OSMa-Bench++: Toward Open-Ended Benchmarking of Semantic Mapping for Manipulation with Prompt-Generated Synthetic Scenes

Regina Kurkova[1*], Maxim Popov[1] and Sergey Kolyubin[1]

***Abstract*— Semantic mapping methods are increasingly used as intermediate scene representations for downstream robotic reasoning and manipulation, yet their evaluation is still largely tied to fixed benchmark datasets with limited coverage of manipulation-relevant corner cases. In this work, we extend OSMa-Bench [1] toward controllable benchmarking with prompt-generated synthetic indoor scenes. Our pipeline automatically generates scene descriptions, synthesizes corresponding environments with SceneSmith [2], and adapts the resulting assets into an OSMa-Bench-compatible simulation format. This adaptation requires a nontrivial intermediate layer, including semantic normalization, material and texture repair, shader fallback policies, floor handling, navigation setup, and controlled lighting configuration. A key advantage of the proposed setup is that the original scene-generation prompt is known in advance and can therefore serve as an auxiliary semantic specification of the intended scene. We use this property to extend the VQA component of OSMa-Bench with a prompt-grounded question category. The resulting framework supports targeted stress-testing of semantic scene representations under conditions such as clutter, small objects, partial occlusions, and lighting variation, and makes benchmarking more extensible and better aligned with downstream manipulation requirements. Our code is available at `https://github.com/be2rlab/OSMa-Bench-v2`.**

I. Introduction

Semantic mapping is increasingly used not only for perception, but also as an intermediate representation for downstream robotic reasoning and manipulation. In this setting, it is not sufficient to produce a visually plausible reconstruction or a semantically labeled map: the representation must preserve object presence, relations, layout, and accessibility cues required for action-oriented queries.

Recent progress in open-vocabulary and object-centric scene representation has substantially improved the expressiveness of robotic maps. Methods such as OpenScene [3], OpenMask3D [4], OpenIns3D [5], PLA [6], ConceptGraphs [7], and BBQ [8] show that vision-language features, 3D reconstruction, and scene graph abstractions can support flexible semantic querying in complex environments. At the same time, evaluation of such methods remains constrained by fixed datasets and closed scene collections. However, Habitat-based datasets and embodied QA benchmarks still underrepresent manipulation-relevant corner cases such as cluttered small objects, partial occlusions, and restricted access to target objects.

In our previous work on OSMa-Bench [1], we observed that such corner cases can expose substantial differences between methods even when their behavior appears similar on standard scenes. This motivates moving from fixed-scene evaluation toward controllable generation of targeted scenarios. In this work, we extend OSMa-Bench with a prompt-driven synthetic scene generation pipeline based on SceneSmith [2]. We generate indoor scene descriptions, synthesize corresponding scenes, adapt them to an OSMa-Bench-compatible simulation format, and evaluate semantic representations under multiple conditions.

A key advantage of the proposed setup is that the original scene-generation prompt is known in advance and can therefore serve as an auxiliary semantic specification of the intended scene. We use this property to extend the VQA component of OSMa-Bench with a prompt-grounded category, enabling evaluation not only of what is visible in rendered trajectories, but also of how well the recovered scene representation remains consistent with the intended scene structure. Overall, the proposed framework makes benchmarking more extensible by enabling targeted generation of challenging manipulation-oriented evaluation scenarios.

II. Methodology

Our pipeline **OSMa-Bench++** extends original method with controllable generation of synthetic indoor scenes for manipulation-oriented evaluation of semantic scene representations. The overall workflow is illustrated in Fig. 1. Starting from textual scene descriptions, we synthesize indoor environments with SceneSmith [2], convert the resulting assets into a Habitat-compatible format, generate RGB-D observation sequences with the HaDaGe generator [1] used in OSMa-Bench, run semantic mapping approaches on these sequences, and finally evaluate the resulting representations through segmentation metrics and VQA-based scene graph assessment.

We begin from automatic generation of scene descriptions, corresponding to the prompt generation block in Fig. 1. A large language model first produces an initial pool of candidate prompts describing indoor scenes with object layouts and relations relevant to downstream manipulation. These prompts are then embedded into a semantic feature space and compared using cosine similarity,

$$\mathrm{sim}(p_i, p_j) = \frac{f(p_i)^\top f(p_j)}{\|f(p_i)\|_2 \, \|f(p_j)\|_2}, \tag{1}$$

where $f(\cdot)$ denotes the text embedding model.

This representation is used in two stages. First, near-duplicate prompts are removed to avoid repeatedly generat-

[1] Biomechatronics and Energy-Efficient Robotics (BE2R) Lab, ITMO University, Saint Petersburg, Russia

[*] Corresponding author: rekurkova@itmo.ru

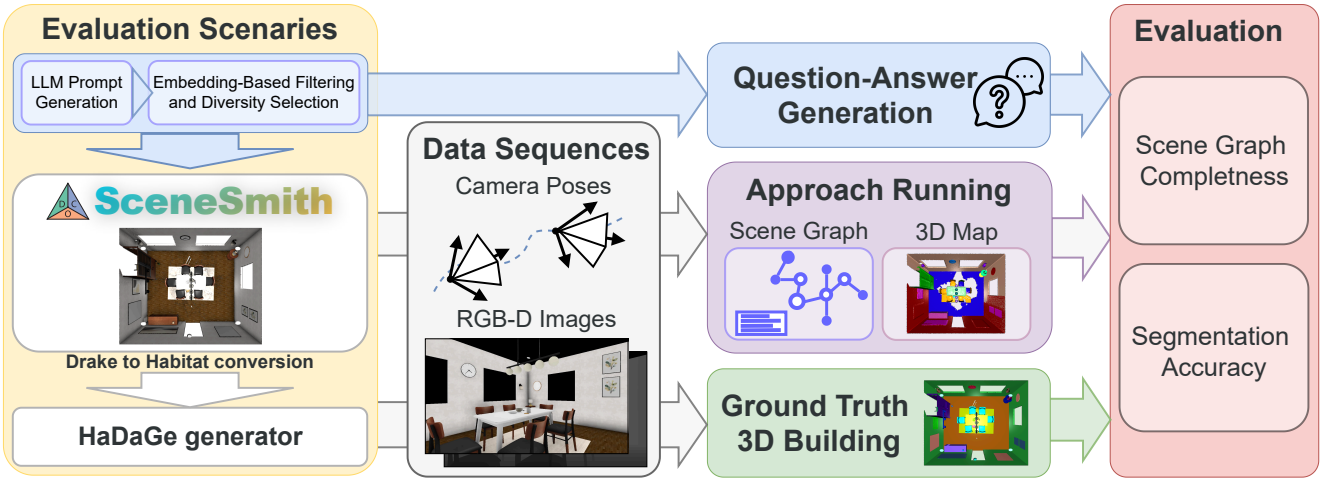

Fig. 1: Overview of the **OSMa-Bench++** pipeline. Scene descriptions are generated and filtered to obtain diverse evaluation scenarios, synthesized with SceneSmith, converted for use in Habitat-based sequence generation with HaDaGe, processed by semantic mapping methods, and evaluated through segmentation and scene graph completeness metrics.

ing semantically similar scenes. Second, from the remaining pool we select the most diverse subset by favoring prompts that are far from those already chosen in the embedding space. In practice, this yields a set of scene descriptions that covers a broader range of layouts and interaction-relevant situations than random sampling or manual prompt writing alone.

Since SceneSmith outputs are in Drake [9] simulation format and are not directly compatible with OSMa-Bench, we convert them into a Habitat-like dataset structure, constructing scene and object configuration files, assigning semantic categories, and resolving asset-level inconsistencies. After conversion, the scenes are passed to the OSMa-Bench [1], which produces RGB-D observation sequences along camera trajectories.

Finally, the generated sequences are processed by the evaluated approaches, which produce 3D maps and scene graphs. The benchmark evaluates both segmentation quality and scene graph completeness through question answering.

In addition to standard VQA categories, we use the original scene-generation prompt as an auxiliary source of semantic specification for question generation. This makes it possible to probe not only what is visible in the rendered sequence, but also how well the recovered representation remains consistent with the intended scene layout described at the prompt level.

III. Experiments

A. Experimental Setup

We evaluate the proposed extension on SceneSmith-generated indoor scenes integrated into OSMa-Bench. We consider two open semantic mapping methods, ConceptGraphs [7] and BBQ [8], which produce structured scene representations for downstream question answering.

The evaluation set is constructed from automatically generated prompts.From an initial pool of 350 prompts, we retain diverse candidates using the filtering procedure in Section II and keep only technically valid scenes. The final benchmark contains 40 scenes, split into 24 *furniture* scenes and 16 *manipuland* scenes. Prompt-to-scene correspondence was manually checked for both categories. Scenes with substantial mismatches were excluded during quality control. In the final set of 40 scenes, only four minor synthesis errors were observed, each corresponding to one missing object. This gives a scene-level full-match rate of $36/40 = 90.0\%$. Assuming at least five prompted objects per retained scene, this corresponds to a conservative lower-bound object-level fidelity of at least 98.0%.

We generate two complementary scene families. The *furniture* subset is used to evaluate whether a representation preserves room-scale structure, object layout, support relations, and coarse spatial dependencies in scenes dominated by large static objects. This subset is also useful early in synthesis because such scenes are easier to generate and export. The *manipuland* subset adds small interaction-relevant objects placed on tables, shelves, counters, and other supporting surfaces. These scenes are more challenging because they increase clutter, introduce fine-grained support and proximity relations, and include small, partially occluded, or densely arranged objects. Together, the two subsets separate room-scale structural failures from small-object and manipulation-oriented failures. For prompt generation and prompt-grounded question generation, we use GPT-4.1 [10].

Following OSMa-Bench [1], we evaluate each scene under four lighting conditions. The *baseline* uses static, non-uniformly distributed light sources natively available in the scene. The *nominal lights* configuration removes all explicit sources, relying solely on mesh-emitted light. The *camera light* condition attaches a directed light source to the camera, producing strong view-dependent shadows. Finally, *dynamic lighting* varies illumination along the robot trajectory, simulating realistic transitions between differently lit areas.

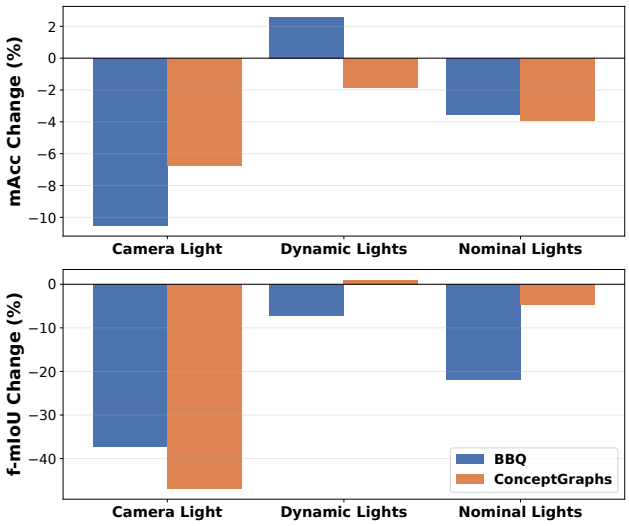

Fig. 2: Illustrative results on SceneSmith-derived **OSMa-Bench++** scenes. Top: relative change in semantic segmentation performance (mAcc) for BBQ [8] and ConceptGraphs [7] under different lighting conditions with respect to the baseline. Bottom: corresponding change in f-mIoU.

B. Prompt-Grounded Evaluation Protocol

In addition to the standard OSMa-Bench evaluation, we introduce a prompt-grounded protocol in which the original scene description serves as an auxiliary source of semantic ground truth. This protocol is particularly useful for categories that are difficult to assess robustly in standard trajectory-based VQA. In the original OSMa-Bench setting, the answer to a question may depend on what is visible from a specific trajectory or viewpoint. This is especially problematic for questions about object counts and inter-object relations, because the relevant objects may be present in the scene but only partially visible, occluded, or never jointly observable from the chosen camera path. As a result, these categories are among the least stable ones when questions are generated purely from view-dependent observations.

For this reason, we focus the prompt-grounded analysis on the *Measurements* and *Relations* between objects categories. These two categories are the most informative for scene graph completeness and are also the most strongly affected by the limitations of view-dependent question generation. Measurement questions probe whether the representation preserves the number of relevant entities, which is sensitive to missed small objects, duplicated instances, and incorrect object merging. Relation questions probe whether the representation preserves the spatial and structural dependencies between entities, such as support, proximity, containment, and relative arrangement. In a prompt-grounded setting, these questions no longer depend on whether a specific camera trajectory happened to expose the required relation.

This design is especially important for manipulation-oriented evaluation. For downstream manipulation, it is often not enough to know that an object exists; the representation must also preserve how many relevant objects are present and how they are arranged with respect to one another. The prompt-grounded protocol therefore complements the original OSMa-Bench VQA evaluation by targeting categories that are semantically central yet difficult to assess robustly through view-dependent observations alone.

C. Results on Segmentation and Prompt-Grounded Evaluation

Table I and Fig. 2 show the relative change in segmentation performance with respect to *Baseline*. *Camera* lighting produces the strongest degradation for both methods, while *Dynamic* lighting has a mild effect with small positive changes in mAcc. The divergence between mAcc and f-mIoU reflects a structural difference between the two approaches: BBQ tends to produce fewer but geometrically precise masks, yielding high f-mIoU yet lower scene-graph completeness, since some objects are filtered out entirely. ConceptGraphs recovers more objects overall, resulting in higher completeness but less precise masks. These complementary failure modes confirm that segmentation robustness cannot be assessed from a single metric alone.

Table II reports prompt-grounded question answering accuracy for BBQ and ConceptGraphs on the *Measurements* and *Relations* categories, separated into the *Furniture* and *Manipuland* subsets. Overall, both methods score below 30%, yet reveal distinct robustness profiles. BBQ, being a more advanced scene-graph method with stronger question-answering capabilities, generally outperforms ConceptGraphs on *Relations* and on the *Manipuland* subset. However, BBQ shows a notable drop under *Dynamic* lighting, which we attribute to its object description strategy: BBQ generates per-object embeddings from a single image, so a frame captured under unfavorable conditions directly degrades the object's representation. ConceptGraphs, by contrast, aggregates evidence across multiple views, which stabilizes its embeddings under *Dynamic* lighting and explains its gains in that condition, particularly on *Furniture* categories.

Table III compares standard image-derived VQA from OSMa-bench with prompt-grounded questions for the same categories. Unlike standard VQA, which inherits trajectory-dependent visibility and perception biases, PromptGT directly evaluates prompt-specified counts and relations.

Overall, these experiments highlight the main advantage of the proposed extension. By generating scenes from prompts, we produce more diverse evaluation environments than a fixed benchmark and assess categories that are difficult to evaluate in a view-dependent setup. In particular, prompt-grounded *Measurements* and *Relations* between objects provide a more stable probe of scene graph completeness, while the relative-to-baseline semantic segmentation analysis shows how these graph-level effects relate to the quality of the underlying 3D semantic reconstruction.

IV. Conclusion

We presented **OSMa-Bench++**, an extension of original OSMa-Bench that enables controllable benchmarking

TABLE I: Semantic segmentation performance on generated sequences under different lighting conditions.

Method	Baseline		Camera Light		Dynamic Lights		Nominal Lights	
	mAcc	f-mIoU	mAcc	f-mIoU	mAcc	f-mIoU	mAcc	f-mIoU
ConceptGraphs [7]	**57.6**	27.9	**53.7**	14.8	**56.5**	28.0	**55.3**	26.6
BBQ [8]	47.0	**68.4**	42.0	**42.8**	48.2	**63.5**	45.3	**53.5**

TABLE II: Accuracy (%) of prompt-grounded question answering for BBQ [8] and ConceptGraphs [7]. Values are computed over all questions in each category across 40 scenes.

Subset	Cat.	Method	Baseline	Camera	Dynamic	Nominal
Furn.	Measur.	CG	12.2	9.2	**15.3**	**14.3**
		BBQ	**15.3**	**24.5**	12.2	12.2
	Relat.	CG	12.9	6.4	**20.5**	17.0
		BBQ	**18.7**	**24.6**	14.6	**18.1**
Manip.	Measur.	CG	12.9	11.4	17.1	12.9
		BBQ	**15.7**	**24.3**	**18.6**	**22.9**
	Relat.	CG	15.0	14.0	13.0	9.0
		BBQ	**16.0**	**21.0**	**14.0**	**19.0**

TABLE III: Question-source ablation for measurement and relational QA Accuracy. Standard denotes image-derived VQA from OSMa-Bench, while PromptGT denotes questions generated from the scene-generation prompt.

Subset	Cat.	Method	Standard	PromptGT
Furn.	Measur.	CG	**30.6**	12.2
		BBQ	**18.4**	15.3
	Relat.	CG	**19.7**	12.9
		BBQ	**24.6**	18.7
Manip.	Measur.	CG	**32.2**	12.9
		BBQ	**33.9**	15.7
	Relat.	CG	**22.2**	15.0
		BBQ	**26.3**	16.0

of semantic mapping methods with prompt-generated synthetic indoor scenes. By combining prompt generation, SceneSmith-based scene synthesis, Habitat-compatible conversion, and HaDaGe-based sequence generation, the proposed pipeline makes it possible to construct evaluation scenarios that are difficult to obtain in fixed datasets, including cluttered layouts, small interaction-relevant objects, and relation-heavy scenes. A central property of this setup is that the original scene-generation prompt is known a priori and can therefore be reused as an auxiliary semantic specification for evaluation.

The experiments show that this addition is useful in two ways. First, it expands the benchmark beyond a closed set of scenes and makes targeted stress-testing possible under controlled conditions. Second, the prompt-grounded protocol provides a more stable way to assess scene graph completeness for measurements and inter-object relations, which are otherwise strongly affected by trajectory-dependent visibility. The results also indicate that BBQ and ConceptGraphs respond differently to lighting changes and scene composition, confirming that graph-level semantic completeness and segmentation robustness capture complementary aspects of representation quality. Overall, **OSMa-Bench++** turns synthetic scene generation into a practical tool for manipulation-oriented evaluation of semantic scene representations.

REFERENCES

[1] M. Popov, R. Kurkova, M. Iumanov, J. Mahmoud, and S. Kolyubin, "Osma-bench: Evaluating open semantic mapping under varying lighting conditions," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[2] N. Pfaff, T. Cohn, S. Zakharov, R. Cory, and R. Tedrake, *Scenesmith: Agentic generation of simulation-ready indoor scenes*, 2026. arXiv: 2602.09153 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2602.09153.

[3] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser, et al., "Openscene: 3d scene understanding with open vocabularies," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 815–824.

[4] A. Takmaz, E. Fedele, R. W. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, "Openmask3d: Open-vocabulary 3d instance segmentation," *arXiv preprint arXiv:2306.13631*, 2023.

[5] Z. Huang, X. Wu, X. Chen, H. Zhao, L. Zhu, and J. Lasenby, "Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation," *arXiv preprint arXiv:2309.00616*, 2023.

[6] R. Ding, J. Yang, C. Xue, W. Zhang, S. Bai, and X. Qi, "Pla: Language-driven open-vocabulary 3d scene understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7010–7019.

[7] Q. Gu et al., "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2024, pp. 5021–5028.

[8] S. Linok et al., "Beyond bare queries: Open-vocabulary object grounding with 3d scene graph," *arXiv preprint arXiv:2406.07113*, 2024.

[9] R. Tedrake et al., *Drake: Model-based design and verification for robotics*, 2019.

[10] OpenAI, "Introducing gpt-4.1 in the api," OpenAI, Tech. Rep., 2025. [Online]. Available: https://openai.com/index/gpt-4-1/.