaussian latent optimization via differentiable rasterization, and cosine-similarity retrieval at query time. LEGaussians [Shi et al., 2024] combined CLIP and DINO with adaptive spatial smoothing. OpenGaussian [Wu et al., 2024] added 3D-consistent SAM-guided instance grouping and a two-stage codebook for coarse-to-fine feature discretization. Semantic Gaussians [Guo et al., 2024] distilled features from multiple 2D encoders. CAGS [Sun et al., 2025] and OpenInsGaussian [Huang et al., 2025] both identify residual limitations of per-crop SAM embedding and propose architectural remedies.

Outdoor and large-scale scenes. Street Gaussians [Yan et al., 2024] and DrivingGaussian [Zhou et al., 2024] address reconstruction for driving scenes but attach no semantic feature fields. OutLangSplat [Anonymous, 2025] is the most recent work on open-vocabulary Gaussian querying in outdoor (UAV) settings.

Vision-language models. CLIP [Radford et al., 2021] remains the dominant backbone. SigLIP 2 [Tschannen et al., 2025] achieves improved zero-shot classification via a sigmoid-based loss. DINOv2 [Oquab et al., 2024] provides self-supervised patch-level features with strong spatial properties. Our work is the first to evaluate SigLIP 2 in a Gaussian feature field.

3 Method

Our pipeline has five stages. Figure 1 gives an overview.

3.1 Stage 1: Scene Reconstruction

We reconstruct camera poses with COLMAP [Schönberger and Frahm, 2016] using sequential matching, and initialize gsplat v1.5.3 [Ye et al., 2024] for 3D Gaussian training (7,000 iterations, NVIDIA A10G, 24 GB VRAM via Modal serverless). Result: N=683,644 Gaussians, SH degree 3, V=292 training views at 900×1600.

3.2 Stage 2: SAM-Guided SigLIP 2 Feature Extraction

For each training frame, SAM ViT-H auto-mask generator produces class-agnostic masks (min area 400 px, $M \approx 90$/frame). Each crop is embedded by SigLIP 2 (1152-dim, float16) and stored as a compact $(\mathbf{I}, \mathbf{T})$ pair: a per-pixel mask-id map $\mathbf{I} \in \mathbb{Z}^{H \times W}$ and embedding table $\mathbf{T} \in \mathbb{R}^{M \times 1152}$. Frame 0150: 92 masks, 242.9 KiB. Total cache: 69 MB (293 frames).

3.3 Stage 3: Per-Scene Autoencoder

Architecture: $1152 \to 256 \to 128 \to 16 \to 128 \to 256 \to 1152$ (GELU, 661,392 parameters). Objective: cosine loss. Training: 200 epochs, Adam (lr=10^{-3}), batch 4096. Compression: 3.15 GB → 43.8 MB (**72**×).

3.4 Stage 4: Feature Field Distillation

Frozen geometry + trainable per-Gaussian 16-dim latents $\mathbf{z}_i$. gsplat rasterizes latents as arbitrary-dim features. Loss:

$$\mathcal{L} = \frac{1}{|P|} \sum_{v \in P} \frac{1}{|\text{valid}_v|} \sum_{p \in \text{valid}_v} \Big(1 - \cos\big(\hat{\mathbf{F}}_v[p],\ \text{AE.encode}(\mathbf{F}_v[p])\big)\Big)$$

where P is a random 60-view subset sampled each epoch. Training: 30 epochs, Adam (lr=10^{-2}). Total wall time: **5h 15m 36s** on A10G.

3.5 Stage 5: Query

Text prompt → SigLIP 2 text tower → unit vector $\mathbf{t}$. Per-Gaussian relevance: robust percentile normalization (1st/99th clamp of cosine similarity). MAD filter (n_σ=3, 2 iters). AABB from 2nd–98th percentile extents. Latency: 36–42 s per prompt on A10G (including model load).

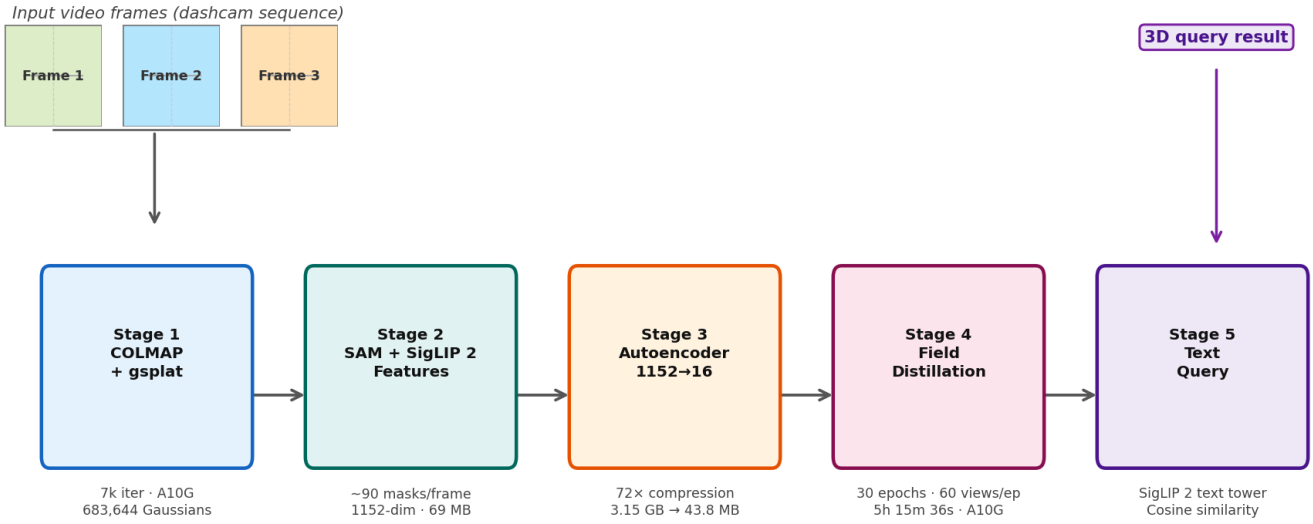


Figure 3. vgrep3d pipeline. Dashcam video frames enter Stage 1 (reconstruction), progress through feature extraction, compression, and distillation, and are queried at Stage 5 to produce a 3D result.

Figure 1: vgrep3d five-stage pipeline: video frames → COLMAP + gsplat → SAM + SigLIP 2 feature extraction → per-scene autoencoder → feature field distillation → text query.

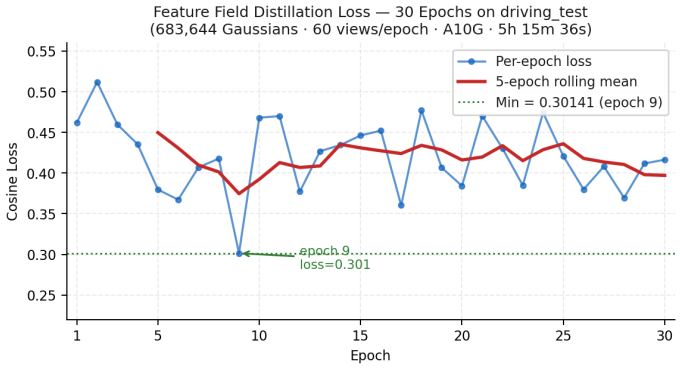


Figure 2: 30-epoch cosine distillation loss. Blue: per-epoch values; red: 5-epoch rolling mean; green dotted: global minimum at epoch 9 (0.30141). Total training: 5h 15m 36s on A10G.

Table 1: Complete 30-epoch distillation loss.

Ep.	Loss	Ep.	Loss
1	0.46215	16	0.45217
2	0.51198	17	0.36026
3	0.45978	18	0.47722
4	0.43537	19	0.40668
5	0.37969	20	0.38421
6	0.36723	21	0.47028
7	0.40708	22	0.43019
8	0.41778	23	0.38481
9	**0.30141**	24	0.47428
10	0.46816	25	0.42090
11	0.47007	26	0.37992
12	0.37714	27	0.40821
13	0.42693	28	0.36947
14	0.43444	29	0.41179
15	0.44633	30	0.41644

4 Experiments

4.1 Feature Field Distillation

Figure 2 and Table 1 report the complete per-epoch distillation loss from `field_train_run1.log` (captured 2026-08-22). Loss does not decrease monotonically due to random 60-view epoch sampling. Mean loss: epochs 1–15: 0.430; epochs 16–30: 0.414.

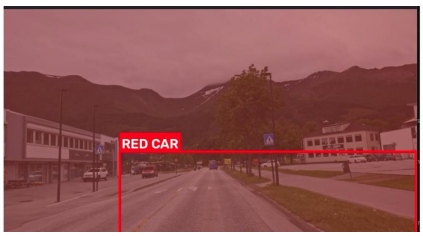
4.2 Qualitative Scene Results

Figure 3 shows the three key qualitative outputs. Panel (a) illustrates the core failure: the semantic heatmap activates across the entire scene for "red car," not just vehicles. Panel (b) shows the quality of the underlying Gaussian reconstruction. Panel (c) shows the hybrid Grounding DINO fallback, which produces tight 2D detections backprojected into 3D — architecturally decoupled from the feature field.

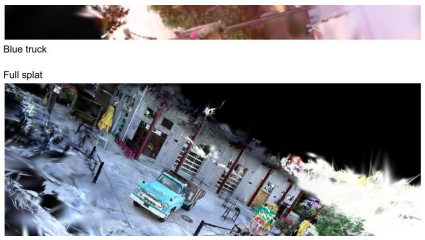
4.3 3D Query Results

Figure 4 and Table 2 report results for all six prompts (latency 36–42 s/prompt on A10G).

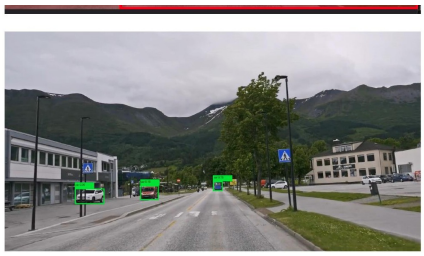
"Sky" is the clear outlier: 52,008 Gaussians (7.6%) with centroid $y{=}-0.239$ versus $+0.06$–$+0.28$ for all other prompts. This directional shift confirms scene-level geometric structure is encoded. "Red car," "stop sign," "person," and "building" each match 63–67% of the scene with near-identical AABB extents — per-object localization is impossible.



(a) "Red car" semantic heatmap



(b) Full Gaussian splat render



(c) Grounding DINO 2D detections

Figure 3: Qualitative results from the vgrep3d pipeline on the driving_test scene. **(a)** The "red car" heatmap overlaid on the dashcam frame. The red tint covers 63.9% of the scene, quantifying the embedding flatness failure mode — the field cannot isolate the vehicle from the background. **(b)** Novel-viewpoint render of the full 683,644-Gaussian scene (gsplat v1.5.3, sh_degree=3). The blue vintage truck is visible in the scene centre. **(c)** Grounding DINO 2D detections on the original dashcam frame — the hybrid workaround. Tight 2D bounding boxes are backprojected to 3D via camera intrinsics, producing accurate 3D localization that the feature field alone cannot achieve.

Figure 2. Query Results — driving_test Scene (N=683,644 Gaussians)

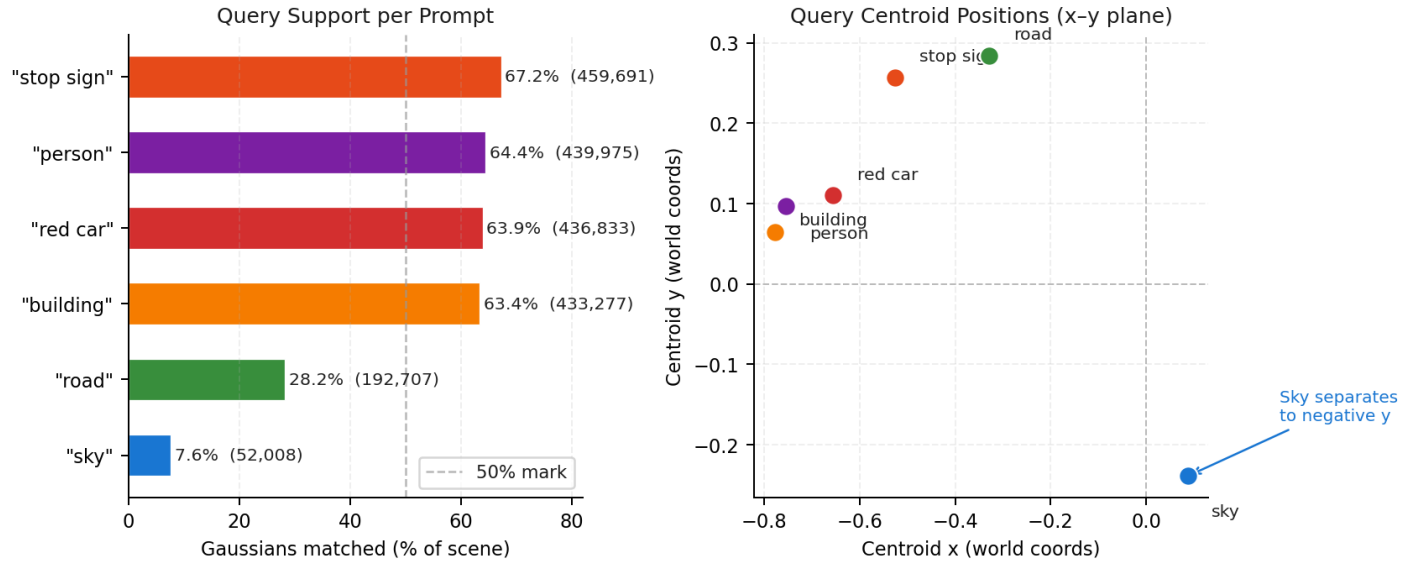


Figure 4: Query results. *Left*: Gaussian support as % of scene per prompt. *Right*: centroid positions in the x–y plane. "Sky" uniquely separates to negative y, confirming scene-level geometric encoding.

Table 2: Query results on the driving_test scene (N=683,644 Gaussians). "% scene" = support / N. Latency: 36–42 s/prompt on A10G (including model load).

Prompt	Centroid (x, y, z)	Support	% scene	AABB $(w \times h \times d)$
"red car"	$(-0.655,\ 0.110,\ 0.116)$	436,833	63.9%	$6.53 \times 2.60 \times 0.95$
"road"	$(-0.328,\ 0.284,\ 0.023)$	192,707	28.2%	$7.10 \times 2.90 \times 0.73$
"stop sign"	$(-0.525,\ 0.257,\ 0.127)$	459,691	67.2%	$6.51 \times 2.89 \times 0.93$
"person"	$(-0.754,\ 0.097,\ 0.120)$	439,975	64.4%	$6.47 \times 2.54 \times 0.94$
"building"	$(-0.778,\ 0.064,\ 0.129)$	433,277	63.4%	$6.46 \times 2.49 \times 0.96$
"sky"	$(\ 0.087, -0.239,\ 0.101)$	52,008	**7.6%**	$5.83 \times 1.39 \times 0.91$

4.4 Autoencoder Compression

The autoencoder achieves 72× compression: 3.15 GB (raw SigLIP 2, float32) → 43.8 MB (16-dim latents). The 69 MB feature cache is not required at query time.

5 Analysis: Why Tight 3D Bounding Boxes Fail

Despite correct scene-level semantics, tight per-object 3D bounding boxes are not achievable. We identify two root causes with direct evidence.

Root cause 1: View diversity. All 292 frames share a forward-

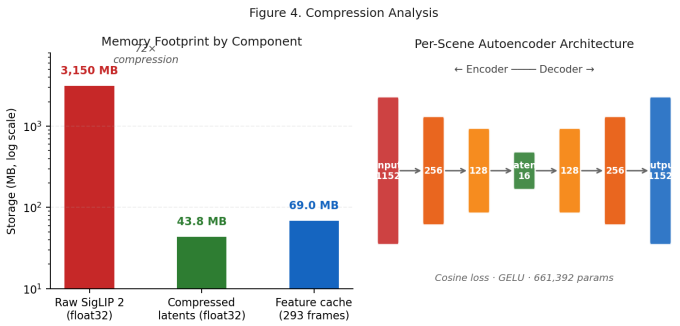


Figure 5: Left: storage footprint (log scale) — 72× compression. Right: per-scene autoencoder architecture.

facing vantage with translational-only camera motion. Gaussians at similar depths receive no contrastive gradients — there is no viewpoint from which they appear in different visual contexts. All published benchmarks use captures orbiting objects or covering 360°.

Root cause 2: SigLIP 2 crop-level embedding flatness. We confirmed the SAM per-mask path ran — frame 0150 contains 92 distinct masks with 1152-dim embeddings. The failure is *not* "we should add SAM." SigLIP 2's image tower, trained for whole-image captioning alignment, produces compact-clustered embeddings for driving-scene crop fragments (car door, road surface, building facade). The distillation objective receives weak contrastive signal, resulting in 63–67% of Gaussians receiving elevated relevance for any given query (Figure **??**).

Hybrid workaround. Grounding DINO 2D detection + depth backprojection yields tight 3D boxes (Figure **??**) but is architecturally decoupled from the feature field — 2D detection dressed in 3D geometry, not 3D-native semantic understanding.

6 Limitations and Future Work

No standard-benchmark evaluation. We report no mIoU or localization accuracy on LERF-OVS or 3D-OVS — the highest-priority extension. Running on the LERF scenes (*ramen*, *teatime*, *figurines*, *kitchen*) would enable direct comparison with LangSplat and OpenGaussian.

Single-scene evaluation. Generalization across scenes with different view diversity and Gaussian counts is unknown.

SigLIP 2 crop-embedding gap. Future work should explore DINOv2 [Oquab et al., 2024] patch features, dense CLIP tokens, or instance-contrastive VLMs better suited to scene fragment embedding.

Instance-contrastive distillation. Replacing per-pixel cosine loss with same-object attraction / different-object repulsion terms (cf. SAGA, Gaussian Grouping) could address semantic flatness more directly.

Capture requirements. A multi-viewpoint rig (stereo or 360°) would dramatically improve per-Gaussian separability for outdoor scenes.

7 Conclusion

We presented vgrep3d, an end-to-end pipeline for open-vocabulary 3D object querying using SigLIP 2 and SAM-guided per-mask feature distillation on a real-world dashcam scene. The pipeline correctly encodes scene-level semantics — "sky" activates only 7.6% of Gaussians with a geometrically distinct centroid — and achieves 72× memory compression. Tight per-object bounding boxes remain out of reach for forward-facing dashcam capture, with two evidence-based root causes. All code, trained artifacts, and figures are at https://github.com/ryamabra/vgrep3d.

References

Anonymous. OutLangSplat: 3D language Gaussian Splatting for UAV outdoor scenes. *arXiv preprint arXiv:2608.04560*, 2025.

J. Guo, X. Ma, Y. Fan, H. Liu, and Q. Li. Semantic Gaussians: Open-vocabulary scene understanding with 3D Gaussian Splatting. *arXiv preprint arXiv:2403.15624*, 2024.

T. Huang, R. Chen, D. Hu, et al. OpenInsGaussian: Open-vocabulary instance Gaussian segmentation with context-aware cross-view fusion. *arXiv preprint arXiv:2510.18253*, 2025.

B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3D Gaussian Splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik. LERF: Language embedded radiance fields. *ICCV*, 2023.

M. Oquab, T. Darcet, T. Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister. LangSplat: 3D language Gaussian Splatting. *CVPR*, 2024.

A. Radford, J. W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. *CVPR*, 2016.

J.-C. Shi, M. Wang, H.-B. Duan, and S.-H. Guan. Language embedded 3D Gaussians for open-vocabulary scene understanding. 2024.

W. Sun, Y. Zhou, J. Jiao, and Y. Li. CAGS: Open-vocabulary 3D scene understanding with context-aware Gaussian Splatting. *arXiv preprint arXiv:2504.11893*, 2025.

M. Tschannen, A. Gritsenko, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Y. Wu, J. Bao, X. Zhang, et al. OpenGaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *arXiv preprint arXiv:2406.02058*, 2024.

Y. Yan, H. Lin, C. Zhou, et al. Street Gaussians: Modeling dynamic urban scenes with Gaussian Splatting. 2024.

V. Ye, R. Bhatt, et al. gsplat: An open-source library for Gaussian Splatting. *arXiv preprint arXiv:2409.06765*, 2024.

X. Zhou, Z. Lin, X. Shan, et al. DrivingGaussian: Composite Gaussian Splatting for surrounding dynamic autonomous driving scenes. 2024.