Fixed External Cameras as Common Prior Maps for Active 3D Scene Graph Generation

Giorgia Modi[12*], Davide Buoso[2], Giuseppe Averta[2], Daniele De Martini[1]
[1]Mobile Robotics Group (MRG), University of Oxford, UK
[2]Visual and Multimodal Applied Learning Lab (VANDAL), Politecnico di Torino, Italy

***Abstract*— Commonly available prior information, such as BIM models, floor plans, and remote sensing images, can provide valuable geometric and semantic context for autonomous robotic systems. In this paper, we treat observations from fixed external RGB cameras as *Common Prior Maps* (CPMs): wide-field views of the environment that initialize a semantic and geometric scene prior before any robot motion begins. We present an RGB-only framework for active, incremental 3D scene graph (3DSG) generation that seamlessly fuses observations from both onboard robot cameras and fixed external cameras within a single hardware-agnostic pipeline. By relying solely on RGB observations processed by a feed-forward 3D reconstruction model, the system treats all cameras - onboard or external - identically, requiring no hardware modifications. A graph-based active semantic exploration framework then directly leverages the partial scene graph to guide the robot toward regions of high semantic uncertainty, progressively completing and refining the prior. Experiments demonstrate that bootstrapping the scene graph with even a single external camera increases initial object recall by up to +79%, and that the richer context of the prior significantly improves the efficiency of subsequent active exploration.**

I. Introduction

Modern robotic systems navigating unstructured human environments – such as homes, hospitals and offices - heavily depend upon rich contextual visual scene understanding to perceive, decide, and act autonomously [1]. Traditional perception pipelines build either low-level dense metric maps that lack explicit semantic content, or produce object-centric outputs that fail to encode relational structure [2]. A structured representation that jointly captures *which* entities are present, *where* they exist in 3D space, and *how* they relate to one another is therefore essential for reliable robot autonomy. 3D scene graphs (3DSGs) have emerged as a powerful representation for this purpose [3], [4]. In a 3DSG, nodes correspond to object instances annotated with semantic labels and metric locations, while edges encode spatial and functional relationships between them. This compact and queryable structure supports downstream tasks including language and task planning [5], navigation [6], and manipulation [7]. However, constructing such representations robustly and efficiently in real-world environments remains challenging, especially when prior knowledge about the scene is only partially available.

This motivates the use of *Common Prior Maps* (CPMs), which have long been available in real-world deployments and offer rich geometric and semantic priors. We argue that observations from fixed external cameras, already deployed as surveillance or monitoring infrastructure in many buildings, can serve as a valuable and underutilized form of CPM for robot scene understanding.

To exploit this prior, the perception pipeline must seamlessly incorporate external camera observations, which are typically available as RGB images. We therefore present a framework for active 3D scene graph generation from RGB inputs alone. By removing the dependency on depth sensors and relying instead on feed-forward 3D reconstruction [8], the system processes all RGB inputs - onboard and external - through a unified pipeline. Within this framework, the 3DSG becomes the semantic and geometric representation extracted from external-camera observations, capturing the dominant layout, objects, and spatial relationships of the environment.

Our contributions are:

1) an RGB-only pipeline for 3D scene graph generation that enables seamless integration of fixed external cameras;
2) an active semantic exploration loop that uses the scene graph to select informative viewpoints;
3) experimental evidence that camera-infrastructure priors substantially improve scene graph completeness and exploration efficiency.

II. Related Work

3D Scene Graphs (3DSGs). Scene graphs couple object semantics with relational context, addressing ambiguities in purely object-centric systems [2]. Grounding this abstraction in 3D enables navigation, manipulation, and language-guided planning [9]. Systems such as Hydra [10] and SceneGraphFusion [11] build hierarchical representations from RGB-D SLAM, while ConceptGraphs [12] introduces open-vocabulary semantics via foundation models, but still assumes access to depths and camera poses. To the best of our knowledge, our system is the first to enable 3DSG generation in an RGB-only setting.

Active Perception. Active perception—controlling sensing to maximize information gain—has long been studied via Next-Best-View (NBV) planning [13]. Geometric

*Correspondent author: giorgiamodi@gmail.com

This work was supported by the EPSRC Impact Acceleration Account (IAA), and the UKRI InnovateUK project 'A general-purpose configuration management system for highly configurable systems validated in automated logistics'.

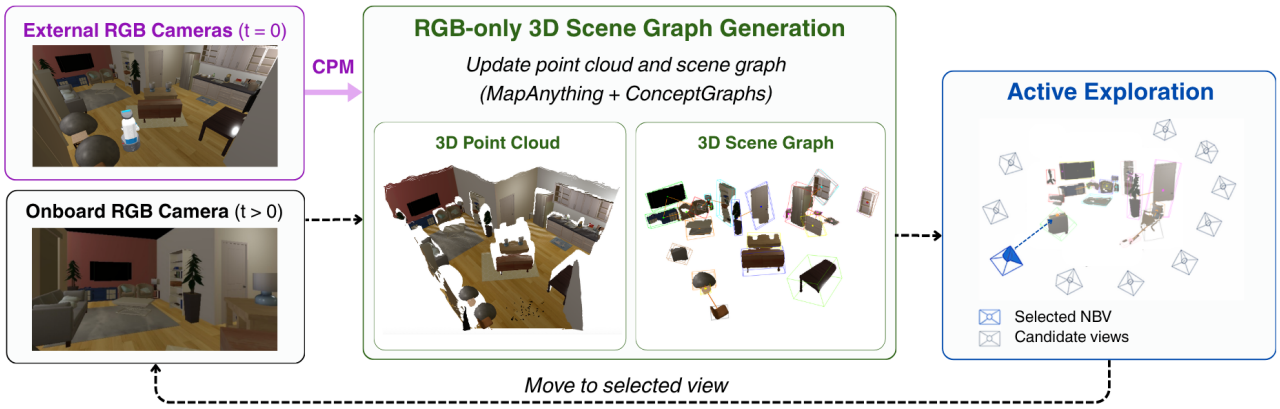

Fig. 1: System Overview. RGB observations from fixed external cameras serve as a Common Prior Map (CPM) and are processed by the RGB-only 3DSG generation pipeline to initialize the point cloud and 3D scene graph. For $t > 0$, the onboard robot camera provides new RGB observations that incrementally update the point cloud and scene graph. The active exploration module then uses the current scene graph to select the Next-Best-View (NBV) and the robot moves to the selected viewpoint to continue exploration.

frontier-based methods [14] and density-based approaches [15] efficiently expand coverage but ignore semantics. Recent work incorporates semantic reasoning: Ding et al. [16] use semantic scene completion to guide exploration toward uncertain regions, while Chen et al. [17] combine semantic and geometric uncertainty via 3D Gaussian Splatting. ASP [18] further enables LLM-based reasoning over scene graphs, turning them into active exploration tools. However, prior work on ASP has not considered settings with external camera inputs, nor RGB-only perception.

External sensing. While robotic perception is typically egocentric, fixed external cameras provide complementary global context. Prior work shows their value for visual servoing [19] and VLM-guided navigation [20], but their integration into 3DSGs remains unexplored.

III. Methodology

The architecture is modular, separating 3D metric reconstruction, scene graph generation, and active exploration into distinct stages.

A. System Overview

The system operates as an incremental perception–action loop, illustrated in Fig. 1. At initialization ($t = 0$), RGB observations from one or more fixed external cameras serve as a *Common Prior Map* (CPM). These observations are processed by the RGB-only 3DSG generation pipeline to initialize a point cloud and scene graph, providing semantic and geometric structure before robot motion begins.

For subsequent steps ($t > 0$), the robot acquires an RGB image from its current viewpoint $v_t \in \mathcal{V}$, where $\mathcal{V}$ denotes the set of navigable viewpoints. The same RGB-only pipeline processes this observation to update the reconstructed point cloud $\mathcal{P}_t$ and the scene graph $\mathcal{G}_t$. Based on the current graph, the active exploration module selects the Next-Best-View (NBV) v_{t+1} by maximizing expected information gain. The robot then moves to the selected viewpoint, acquires a new observation from the onboard camera, and repeats the loop, progressively extending and refining the initial CPM.

B. RGB-Only 3DSG Generation

Depths and Poses Estimation. The pipeline leverages MapAnything [8] to infer scene geometry from a set of RGB images. Given N inputs, the model predicts in a single forward pass a *factored* representation: per-view pixel ray directions R_i, up-to-scale depth maps $\tilde{D}_i$, poses $\tilde{P}_i$ in the reference frame of the first image, and a global metric scale factor $m \in \mathbb{R}$. These outputs are composed to back-project 2D pixels into a metrically scaled 3D point cloud. Since all poses are expressed in the reference frame of the first input image, fixing this reference - e.g., a known external camera location - is sufficient to anchor all subsequent observations in a consistent global frame, supporting reliable robot localization within the CPM's coordinate system.

Open-Vocabulary Scene Graph Construction. RGB images, together with estimated depths and poses, are passed to ConceptGraphs [12]: SAM [21] extracts class-agnostic instance masks, CLIP [22] embeds each region into a semantic descriptor, the masks are projected into 3D using the MapAnything-derived depth and poses, and multi-view association incrementally merges detections into 3-D object nodes $o_j \in \mathcal{O}_t$. While the original ConceptGraphs pipeline queries proprietary LLMs to infer relationships between objects [12], here spatial edges are instead derived through a deterministic, geometry-only procedure over oriented 3-D bounding boxes of the object nodes: for each ordered object pair, a fixed-priority set of predicates is evaluated - *on top of / supported by* (vertical contact with footprint overlap), *under / over* (vertical adjacency), *inside* (volumetric containment), and *next to* (horizontal proximity at similar height) - and at most one edge per ordered pair is assigned, yielding a reproducible scene graph. All geometric relation thresholds (e.g., maximum vertical gap, minimum footprint overlap, and horizontal proximity) are fixed across experiments and tuned

empirically once per dataset.

C. Active Semantic Perception

The framework incorporates Active Semantic Perception (ASP) [18] to actively explore the unknown environment using the current 3DSG. At each exploration step, ASP reasons over the scene graph produced so far by the RGB-only pipeline and uses it as the structured state for action selection. An LLM samples plausible completions of the unobserved scene conditioned on the current graph, enabling the system to infer what may lie beyond the currently explored regions. For each candidate viewpoint x, ASP computes expected information gain as the mutual information between the predicted observation Y_{k+1} and the current graph G_k:

$$I(Y_{k+1}; G_k \mid x) = H(Y_{k+1} \mid x) - H(Y_{k+1} \mid x, G_k) \quad (1)$$

where $H(\cdot)$ denotes Shannon entropy. By selecting the viewpoint that maximizes information gain, ASP prioritizes locations that are expected to resolve semantic ambiguities in the partially observed scene (e.g., behind a door leading to a kitchen). In this setting, the 3DSG serves not only as the target representation, but also as the structured model that guides exploration.

D. External Cameras as Common Prior Maps

Because the perception framework is strictly visual, observations from external RGB cameras can be processed by MapAnything identically to onboard egocentric images. In this work, we treat such observations as CPMs: a source of semantic and geometric prior information available before robot deployment. When used to initialize the system, this CPM provides partial semantic and geometric structure before robot motion begins, allowing ASP to exploit richer initial context and select more informative viewpoints from the earliest stages of exploration.

IV. Experiments and Results

We evaluate the proposed method in three experimental settings: static pipeline validation, active exploration, and multi-view external camera initialization. Across all settings, node quality is assessed automatically using joint semantic and geometric matching: predicted and ground-truth labels are embedded with a SentenceTransformer [23], and matches are accepted only when both semantic similarity and 3D localization constraints satisfy fixed thresholds. Quantitative evaluation is restricted to nodes, since datasets used in this work do not provide ground-truth relational annotations, while edges are generated by a deterministic geometry-based module that was extensively tested during development.

A. Static RGB-Only 3DSG Pipeline Validation

We first validate the core RGB-only 3DSG generation pipeline in a static setting, independently of both external camera initialization and active exploration. Specifically, we compare the original ConceptGraphs (CG), which uses depth observations and camera poses, with our RGB-only variant (CG-RGB), which replaces them with MapAnything predictions. The goal is to isolate the effect of this substitution and verify that the resulting RGB-only pipeline preserves scene-graph node quality. Evaluation was conducted on the Replica dataset [24], and Table I reports the results.

TABLE I: Quantitative evaluation of scene graph node accuracy across seven Replica scenes (`room0-2`, `office0-3`). CG is the original ConceptGraphs, CG-RGB is the proposed RGB-only variant, with predicted depths and poses. Values are reported as mean $\pm$ standard deviation across scenes.1

Experiment	Precision ↑	Recall ↑	F1-score ↑
CG	$\mathbf{0.686 \pm 0.05}$	0.401 ± 0.10	0.499 ± 0.08
CG-RGB (ours)	0.615 ± 0.07	$\mathbf{0.436 \pm 0.12}$	$\mathbf{0.500 \pm 0.08}$

The results show strong parity between the two variants, with nearly identical F1 scores. MapAnything's noisier depth predictions introduce minor geometric inconsistencies, slightly lowering precision but preventing ConceptGraphs from overly aggressive merging of closely situated objects, thereby improving recall. Overall, these findings indicate that replacing ground-truth depth and pose with MapAnything estimates preserves the quality of the scene graph.

B. Active Exploration with CPM Initialization

The RGB-only active exploration pipeline reported in Fig. 1 was evaluated within the Habitat simulator [25] using the six `apartment` scenes from the ReplicaCAD dataset [26], where local navigation to selected viewpoints is handled via Habitat's built-in navigation utilities. `gemini-2.5-pro` [27] was used as LLM for ASP. Experiments were run from 10 distinct starting positions over 30 exploration steps. We considered two settings: ASP operating from onboard robot observations only, and ASP initialized with one additional fixed external camera observation acting as a CPM prior. Since performance trends were consistent across the six apartments, the results reported in Fig. 2 focus on `apartment 4` as a representative scene.

Operating exclusively on onboard observations, ASP identified approximately 98 object nodes by step 30, corresponding to a recall of about 0.47. When initialized with a single overhead external camera, ASP started from a substantially richer scene graph, increasing the initial node count from approximately 23 to 40 (about +74%) and the initial recall from about 0.15 to 0.26 (about +79%), and maintained this advantage throughout exploration. By step 30, the CPM-initialized setting reached approximately 107 nodes and a recall of about 0.56, with the trajectory remaining consistently above the onboard-only case in both node count and recall. These results indicate that even a *single* external camera can provide a strong semantic and geometric prior, improving scene graph completeness and supporting more effective semantic exploration from the first step.

C. External Camera Infrastructure as Standalone CPM

To quantify the standalone value of camera-infrastructure CPMs, we evaluated the RGB-only 3DSG generation pipeline using only fixed external cameras (one to three),

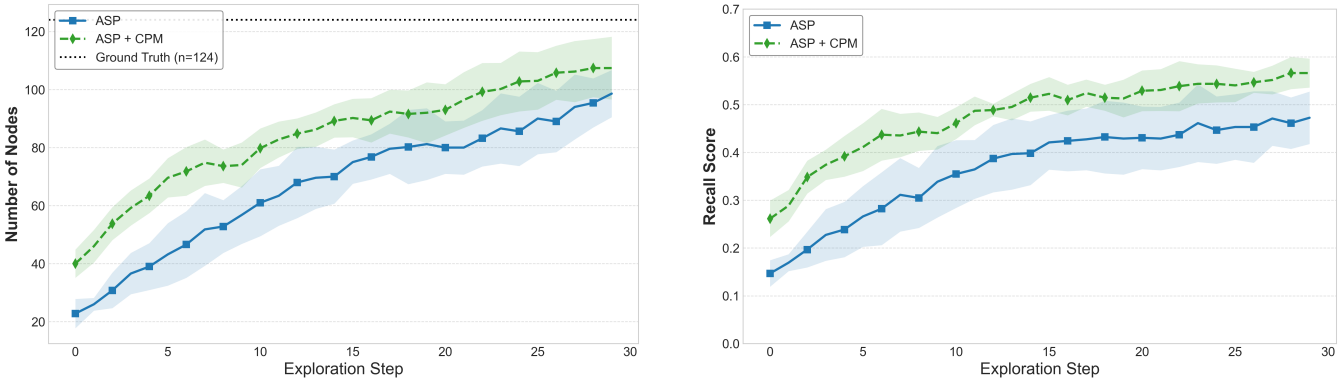

Fig. 2: Evolution of number of predicted nodes (left) and recall (right) for ASP, with and without CPM initialization from a single external camera (+CPM), over 30 exploration steps in `apartment 4`. Shaded areas denote standard deviation over different initial robot poses. The prior provides a strong initialization and sustains improved performance throughout exploration.

without a mobile robot or active exploration. In total, 270 experiments were conducted across 90 ReplicaCAD scenes. The scenes were split into complex `apartments` (dense clutter, ∼123 objects, used also in the previous experiment) and simpler `furnished rooms` (∼25 larger objects). Table II reports the results.

TABLE II: Scene graph accuracy using one to three fixed external cameras (#Cam.) as standalone CPMs on ReplicaCAD `apartments` (Apt.) and `furnished rooms` (Fur.) scenes. No mobile robot is used. Values are mean ± std across scenes.

Scene	#Cam.	Precision ↑	Recall ↑	F1-score ↑
Apt.	1	**0.770 ± 0.054**	0.198 ± 0.030	0.315 ± 0.041
	2	0.741 ± 0.036	0.232 ± 0.030	0.353 ± 0.036
	3	0.698 ± 0.031	**0.301 ± 0.029**	**0.421 ± 0.030**
Fur.	1	**0.566 ± 0.087**	0.398 ± 0.071	0.465 ± 0.073
	2	0.540 ± 0.077	0.487 ± 0.071	0.510 ± 0.067
	3	0.486 ± 0.060	**0.555 ± 0.065**	**0.516 ± 0.055**

While 3 static viewpoints cannot match the recall of a 30-step active robotic exploration, they can successfully establish the environment's dominant structure. In complex `apartments`, three cameras achieved a recall of 0.301, while in `furnished rooms`, the recall reached 0.555. Precision declines slightly as cameras are added due to overlapping multi-view duplicate fragments, but the net F1-score confirms that fixed external RGB cameras can function as a high-quality CPM for semantic-geometric scene initialization, even in the absence of any mobile robot.

V. Conclusion

This paper presented an RGB-only framework for active 3D scene graph construction that uses observations from fixed external cameras as Common Prior Maps. The proposed pipeline builds 3D scene graphs directly from external RGB views, including already deployed surveillance cameras, and uses them as priors for downstream exploration and planning without additional hardware cost. In this way, it advances scene understanding by providing a structured semantic and geometric representation of the environment that can be incrementally refined through subsequent robot observations. In addition, because these cameras provide stable wide-field views, they can also improve the robustness of localization and mapping by anchoring the reconstruction in a consistent global reference frame and by offering broader spatial context across rooms. Experiments show that even a single external camera delivers substantial gains in initial scene coverage and improves the efficiency and quality of subsequent active exploration. Overall, the proposed approach enables robotic systems to leverage widely available camera infrastructure as a powerful source of geometric and semantic priors, opening practical new directions for scalable, infrastructure-assisted robot autonomy.

References

[1] J. Ni, Y. Chen, G. Tang, J. Shi, W. Cao, and P. Shi, "Deep learning-based scene understanding for autonomous robots: A survey," *Intelligence & Robotics*, vol. 3, no. 3, pp. 374–401, 2023.

[2] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5410–5419.

[3] H. Li, G. Zhu, L. Zhang, Y. Jiang, Y. Dang, H. Hou, P. Shen, X. Zhao, S. A. A. Shah, and M. Bennamoun, "Scene graph generation: A comprehensive survey," *Neurocomputing*, vol. 566, p. 127052, 2024.

[4] J. Bae, D. Shin, K. Ko, J. Lee, and U.-H. Kim, "A survey on 3d scene graphs: Definition, generation and application," in *International Conference on Robot Intelligence Technology and Applications*. Springer, 2022, pp. 136–147.

[5] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf, "Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning," *arXiv preprint arXiv:2307.06135*, 2023.

[6] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, "Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[7] H. Jiang, B. Huang, R. Wu, Z. Li, S. Garg, H. Nayyeri, S. Wang, and Y. Li, "Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation," *arXiv preprint arXiv:2402.15487*, 2024.

[8] N. Keetha, N. Müller, J. Schönberger, L. Porzi, Y. Zhang, T. Fischer, A. Knapitsch, D. Zauss, E. Weber, N. Antunes *et al.*, "Mapanything: Universal feed-forward metric 3d reconstruction," *arXiv preprint arXiv:2509.13414*, 2025.

[9] I. Catalano, C. C. Zumaya, J. A. Placed, J. Civera, W. M. Bessa, and J. Peña-Queralta, "3d scene graphs in robotics: A unified representation bridging geometry, semantics, and action," *Authorea Preprints*, 2025.

[10] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3d scene graph construction and optimization," *arXiv preprint arXiv:2201.13360*, 2022.

[11] S.-C. Wu, J. Wald, K. Tateno, N. Navab, and F. Tombari, "Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7515–7525.

[12] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa *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.

[13] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," *Autonomous Robots*, vol. 42, no. 2, pp. 177–196, 2018.

[14] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*. IEEE, 1997, pp. 146–151.

[15] R. Border and J. D. Gammell, "The surface edge explorer (see): A measurement-direct approach to next best view planning," *The International Journal of Robotics Research*, vol. 43, no. 10, pp. 1506–1532, 2024.

[16] H. Ding, X. Liang, Y. Fang, Y. Wu, J. Shi, J. Huo, W. Li, J. Wu, Y.-K. Lai, and Y. Gao, "Sea: Semantic map prediction for active exploration of uncertain areas," *arXiv preprint arXiv:2510.19766*, 2025.

[17] L. Chen, H. Zhan, H. Yin, Y. Xu, and P. Mordohai, "Understanding while exploring: Semantics-driven active mapping," *arXiv preprint arXiv:2506.00225*, 2025.

[18] H. Tang and P. Chaudhari, "Active semantic perception," *arXiv preprint arXiv:2510.05430*, 2025.

[19] L. Robinson, M. Gadd, P. Newman, and D. D. Martini, "Robot-relay: Building-wide, calibration-less visual servoing with learned sensor handover networks," *Autonomous Robots*, vol. 50, no. 1, p. 3, 2026.

[20] D. Buoso, L. Robinson, G. Averta, P. Torr, T. Franzmeyer, and D. De Martini, "Select2plan: Training-free icl-based planning through vqa and memory retrieval," *IEEE Robotics and Automation Letters*, 2025.

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[23] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[24] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[25] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9339–9347.

[26] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. S. Chaplot, O. Maksymets *et al.*, "Habitat 2.0: Training home assistants to rearrange their habitat," *Advances in neural information processing systems*, vol. 34, pp. 251–266, 2021.

[27] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.