Beyond Bare Queries: Open-Vocabulary Object Grounding with 3D Scene Graph

Sergey Linok[1], Tatiana Zemskova[1,2], Svetlana Ladanova[1], Roman Titkov[1], Dmitry Yudin[1,2], Maxim Monastyrny[3], Aleksei Valenkov[3]

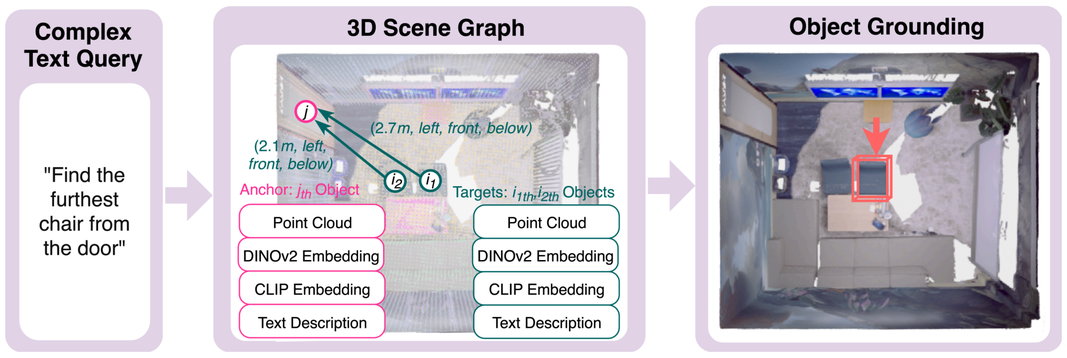
[1] Center for Cognitive Modeling, Moscow Institute of Physics and Technology, Dolgoprudny, Russia
[2] AIRI, Moscow, Russia
[3] Sberbank of Russia, Robotics Center, Moscow, Russia

group contact:
linok.sa@phystech.edu

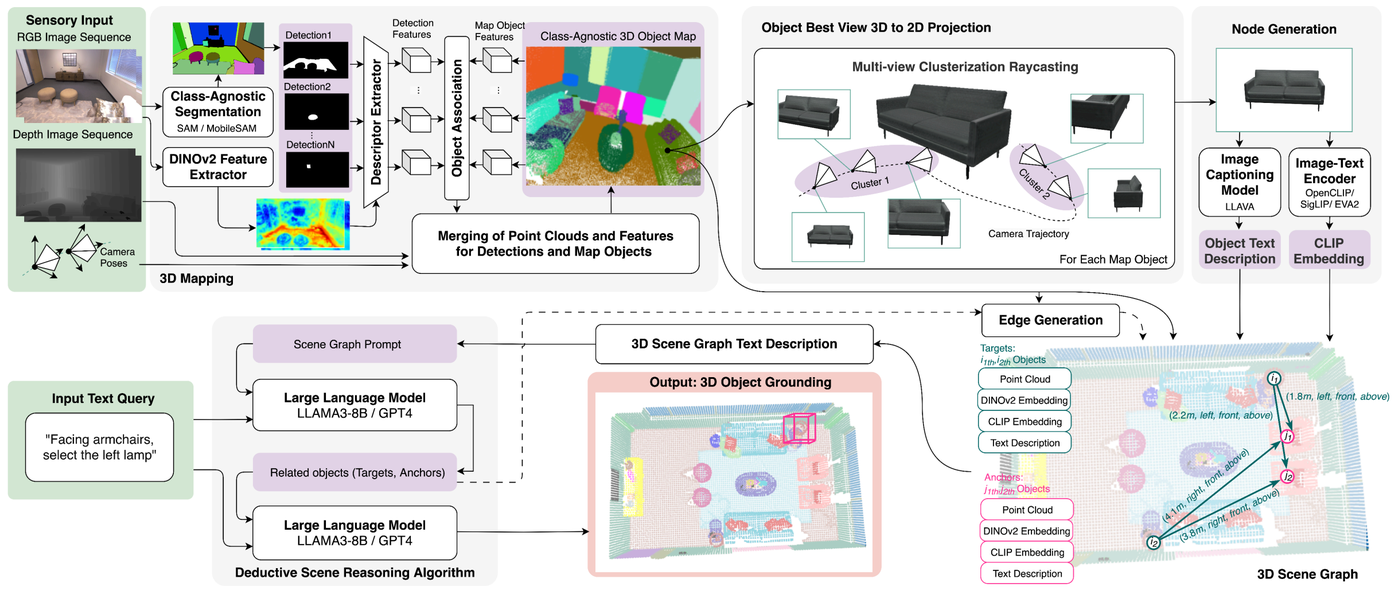
BBQ

Problem Definition

Research Question: How to represent 3DEnv for multi-hop open-vocabulary references?
Our Solution: 3D scene graph

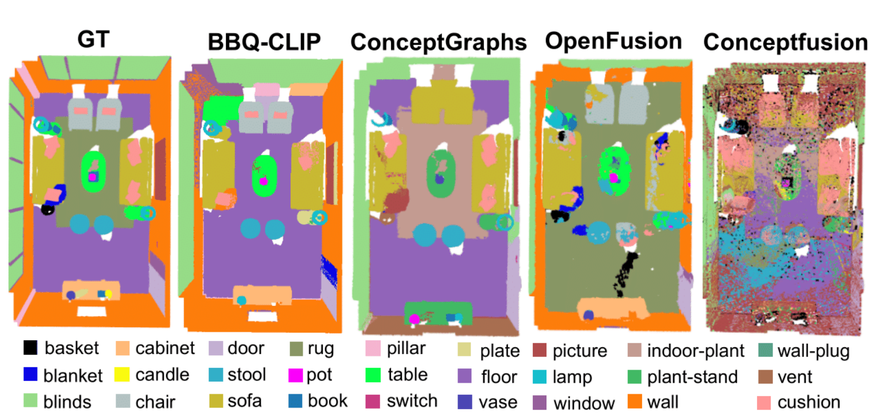


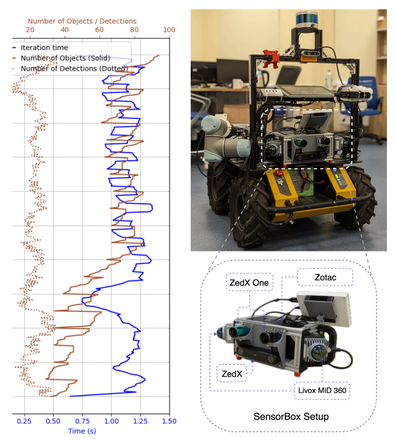
BBQ Overview



Our contributions:

- DINO features are great for 2.5D spatial scene understanding and can be applied for object aggregation.
- Aggregated mask, projected onto the best view, are sufficient to describe the object.
- With a carefully selected models and our design choices we can achieve close to realtime performance for 3D object mapping on the real robot.
- For spatial multi-hop reasoning, metric edges are a “cheap” source of valuable relations.
- Semantic spatial relations help to discern view-dependent relations based on the user’s query.
- The results on the Sr3D+, Nr3D, and ScanRefer datasets demonstrate the effectiveness of the BBQ modular method.





Results

3D open-vocabulary semantic segmentation benchmark

	Methods	Replica mAcc↑	Replica mIoU↑	Replica fmIoU↑	ScanNet mAcc↑	ScanNet mIoU↑	ScanNet fmIoU
Privileged	OpenFusion	0.41	0.30	0.58	0.67	0.53	0.64
Zero-Shot	ConceptFusion	0.29	0.11	0.14	0.49	0.26	0.31
	OpenMask3D	-	-	-	0.34	0.18	0.20
	ConceptGraphs	0.36	0.18	0.15	0.52	0.26	0.29
	BBQ-CLIP	**0.38**	**0.27**	**0.48**	**0.56**	**0.34**	**0.36**

Graph edges ablation study on Nr3D dataset (GT objects)

LLM	Edge	Recall@1 (Overall)	Recall@1 (View Independent)	Recall@1 (View Dependent)
Llama3-8B	-	36.1	36.1	36.0
Llama3-8B	Metric	<u>43.8</u>	<u>43.0</u>	46.3
Llama3-8B	Semantic	41.4	37.8	**53.0**
Llama3-8B	Metric+Semantic	**45.5**	**43.3**	<u>52.4</u>
GPT-4o	-	61.8	67.8	43.7
GPT-4o	Metric	**68.6**	**73.9**	52.4
GPT-4o	Semantic	50.5	49.2	<u>54.9</u>
GPT-4o	Metric+Semantic	<u>68.4</u>	<u>70.3</u>	**62.1**

Grounding accuracy on Sr3D+/Nr3D dataset

Sr3D+

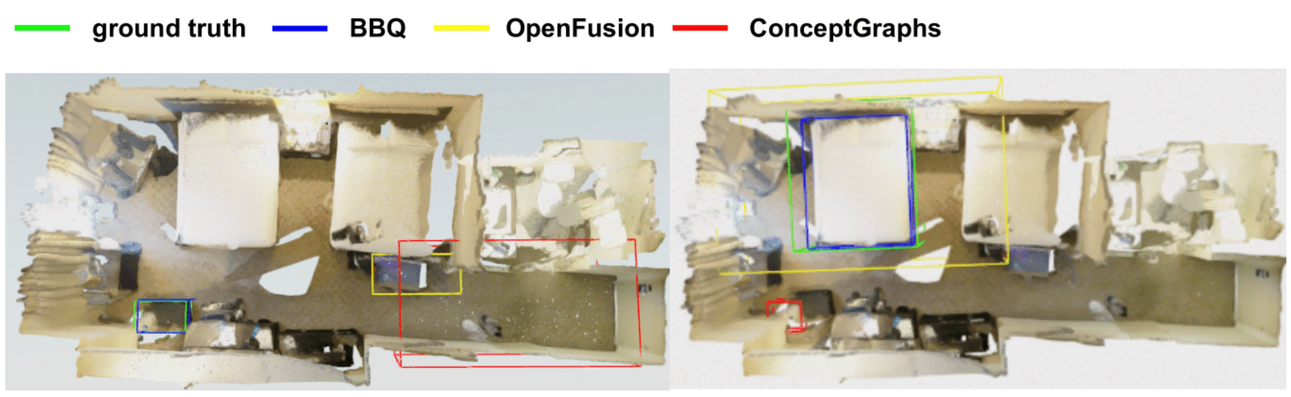
Methods	Overall A@0.1	Overall A@0.25	Easy A@0.1	Easy A@0.25	Hard A@0.1	Hard A@0.25	View Dep. A@0.1	View Dep. A@0.25	View Indep. A@0.1	View Indep. A@0.25
OpenFusion	12.6	2.4	14.0	2.4	1.3	1.3	3.8	2.5	13.7	2.4
BBQ-CLIP	14.4	8.8	15.4	9.0	6.7	6.7	11.4	5.1	14.4	8.8
ConceptGraphs	13.3	6.2	13.0	6.8	16.0	1.3	15.2	5.1	13.1	6.4
BBQ	**34.2**	**22.7**	**34.3**	**22.7**	**33.3**	**22.7**	**32.9**	**20.3**	**34.4**	**23.0**

Nr3D

Methods	Overall A@0.1	Overall A@0.25	Easy A@0.1	Easy A@0.25	Hard A@0.1	Hard A@0.25	View Dep. A@0.1	View Dep. A@0.25	View Indep. A@0.1	View Indep. A@0.25
OpenFusion	10.7	1.4	12.9	1.4	5.1	1.5	8.5	0.0	11.4	1.9
BBQ-CLIP	15.3	9.4	18.1	11.0	8.1	5.6	8.1	6.1	17.2	10.5
ConceptGraphs	16.0	7.2	18.7	9.2	9.1	2.0	12.7	4.2	17.0	8.1
BBQ	**28.3**	**19.0**	**30.5**	**21.3**	**22.8**	**13.2**	**23.6**	**18.2**	**29.8**	**19.3**

Grounding accuracy on ScanRefer dataset

Methods	A@0.25	A@0.5
LERF	4.4	0.3
OpenScene	13.0	5.1
LLM-Grounder	17.1	5.3
BBQ	**19.4**	**11.6**



a) Query: **"Object used to transport clothes while traveling, found on the right of the TV"**
b) Query: **"Select the bed that is near the backpack"**

Conclusion

- With BBQ, we advance the limits of 3D scene perception by integrating language models with general world knowledge and our scene-specific graph representation.
- Results on Nr3D, Sr3D, ScanRefer, and real-world data demonstrate that leveraging metric and semantic scene edges enables more comprehensive and flexible 3D scene understanding.
- We hope our efficient code will support BBQ in real-world robotics, bridging communication between humans and autonomous agents.
- It should be noted that our method assumes static, room-sized environments; in larger spaces, 3D scene graphs may not capture spatial relations accurately.
- Conducted experiments highlight that our 3D object-centric map construction method is limited in its ability to successfully distinguish tiny objects in the image. Therefore, BBQ requires more scene exploration where the camera is placed closer to objects of interest to successfully map such instances.

Literature

Linok, S., Zemskova, T., Ladanova, S., Titkov, R., Yudin, D., Monastyrny, M., & Valenkov, A. (2024). Beyond Bare Queries: Open-Vocabulary Object Grounding with 3D Scene Graph. arXiv preprint arXiv:2406.07113.