# VLA-3D: A Dataset for 3D Semantic Scene Understanding and Navigation

Haochen Zhang, Nader Zantout, Pujith Kachana, Zongyuan Wu, Ji Zhang, Wenshan Wang
Robotics Institute, Carnegie Mellon University
{haochen4, nzantout, pkachana, zongyuaw, zhangji, wenshanw}@andrew.cmu.edu

***Abstract*—With the recent rise of Large Language Models (LLMs), Vision-Language Models (VLMs), and other general foundation models, there is growing potential for multimodal, multi-task embodied agents that can operate in diverse environments given only natural language as input. One such application area is indoor navigation using natural language instructions. However, despite recent progress, this problem remains challenging due to the spatial reasoning and semantic understanding required, particularly in arbitrary scenes that may contain many objects belonging to fine-grained classes. To address this challenge, we curate the largest real-world dataset for Vision and Language-guided Action in 3D Scenes (VLA-3D), consisting of over 11.5K scanned 3D indoor rooms from existing datasets, 23.5M heuristically generated semantic relations between objects, and 9.7M synthetically generated referential statements. Our dataset consists of processed 3D point clouds, semantic object and room annotations, scene graphs, navigable free space annotations, and referential language statements that specifically focus on view-independent spatial relations for disambiguating objects. The goal of these features is to specifically aid the downstream task of navigation, especially on real-world systems where some level of robustness must be guaranteed in an open world of changing scenes and imperfect language. We also aim for this dataset to aid the development of interactive agents that can both respond to commands and ask and answer questions regarding a scene. We benchmark our dataset with current state-of-the-art models to obtain a performance baseline. All code to generate and visualize the dataset is publicly released [1]. With the release of this dataset, we hope to provide a resource for progress in semantic 3D scene understanding that is robust to changes and one which will aid the development of interactive indoor navigation systems.**

## I. Introduction

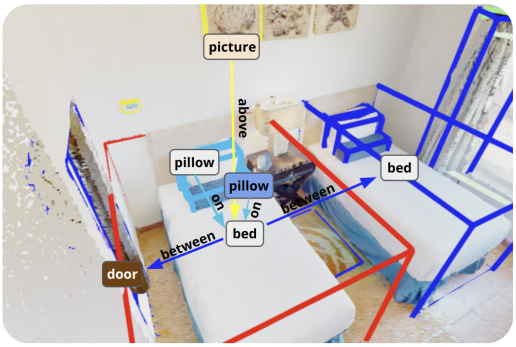
Methods combining vision and language have been evolving rapidly with the advent of both Large Language Models (LLMs) [1, 29, 28] and Vision-Language Models (VLMs) [24, 26, 23] pre-trained on significant amounts of data, tackling various 2D tasks such as Visual Question Answering (VQA) [4], image retrieval [20], and image captioning [24]. As we progress towards generalizable embodied intelligence, there is a need for methods that are capable of reasoning in 3D-space and interacting with humans. Using natural language for example, humans are able to refer to objects in a 3D scene in a way that disambiguates the target object, often using the utterance of "least effort" [33] and making use of relative spatial relationships. An agent that can similarly solve such a problem would be particularly valuable in robotics fields such as indoor-navigation with applications as in-home assistants.

[1]https://github.com/HaochenZ11/VLA-3D

The pursuit of such agents that can identify and understand 3D scenes, consolidate visual input with language semantics, and display robust performance for real-world deployment, however, presents various challenges. First, the scene can have hundreds of objects, contain objects belonging to fine-grained classes, and have many similar objects [25]. Second, human referential language often involves spatial reasoning, affordances, open-vocabulary language, and may even be incorrect or refer to something that does not exist, e.g. *"the remote on the table"* when the remote is actually on the sofa. Third, the scale of available vision-language data in the 3D space pales in comparison to the amount of 2D data, which was crucial to the success of 2D vision-language learning methods [21, 8]. Despite impressive recent advancements with foundation models, such problems remain difficult when applied to robotics as current methods fail to offer the accuracy and robustness needed for real-world deployment [14].

To this end, we propose a novel dataset based on 3D scenes from a diverse set of existing scans of indoor environments that provides a unique resource for training referential object grounding methods. Building on top of the scans, we provide 1) point clouds as they enable learning directly from 3D geometric and visual information [5], 2) extracted object-level attributes and semantic class labels for discriminating and categorizing objects, 3) large-scale scene graphs with spatial relations for a structured representation of a scene, 4) referential language statements to support vision-language grounding using natural language, and 5) traversable free space annotations to explicitly connect to downstream navigation tasks. The inclusion of dense scene graphs and traversable free space are two features that particularly distinguish our dataset from previous object-referential datasets. The scene graphs allow for a robust representation for the semantics in each scene that can be used to guide the grounding task and also to infer when the language statement is invalid. Free space annotations give the ability to generate referential statements that refer not just to objects but to spaces or paths.

Along with our dataset, we also release the code for the entire dataset generation process, demonstrating that synthetic heuristic-based generation methods can aid the efficient generation of large-scale datasets. A custom dataset visualizer tool is also provided to visualize individual scenes and regions from our dataset. With our dataset, we test two state-of-the-art (SOTA) referential object grounding baseline models on our data to verify that such low-level semantic understanding




**(a) Scene with scene graph**

"The white bed that is between the other bed and the door frame"

**(b) Referential statement**

Fig. 1. Sample region from the dataset visualized with (a) a scene graph (a) and (b) a corresponding referential statement

remains a challenging problem and provide a starting point to the identification of where SOTA methods may fail. This can then aid the subsequent development of higher-fidelity 3D vision-language methods that reason over real-world scenes. A sample from our dataset is shown in Figure 1.

## II. Related Work

### *Object Referential Datasets*

The referential object-grounding task has been defined and explored in datasets such as CLEVR [19] in the 2D space and ReferIt3D [2], ScanRefer [9], SceneVerse [17] in the 3D space. While these datasets are similar in the style of their referential statements, the statements are often unintuitive and unnatural compared to human referential statements. E.g, using clock bearings to describe spatial location or using many redundant or subjective descriptors such as "*comfy*" [2]. Both ReferIt3D and ScanRefer are also of a smaller scale and focus only on a single scene data source in which all scenes are single-room, making them less suitable for downstream navigation tasks. SceneVerse scales the data up by curating a much larger dataset and generating statements synthetically, then using an LLM for rephrasing. Despite the increase in scale, the LLM-rephrased statements are often unnatural (e.g. *"the chair stands proudly against the wall"*), and the templates lack explicit references to attributes like size, color, and shape which humans often use for object reference. As a result, models trained on SceneVerse still performed poorly on the Nr3D benchmark [17].

### *Semantic Scene Graph Datasets*

Generating scene graphs from 3D scenes has also been explored in 3DSSG [30], Hydra [15], HOV-SG [31], and ConceptGraphs [12]. 3DSSG focuses on predicting scene graphs automatically, resulting in generated graphs that can miss relations or generate redundant ones. The main use case is scene retrieval from a set of scenes which is different from the navigation paradigm where relations must help disambiguate objects or locations within a single scene. In Hydra, a system is developed to build 3D scene graphs in real-time but does not include explicit language-grounding. While HOV-SG and ConceptGraphs both build open-vocabulary scene graphs, the language-guided navigation task they're designed for involves referring to an object mainly using region references rather than fine-grained inter-object relations.

### *Instruction-Following Datasets*

Multiple works have also explored language-guided navigation through instruction-following statements, often specifying a series of steps to move between regions in a large scene. Common datasets include Room Across Room [22] and Room-2-Room [3]. While these datasets involve spatial and directional references, the task is different from ours as it focuses on a series of distinct steps between rooms rather than explicit fine-grained inter-object spatial relations. The instruction-following task not only requires knowledge of what objects are present, but also exactly where objects and rooms are relative to each other in a multi-room scene, making it difficult to be robust to scene changes or imperfect language.

### *Referential Object Grounding*

A number of papers have explored the task of learning referential object grounding, mainly on either the ReferIt3D benchmark or the ScanRefer task. These include BUTD-DETR [16], MVT [14], ViL3DRel [9], 3D-VisTA [32], and GPS trained on SceneVerse [17]. The best performing method, GPS, however, still only achieves an accuracy of 64.9% on natural language statements [17], which is far below the acceptable threshold for real-world deployment. All of these models are also based on a similar transformer architecture, and with the exception of GPS, cannot handle open-vocabulary object names, which is unideal for a real-world use case. These models' ability to only choose the most likely object from a list also makes them incapable of handling situations where the language input has mistakes or is only partially valid.

## III. VLA-3D Dataset

### *A. Overview*

To aid the development of robust and interactive indoor navigation agents, we introduce a synthetically-generated, publicly released dataset for Vision and Language-guided Action in 3D Scenes (VLA-3D). Our dataset is based on 3D scans from five real-world datasets: ScanNet [10], Matterport3D [7], Habitat-Matterport 3D (HM3D) [25], 3RScan [18], and ARKitScenes [6], as well as scenes generated in Unity. Figure 2 shows a breakdown of the number of regions from each data source. For each scene, we provide:

- Scene point cloud
- List of objects with semantic class labels, bounding box, and color(s)
- List of traversable free spaces
- List of regions with semantic labels and bounding boxes
- Scene graph of spatial relations split by room

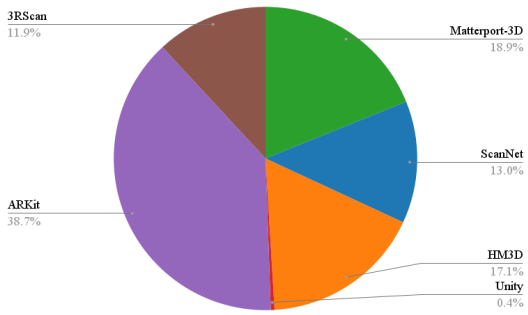



Fig. 2. Breakdown of regions from each data source

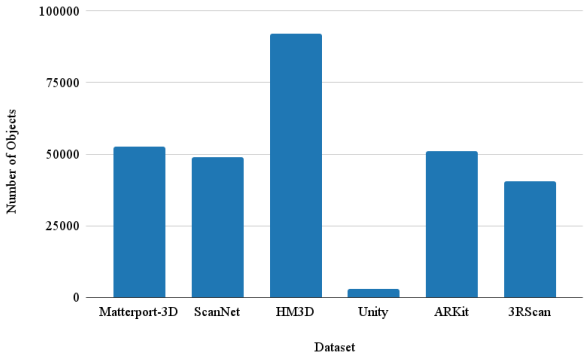



Fig. 3. Total number of objects in each dataset processed

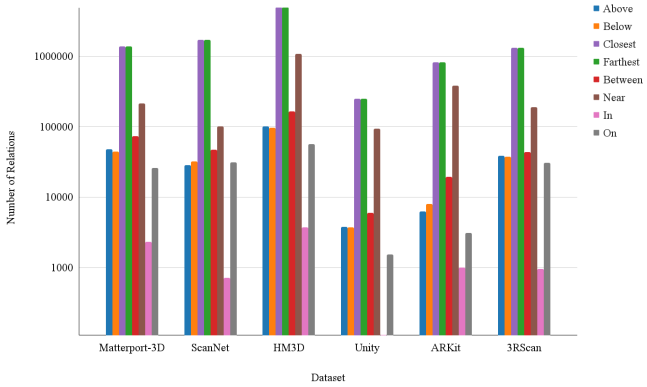



Fig. 4. Total number of each relation type from each dataset processed

TABLE I
SUMMARY OF SEMANTIC RELATIONSHIP TYPES IN VLA-3D

| Relation | Definition | Synonyms | Properties |
|---|---|---|---|
| Above | Target is above the anchor | Over | |
| Below | Target is below the anchor | Under, Beneath, Underneath | |
| Closest | Target is the closest object of a certain class to the anchor | Nearest | Inter-class |
| Farthest | Target is the farthest object from a certain class to the anchor | Most distant, Farthest away | Inter-class |
| Between | Target is between two anchors | In the middle of, In-between | Ternary |
| Near | Target is within a threshold distance of the anchor | Next to, Close to, Adjacent to, Beside | Symmetric |
| In | Target is inside the anchor | Inside, Within | |
| On | Target is above and in contact with the anchor in the Z-axis | On top of | |

- Language statements with ground-truth annotation

Two key features of our dataset are providing large-scale scene graphs for each scene that are robust to scene changes and enables identification of similar objects, as well as incorporating traversable free space as referential targets in addition to just objects. In total, our dataset contains 7635 scenes which contain over 11.5k regions, defined as separate rooms in a scene. A total of over 286k objects from 477 unique classes exist in the dataset, along with 23.5M inter-object spatial relations and 9.7M referential statements. Figure 3 shows the total number of objects in each dataset source while Figure 4 shows the number of each spatial relation generated per dataset.

The data curation process is further detailed below and an overview is shown in 5.

### B. 3D Scan Processing

To generate point cloud files, scene-level point clouds were obtained from the vertices defined in the original PLY files for ScanNet, Matterport3D, and ARKitScenes. For HM3D, Unity, and 3RScan scenes, point clouds were sampled uniformly from the original mesh files while colors were sampled from the textures. Regions and objects were identified leveraging the semantic information in the original meshes. ScanNet, 3RScan, and ARKitScenes each have a single room per scene while region segmentations are provided in Matterport-3D and HM3D, and custom-segmented for Unity scenes. For each object labeled in the scenes, an open-vocabulary class name is stored and the semantic class is mapped to both the NYU40 [13] and NYUv2 [27] schemas with the provided mappings for flexibility [2]. The dominant three colors (if any) were obtained for each object based on the object point cloud and a color clustering algorithm.

To provide extra navigation targets, each scan was also processed to generate the horizontally traversable free space. Separate traversable regions in a room are chunked into subregions, for which spatial relations with other objects in the scene are generated to create unambiguous references to these spaces (e.g. “the space near the table”).

### C. Scene Graph Generation

Eight different types of semantic spatial relations were calculated using heuristics based on the yawed object bounding boxes to generate a scene graph of relations. Relations are generated exhaustively for every pair or triple of objects within a region, then filtered afterwards depending on the semantic classes involved. All relations are binary except for the “*between*” relation, which is ternary.

Table I defines the types of spatial relations used.

[2]For the Unity scenes, the ground-truth semantic labels were cleaned then manually mapped to the class schemas by five data annotators. A validation round was done to standardize the labels.

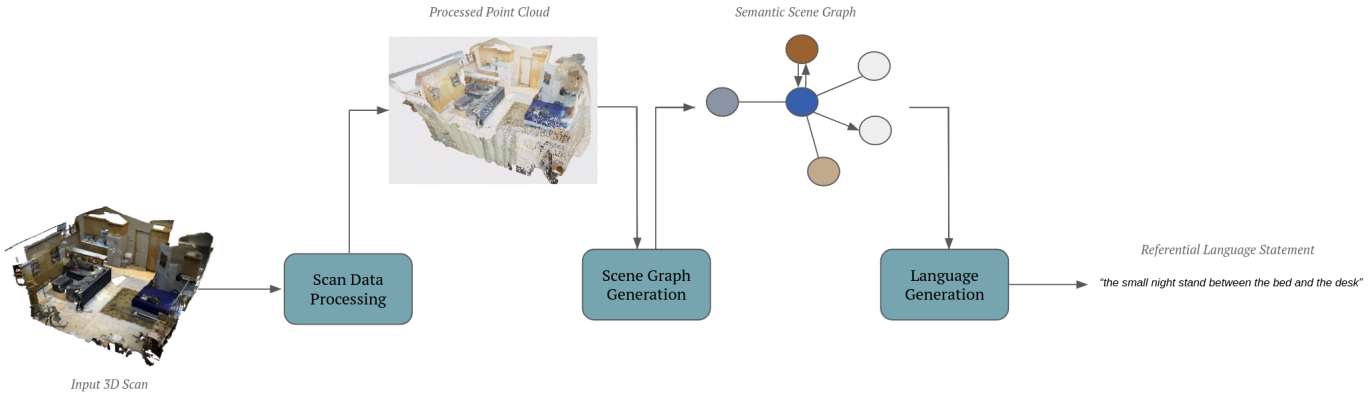



Fig. 5. Data processing pipeline consisting of: 3D Scan Processing, Scene Graph Generation, and Language Generation

TABLE II
REFERENTIAL OBJECT GROUNDING ACCURACY OF TWO BASELINE MODELS ON VLA-3D, NR3D, AND SR3D

| Baseline Model | VLA-3D | Nr3D | Sr3D |
|---|---|---|---|
| MVT | 22.5% | 59.5% | 64.5% |
| 3D-VisTA | 28.9% | 64.2% | 76.4% |

### D. Language Generation

Referential language statements were synthetically generated based on the computed scene graph using a template-based generation method. From the table above, synonyms for each relation are used to add variety into the statements. Every statement has at least one semantic relation and only uses object attributes if needed to distinguish the target object. The generated statements are also:

1) **View-independent**: The relation predicate for the target object does not depend on the perspective from which the scene is viewed from.
2) **Unambiguous**: Only one possibility exists in the region for the referred target object.
3) **Minimal**: Following Grice's maxim of manner [11], statements use the least possible descriptors to disambiguate the target object.

## IV. BASELINE EVALUATION

To verify the difficulty of our dataset, we evaluate the pre-trained checkpoints of two SOTA open-source baseline models on our data directly: MVT and 3D-VisTA. MVT is the best-performing method on the official ReferIt3D benchmark while 3D-VisTA is a more recent method that has since outperformed MVT. The test results are shown in Table II. The test performances on both Nr3D and Sr3D (which the models are trained on) are also shown as a point of comparison.

The results of both models are much lower on our dataset compared to their performance on the ReferIt3D benchmark, likely due to the fact that they are directly evaluated not just on new language data but also on unseen scenes with many more fine-grained objects than what they were trained on. Upon examining failure cases, we observe that failures are either due to object classification errors, language semantic reasoning errors (e.g. mixing up target and anchor object), or errors in spatial reasoning (e.g. choosing "distractor objects" of the same semantic class but incorrect spatial relation). This disconnect in performance indicates the poor cross-domain generalizability of existing methods, especially to complex real-world scenes, and delineates the need for more diverse language data to improve 3D visual grounding models and enable their use in more complex tasks like interactive indoor navigation. It also verifies VLA-3D as a challenging benchmark for progress towards this goal.

## V. CONCLUSION

Aiming to advance progress in semantic scene reasoning and understanding in robotics applications, we introduce a large-scale novel dataset of object-referential natural language statements along with spatial scene graphs for a diverse set of 3D scenes. This dataset contains a variety of spatial relationships and language statements on the scale of millions and is suited for the sub-task of referential object grounding guided by structured scene representations. Future extensions to VLA-3D include augmenting the statements with LLMs, adding 3D scan data from other real-world sources, generating compound relational statements, generating view-dependent statements, and extending the statements beyond referential object-grounding to include the action component explicitly. Further research using this dataset for training could involve the development of generalizable system-integrated modules with the capabilities of answering questions about the scene, identifying items not in the scene, and suggesting alternative objects with similar attributes, location, or affordances. Overall, our dataset establishes a resource for the development of generalizable methods that extract observations from 3D scenes and reason about them using open-vocabulary natural language, which aids the development of interactive indoor navigation agents that can operate in changing environments, both alongside and with humans.

## References


[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 422–440. Springer, 2020.

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016.

[6] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.

[9] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020.

[10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[11] Richard E. Grandy and Richard Warner. Paul Grice. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2023 edition, 2023.

[12] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.16650*, 2023.

[13] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 564–571, 2013.

[14] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15524–15533, 2022.

[15] Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. *arXiv preprint arXiv:2201.13360*, 2022.

[16] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pages 417–433. Springer, 2022.

[17] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. *arXiv preprint arXiv:2401.09340*, 2024.

[18] Nassir Navab Federico Tombari Matthias Niessner Johanna Wald, Armen Avetisyan. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, 2019.

[19] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[20] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27, 2014.

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[22] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*, 2020.

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[25] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

[26] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[27] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012.

[28] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[29] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[30] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2020.

[31] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram 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.

[32] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023.

[33] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.