EmbodiedSplat : Online Feed-Forward Semantic 3DGS for Open-Vocabulary 3D Scene Understanding

Seungjun Lee *Zihan Wang* *Yunsong Wang* *Gim Hee Lee*

National University of Singapore

CVPR JUNE 3-7, 2026 DENVER COLORADO

Project Page

Linkedin

Introduce our EmbodiedSplat

+300 views

Streaming Images

Build and Understand at Once!

EmbodiedSplat

Feed Forward

up to 5-6 FPS

Where can I sit?

Online Whole-Scene 3DGS Reconstruction

Novel View Synthesis

Depth Rendering

Bag

Chair

3D Semantic Segmentation

2D-rendered Segmentation

Open-Vocabulary Scene Understanding

Online 3DGS reconstruction and *Semantic Reasoning* at the same time!

Perception model for Embodied Agent requires:

1. ***Online***: Process streaming images incrementally to reconstruct the 3D scene along its exploration
2. ***Real-time***: High inference speed is required to be synchronized with agent's exploration
3. ***Highly-generalizable***: Model should be generalizable to the novel scenes.
4. ***Whole-scene Reconstruction***: Model should support long video for the long-term planning and actions.
5. ***Open-vocabulary Understanding***: Perceive a wide range of objects described with natural-form of language.

Current semantic 3DGS is not applicable for embodied scenarios!

Methods	Online	Near Real-time	Feed-Forward	Whole-scene Reconstruction	Open-vocabulary
LangSplat-v1, v2	✗	✗	✗	✓	✓
OpenGaussian	✗	✗	✗	✓	✓
InstanceGaussian	✗	✗	✗	✓	✓
Dr. Splat	✗	✗	✗	✓	✓
LSM	✗	✓	✓	✗	✓
SIU3R	✗	✓	✓	✗	✓
Online-LangSplat	✓	😕	✗	✓	✓
EA3D	✓	😕	✗	✓	✓
EmbodiedSplat	✓	✓	✓	✓	✓

Our EmbodiedSplat supports all of them!

- +300 images → Semantic 3DGS
- Online, Whole-scene recon
- Feed-forward design
- Near real-time
- Memory efficient design

Support:

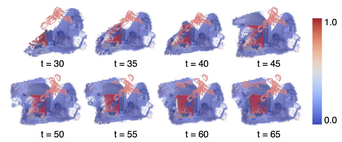
1. *3D Semantic Segmentation*
2. *2D-rendered segmentation*
3. *Novel-view RGB synthesis*

Our contribution is…

- Novel framework for embodied 3D perception which enables online, whole-scene reconstruction for language-embedded 3DGS with up to 5-6 FPS inference speed.
- Combination of 2D CLIP Features with rich semantic capabilities and 3D CLIP Features with geometric prior.
- Sparse Coefficient Field with CLIP Global Codebook to store the per-Gaussian language embeddings compactly.

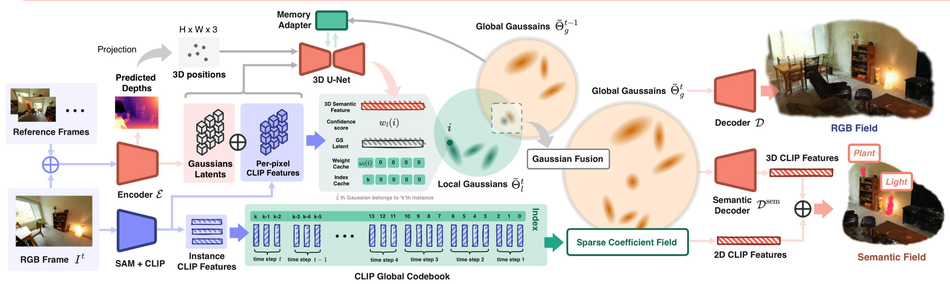
Check Webpage!

Online 3D Semantic Reasoning



2D-rendered Segmentation

Method



3D CLIP Features

2D CLIP features lack the 3D geometry prior.

Adopt 3D U-Net and Memory adapter to aggregate 3D geometry prior.

→ Distill CLIP features into 3D module.

Ensemble!

CLIP Global Codebook and *Online Sparse Coefficient Field*

Binding full CLIP features to every Gaussians are memory-expensive!

- Auto encoder-decoder, PQ → Need additional pretraining
- Per-scene Optimized Codebook → Not generalizable

We propose CLIP global codebook and online sparse coefficient field for memory-efficient semantic Gaussians. No pretraining, Highly generalizable!

Inference Strategy

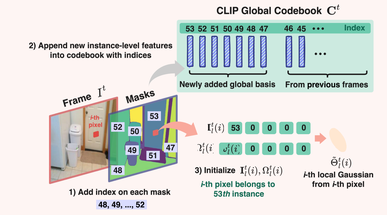
After reconstructing the semantic Gaussians, per-Gaussian CLIP feature is recovered through sparse linear combination of codebook vectors:

$$\mathbf{s}_g^T(i) = \sum_{j=1}^{L-1} \Omega_g^T(i,j)\, \mathbf{C}^T(\mathbf{I}_g^T(i,j)), \quad \sum_{j=1}^{L-1} \Omega_g^T(i,j) = 1. \quad (4)$$

Codebook vectors store original CLIP features.

→ No information loss but memory-efficient!

Sparse Coefficient Field Initialization



Online Fusion

Global j

Local GS i

Confidence Weighted-sum

Algorithm 1.

Fig. 9

Fast 3D Search

Computing cosine similarity for every Gaussians are expensive! → $O(MD)$

Instead, we compute cosine similarity between codebook vectors and text. And, we reuse it for all Gaussians, leveraging the sparse coefficient field.

$$\tilde{\mathbf{s}}_g^T(i) \approx \sum_{j=1}^{L-1} \Omega_g^T(i,j)\, \mathbf{C}^T(\mathbf{I}_g^T(i,j)), \quad \left\|\mathbf{C}^T(k)\right\|_2 = 1.$$

$$\cos(\mathbf{t}, \tilde{\mathbf{s}}_g^T(i)) \approx \sum_{j=1}^{L-1} \Omega_g^T(i,j)\, (\mathbf{t}, \mathbf{C}^T(\mathbf{I}_g^T(i,j)))$$

$$\approx \sum_{j=1}^{L-1} \Omega_g^T(i,j) \cos(\mathbf{t}, \mathbf{C}^T(\mathbf{I}_g^T(i,j))).$$

Complexity reduces to $O(KD + M(L-1))$, Where K is codebook size and $K \ll M$

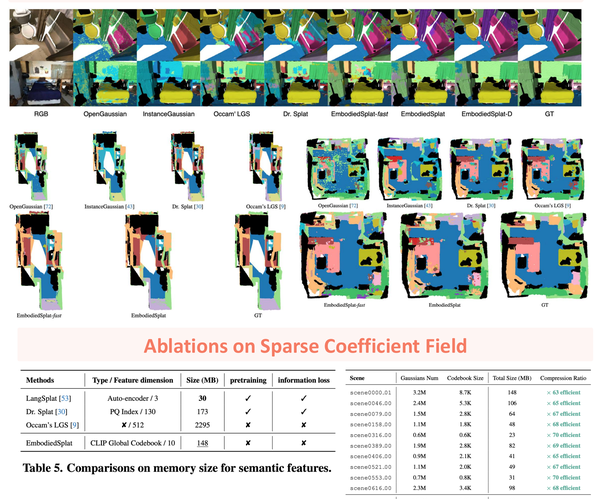
More fast search in 3D Gaussian!

Results

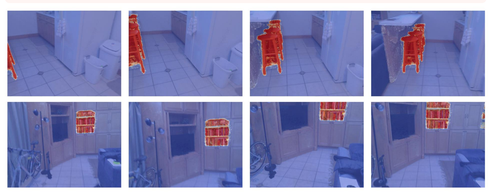
Quantitative results on ScanNet, ScanNet200, ScanNet++

Method	Search Domain	ScanNet [12] 10 classes mIoU	ScanNet [12] 10 classes mACC	ScanNet [12] 15 classes mIoU	ScanNet [12] 15 classes mACC	ScanNet [12] 19 classes mIoU	ScanNet [12] 19 classes mACC	ScanNet200 [56] 70 classes mIoU	ScanNet200 [56] 70 classes mACC	ScanNet++ [77] 20 classes mIoU	ScanNet++ [77] 20 classes mACC	Scene-reconstruction Time (363 images)	Per-Scene / Generalizable	on / off
LangSplat [53]	2D	6.52	20.11	3.25	13.16	1.34	7.44	0.72	4.39	2.21	10.18	~ 6 hr	Per-Scene	Offline
LEGaussians [59]	2D	6.79	18.13	4.13	15.49	2.53	5.67	1.39	5.45	2.93	9.34	~ 6 hr	Per-Scene	Offline
Online-LangSplat [31]	2D	7.13	21.56	3.89	14.52	3.45	8.97	2.45	4.12	4.51	11.34	5.4 min (1.12 FPS)	Per-Scene	Online
OpenGaussian [72]	3D	29.50	44.61	23.74	39.14	22.52	35.02	15.15	25.66	25.65	37.03	~ 2.5 hr	Per-Scene	Offline
Occam's LGS [9]	3D	42.14	70.28	35.04	63.71	30.49	57.91	20.32	40.49	34.08	61.19	~ 2 hr	Per-Scene	Offline
Dr. Splat [30]	3D	39.21	66.66	31.84	60.58	28.38	55.85	19.29	33.84	39.85	58.34	~ 2 hr	Per-Scene	Offline
InstanceGaussian [43]	3D	29.77	52.32	28.79	50.07	26.57	48.63	23.20	38.32	29.98	47.47	~ 3 hr	Per-Scene	Offline
EmbodiedSplat (RGB)	3D	**49.81**	**76.13**	**49.23**	**75.47**	**46.22**	**70.37**	**31.16**	48.38	41.93	61.50	8 min (0.75 FPS)	Generalizable	Online
EmbodiedSplat-*fast* (RGB)	3D	47.86	77.62	43.21	73.85	41.03	70.12	30.46	**55.31**	**45.53**	**71.42**	**1 min 10 sec (5.18 FPS)**	Generalizable	Online
EmbodiedSplat (RGB-D)	3D	57.41	82.45	55.18	80.27	52.12	75.66	34.75	52.36	44.03	66.27	8 min (0.75 FPS)	Generalizable	Online
EmbodiedSplat-*fast* (RGB-D)	3D	51.05	80.15	46.92	77.15	43.89	72.73	32.43	58.14	51.09	78.68	1 min 10 sec (5.18 FPS)	Generalizable	Online

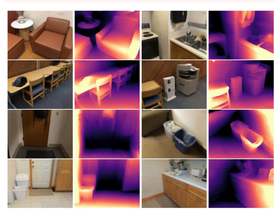
Qualitative Comparisons on 3D segmentation



2D-rendered Segmentation



RGB Synthesis



Ablations on Sparse Coefficient Field

Methods	Type / Feature dimension	Size (MB)	pretraining	information loss
LangSplat [53]	Auto-encoder / 3	**30**	✓	✓
Dr. Splat [30]	PQ Index / 130	173	✓	✓
Occam's LGS [9]	✘ / 512	2295	✘	✘
EmbodiedSplat	CLIP Global Codebook / 10	148	✘	✘

Table 5. Comparisons on memory size for semantic features.

Scene	Gaussians Num	Codebook Size	Total Size (MB)	Compression Ratio
scene0000_01	3.2M	8.7K	148	× 63 efficient
scene0046_00	2.4M	5.3K	106	× 65 efficient
scene0079_00	1.5M	2.6K	64	× 67 efficient
scene0158_00	1.1M	1.8K	48	× 68 efficient
scene0316_00	0.6M	0.6K	23	× 70 efficient
scene0389_00	1.9M	2.8K	82	× 69 efficient
scene0406_00	0.9M	2.1K	41	× 65 efficient
scene0521_00	1.1M	2.0K	49	× 67 efficient
scene0553_00	0.7M	0.8K	31	× 70 efficient
scene0616_00	2.3M	3.4K	98	× 68 efficient
Average	1.57M	3.0K	69	× 67 efficient