

Spatial Q&A: Leveraging and Benchmarking LLM-based Solutions

Master Thesis

Emilia Szymańska Robotics, Systems and Control Department of Mechanical and Process Engineering D-MAVT

Advisors:Dr. Mihai Dusmanu, Dr. Mahdi Rad, Dr. Jan-Willem BuurlageSupervisor:Prof. Dr. Marc Pollefeys
Microsoft Zürich
Computer Vision and Geometry Group, ETH Zürich

September 15, 2024

Abstract

Answering questions about the spatial properties of the environment poses challenges for existing language and vision foundation models due to a lack of understanding of the 3D world notably in terms of relationships between objects. To push the field forward, multiple 3D Q&A datasets were proposed which, overall, provide a variety of questions, but they individually focus on particular aspects of 3D reasoning or are limited in terms of data modalities. To address it, this thesis presents Space3D-Bench – a collection of 1000 general spatial questions and answers related to scenes of the Replica dataset which offers a variety of data modalities: point clouds, posed RGB-D images, navigation meshes and 3D object detections. To ensure that the questions cover a wide range of 3D objectives, we propose an indoor spatial questions taxonomy inspired by geographic information systems and use it to balance the dataset. Moreover, we provide an assessment system that grades natural language responses based on predefined ground-truth answers by leveraging a Vision Language Model's comprehension of both text and images to compare the responses with ground-truth textual information or relevant visual data. Finally, we introduce a baseline called RAG3D-Chat integrating the world understanding of foundation models with Retrieval Augmented Generation, achieving an accuracy of 67% on the proposed dataset.

Acknowledgements

I would like to express my heartfelt gratitude to all members of the Microsoft Zurich lab, led by Prof. Marc Pollefeys, for making these six months of thesis work highly enjoyable. I hope you will successfully keep SAILing far into the yet undiscovered ocean of research. I want to thank my supervisors, especially Mihai Dusmanu and Mahdi Rad, for their outstanding overseeing of this Master thesis project – thank you for all the support, prolonged weekly meetings, and overtime work on both the thesis and the paper. It would not have been possible without you!

Contents

1	Intr	oductio	1	1
2	Rela 2.1 2.2	ited Wo Spatial Spatial	rk Q&A Benchmarks and Datasets	3 3 4
3	Met	hod		7
	3.1	Benchi	nark	7
		3.1.1	Dataset	7
		3.1.2	Question Taxonomy	9
		3.1.3	Replica Dataset Preprocessing	12
		3.1.4	Automatic Assessment	16
	3.2	RAG3	D-Chat	18
		3.2.1	Image Module	18
		3.2.2	Text Module	19
		3.2.3	Navigation Module	20
		3.2.4	SQL Module	21
		3.2.5	Semantic Kernel	23
4	Exp	eriment	S	25
	4.1	Evalua	tion of the Automatic Assessment	25
		4.1.1	User Study	25
		4.1.2	LLM Replacement in the Automatic Assessment System	26
	4.2	RAG3	D-Chat	27
		4.2.1	Evaluation on Space3D-Bench	27
		4.2.2	Ablation Study	29
5	Disc	ussion		31
6	Con	clusion	:	33
A	Resi	ults of th	ne User Study	35
B	Auto	omatic A	Assessment Prompts	51
С	Bias	Study		57
D	Mod	lificatio	ns to the Replica Dataset	59
				- 1

Ι

Е	Semantic K	ernel Modules' Descriptions	61
	E.0.1	Image Module	61
	E.0.2	Text Module	61
	E.0.3	Navigation Module	61
	E.0.4	SQL Module	62

List of Figures

1.1	Questions from Space3D-Bench with answers generated by RAG3D-Chat	2
3.1	Distribution of detected objects across selected scenes in Replica	8
3.2	Question distribution based on their lengths	8
3.3	Statistics of questions in Space3D-Bench	9
3.4	Navigation mesh overlaid on the apartment	13
3.5	Object-to-room assignments	14
3.6	An issue with the semantic segmentation of Replica's point clouds when generating object	
	descriptions	15
3.7	Camera positions for saving images of a point cloud	15
3.8	Example of an image passed to the VLM when generating object descriptions	16
3.9	Automatic assessment procedure	17
3.10	Overview of RAG3D-Chat	18
3.11	Overview of the Image Module	19
3.12	Overview of the Text Module	20
3.13	Overview of the Navigation Module	21
3.14	Overview of the SQL Module	22
3.15	Example of modules chained together by the planner	23
4.1	Results of the user study with respect to the GPT4-based assessment system	25
4.2	Results of the user study with respect to the Phi3-based assessment system	26
4.3	Results of the baseline on each category of questions in Space3D-Bench	27
4.4	Analysis of the plans generated by SK	28
4.5	The longest plan generated by the SK planner	28
C.1	The apartment considered in the bias study	57

List of Tables

2.1	Comparison of the datasets most similar to Space3D-Bench	4
3.1	SQL table schema for <i>detected_objects</i> table	21
3.2	SQL table schema for <i>rooms</i> table	22
4.1	Results of the ablation study	29
B .1	Acceptance criteria for location-related answers	52
B.2	Acceptance criteria for pattern-related answers	52
B.3	Acceptance criteria for measurement-related answers	53
B. 4	Acceptance criteria for relation-related answers	54
B.5	Acceptance criteria for prediction-related answers	55
B.6	Acceptance criteria for path-related answers	56
D.1	Changes of class names in multi-room Replica scenes	59
D.2	Changes of class names in one-room Replica scenes	60

Chapter 1

Introduction

Foundation models are becoming a standard tool in the field of artificial intelligence by providing adaptability and a high level of performance in a variety of down-stream tasks. Notably, the field of 3D scene understanding has experienced rapid advancements fueled by these large pre-trained models. More and more applications are emerging in terms of 3D reasoning, spatial awareness, and environment interaction [21]. These spatial competences have provided significant improvements for mixed reality [12, 37, 4], robotics [39, 5, 7], autonomous vehicles [23, 24, 9], inclusive technologies [17], or navigation [48, 2, 13].

Numerous approaches have been proposed by the research community to address 3D tasks. As these foundation models proved to be capable of understanding other modalities than text, one group of solutions involves proposing multi-modal models, integrating images [14], videos [47], or 3D data such as point clouds or meshes [40]. Another strong trend in the field is combining the strengths of existing models with such tools as context retrieval [28] or zero-shot learning [46, 45]. The progress in understanding the spatial properties is evident, however, robustness and alignment still remain a challenge.

To measure the performance of these spatially-aware systems, multiple 3D Question and Answer (Q&A) datasets have been proposed [3, 50, 22, 8, 42, 19]. They vary in their assessment objective, size, scene types, and provided data. Although, from the holistic perspective, the datasets provide a large variety, individually they are either limited in terms of accompanying data modalities, focus on a narrow aspect of 3D reasoning, or do not have a balanced question distribution in terms of objectives.

To address these limitations, we present a dataset composed of 1000 questions with ground truth answers. To assure that the questions cover a wide range of 3D objectives, we adapt an existing taxonomy of spatial question [26] used in Geographic Information Systems and adjust it to the indoor scenes scenario. We balance the number of questions with respect to the presented categories. The questions are associated with thirteen selected scenes from the Replica dataset [36], that gives access to a variety of data, such as 3D object detections, navigation meshes, and point clouds. Additionally, as Replica is integrated into Habitat Sim's environment [35], data such as videos, and RGB-D or semantically-segmented images with camera poses can be seamlessly extracted.

To complement the dataset's functionality, we develop an automatic assessment system based on a Vision Language Model (VLM) that evaluates the responses from a question answering system against the dataset's ground truth. To establish the assessment's correctness and reliability, we conduct an extensive user study of 60 participants, on a subset of 40 questions that are randomly sampled. As the result of the survey shows, our evaluation system agrees on 97.5% of the cases with users, which confirms the reliability of our evaluation protocol.

To demonstrate a baseline performance on the created dataset, we propose RAG3D-Chat – Retrieval Augmented Generation (RAG) for 3D Chat – a system that utilizes RAG [18] and VLMs to identify the relevant scene context from images, texts, and an SQL database, and also has the capability to answer questions



Figure 1.1: Questions from Space3D-Bench with answers generated by RAG3D-Chat. The dataset supports a variety of spatial tasks, including object location, measurements, pattern identification, navigation, spatial relationships, and predictions.

regarding navigable distances. We employe a planner based on a Large Language Model (LLM) to chain available functionalities of the system to answer complex questions. This system scores 67% of accuracy on the dataset, which proves that there is room for improvement in the robustness of 3D spatial Q&A. Selected questions from Space3D-Bench dataset, answered by RAG3D-Chat, are presented in Fig. 1.1.

In summary, this paper introduces the following contributions:

- We propose a dataset of 1000 diverse spatial questions and answers, based on the scenes of the Replica dataset, which offer a variety of data modalities. We present an application of a geographic spatial questions taxonomy to indoor scenes, and balance our questions accordingly.
- We provide a VLM-based assessment system that evaluates natural language responses given ground truth answers. To confirm the reliability of our proposed system, we conducted an extensive user study.
- We leverage the strengths of foundation models, and combine them with Retrieval-Augmented Generation, to present a baseline achieving 67% of accuracy on the proposed dataset.
- We release the dataset with the assessment system, to encourage the research community to address the challenges of 3D question answering by developing and evaluating their spatial Q&A systems.

We provide the project's webpage with linked repositories containing the source codes of RAG3D-Chat and Space3D-Bench under the following link: https://space3d-bench.github.io/.

Chapter 2

Related Work

2.1 Spatial Q&A Benchmarks and Datasets

A variety of spatial Q&A datasets with associated benchmarks have been constructed to tackle different aspects of spatial question answering. These datasets collectively offer diverse modalities. SpartQA [27], for example, provides textual stories describing scenes with 2D geometrical figures, based on which questions regarding spatial relationships are asked. However, since the research community extended the applications of foundation models beyond text, more complex, real-life scenes were included into Q&As. For example, multi-view images paired with questions in 3DMV-VQA [15] allow for evaluation of a system's abilities with respect to object counting and existence, relations and comparisons. One of the limitations of this dataset is the fact that in HM3DSEM [41], on top of which 3DMV-VQA is built, very close objects tend to be annotated as one semantic instance. A similar nature of problems is present in the Replica dataset, therefore curated object detection are provided with this project.

M3DBench [19] interleaves modalities in instruction-response pairs, combining texts, coordinates, images, and 3D objects, and thus offering a promising benchmark for general multi-task systems. However, within the question-answering-related tasks no specific taxonomy of questions was mentioned with respect to which their generation was balanced. Additionally, as for the beginning of September 2024, M3DBench dataset has not been yet released, although the paper and the repository for the dataset were made public in December 2023.

ScanNet [10], giving access to posed RGB-D image sequences, surface reconstructions and instancelevel semantic segmentations, has been selected as a source of indoor scene context for some Q&A datasets. SQA-3D [22] based its question on egocentric situation awareness within ScanNet's environments, thereby focusing on one category of QA tasks. Azuma's *et al.* ScanQA [3] introduced a task – with a corresponding dataset – of combining the answer on 3D scans with 3D bounding boxes. Simultaneously-developed Ye's *et al.* ScanQA [43] leveraged human annotations to correlate the scenes with questions and free-form answers. Both ScanQA datasets form a complementary set of questions, however, the prediction- and distance-related questions are missing.

3RScan [38] similarly to ScanNet offers posed RGB-D sequences, instance-level semantic segmentation, object alignment and 3D meshes for indoor spaces, other datasets were also based upon it. One example of such is CLEVR3D [42], covering Q&A in the aspects of objects' attributes and their spatial relationships. In this case as well prediction-focused instructions were not present.

Some benchmarks aim to provide an extensive evaluation on approaching specific tasks by combining various datasets. LAMM [44] integrates 3 point-cloud-related datasets to provide a comprehensive benchmark for 3D tasks, including 3D question answering. LV3D [9], on the other hand, focuses on multi-turn Q&A, fusing fifteen 2D and 3D object recognition datasets. Both datasets have significant sizes, which were

											5
									nes	on all in	Ш°.
						~	স	è .	Hest. N	COL CHEST	restil
				న	ot .	inas	. clone	ation	. at 511	tion red	÷.
datasat nama	Hinstructions	gonaration	.ndc	U, JIC	50 0C	5 JOIL	it with	Se Isi	Mr. red	ile distali	
uataset name	#Instructions	generation	v	0	•	×,	v	y .	₹.		
3DMV-VQA [15]	50K	automatic	\checkmark	X	\checkmark	\checkmark	X	X	X	X	
M3DBench [19]	320K	automatic	\checkmark	X	\checkmark	\checkmark	X	1	\checkmark	X	
SQA-3D [22]	33.4K	human	✓	X	1	\checkmark	X	1	1	X	
ScanQA Azuma et el. [3]	41K	automatic	1	X	1	1	X	1	X	X	
ScanQA Ye et el. [43]	10K	human	1	X	1	1	X	1	X	X	
CLEVR3D [42]	171K	automatic	1	X	1	\checkmark	X	1	X	X	
LAMM [44]	186K	automatic	1	\checkmark	1	\checkmark	X	X	X	X	
LV3D [9]	40.9M	automatic	1	1	1	1	X	X	1	×	
NuScenes-QA [33]	460K	automatic	X	1	1	1	X	X	X	X	
Space3D-Bench (ours)	1K	human	1	X	✓	✓	1	✓	1	\checkmark	

Table 2.1: **Comparison of the datasets most similar to Space3D-Bench.** They are compared with respect to the number of available instructions (questions), method of instruction generation, type of considered scenes (outdoor, indoor), availability of different data modalities (RGB images, point clouds, navigation meshes), compatibility with Habitat Sim, and the existence of prediction- and distance-related questions.

possible to be achieved due to the automatic generation of instruction-response pairs. This approach has a major limitation – at times the resulting instruction can lack coherence or be ambiguous.

Not only indoor spaces are addressed in spatial Q&As. NuScenes-QA [33], for instance, addresses visual question answering in autonomous driving by proposing question-answer pairs based upon outdoor environments from nuScenes [6]. However, in this thesis focuses on rooms and apartments, and the dynamic, outdoor scenes are to be explored in the future research.

The differences between the datasets having similar available modalities to the proposed Space3D-Bench are presented in Table 2.1. Although Space3D-Bench is smaller in size, it offers a human-formulated Q&A, balanced among categories suggested in spatial geography research. As it is based upon Replica, which is compatible with Habitat Sim, additional data collection is possible when needed by the dataset users. That leaves the door open for future approaches in terms of question answering systems, not limiting them to the default Replica's modalities.

2.2 Spatial Questions Answering

A multitude of approaches have been experimented with in order to solve spatial Q&A challenge. One such method is 3D Concept Learning and Reasoning (3D-CLR) [15], which combines neural fields with 2D-pretrained vision-language models and neural reasoning operators to answer questions based on multi-view images of a scene. However, as it relies solely on view-constrained images without incorporating additional object detection data, it struggles with questions involving small objects and spatial relations.

Azuma *et al.* introduced a different approach, which employs a fused descriptor, ScanQA [3], that links language expressions to the 3D scan's geometric features, enabling the regression of 3D bounding boxes to determine the objects described in the questions. Despite this, the descriptor was neither trained nor tested on tasks such as prediction, pattern identification, or distance estimation, leaving its applicability to these

areas unclear.

In another method, 3DQA-TR [43], a language tokenizer embeds questions with two encoders extracting appearance and geometry information, to finally fuse modalities with 3D-L BERT to answer a question. However, the scarcity of human-related data in its training made it less effective for queries involving people.

Another integration of BERT is present in GPT4Point [32], where a Point-Q-Former [11] aligns point-text feature, later to be analysed by a Language Model enhancing the model's ability to infer text. While this method yielded promising results, its primary application was limited to reasoning about single objects rather than interpreting full scenes.

On the other hand, 3D-VisTA [50] leverages self-supervised pre-training via masked language/object modeling to effectively learn the alignment between texts and point clouds. In terms of 3D question answering, it was tested on ScanQA of Azuma *et al.*, hence not providing evaluation on prediction- and distance-related questions.

LAMM [44], a multi-modality language model, encodes each modality by a corresponding pre-trained encoder, followed by a trainable projection layer and LoRA parameters [16], to eventually be decoded by a shared LLM. However, this approach struggled with counting tasks involving a large number of objects.

PointLLM [40] as well uses a pre-trained encoder, although only on point clouds, whose extracted features are used by a pre-trained LLM for reasoning and generating responses. Like GPT4Point, its primary focus was single-object reasoning, with limited application to full scene interpretation.

An intriguing integration of scene graphs can be found in TransVQA3D [42], which applies a crossmodal Transformer to fuse the features of language and object. This approach then incorporates scene graph initialization and performs scene graph-aware attention, excelling at pattern identification and spatial relationship tasks, though its applicability to measurement and distance-based questions remains unproven.

NuScenes-QA baseline framework [33] processes multi-view images and point clouds to obtain Bird's-Eye-View features, crops objects embeddings (based on the detected 3D bounding boxes), and forwards these features to a transformer-based Q&A model. However, since the training set included only outdoor scenes, this method is specifically tailored for autonomous driving applications.

Cube-LLM [9] applies changes to a multi-modal LLM by replacing a visual encoder, finetuning it on specific datasets, and using different resolutions and normalization techniques for inputs to enhance performance in 3D-related tasks. Despite its versatility, Cube-LLM was fine-tuned primarily for outdoor scenes related to autonomous driving.

Finally, M3DBench's baseline [19] utilizes a scene perceiver to extract scene tokens from 3D visual input, encodes multi-modal instructions into instruction tokens, which are then concatenated and fed into a frozen LLM, which generates the corresponding responses subsequently. However, similar to other methods, M3DBench did not include predictions or navigable distances in its training or testing.

The performances of the described systems vary, depending on the end-task and available modalities. This thesis proposes another family of solutions for spatial problems, which in contrary to all the presented state-of-the-art solutions does not need any prior training, therefore avoiding the costs associated with GPU usage. An approach based on Retrieval-Augmented Generation is presented, and its effectiveness is evaluated on the created dataset. To the best of our knowledge, spatial question answering for indoor spaces has not been addressed with a RAG-based method yet.

Chapter 3

Method

3.1 Benchmark

This thesis introduces Space3D-Bench, a dataset composed of 1000 spatial questions with ground truth answers for the Replica dataset scenes [36]. Replica's object detections are improved by correcting class name labelling, removing irrelevant objects, adding object-to-room assignments, and moving coordinates from Habitat Sim's to Replica's coordinate system. Data on rooms' centers and sizes is provided, as well as framework-independent navigation meshes are cleared of artifacts. To automatically assess the correctness of the answers, an evaluation system based on both text and vision language models is proposed. The following sections detail the dataset, the taxonomy of its questions, preprocessing performed on the Replica dataset, and the automatic evaluation system.

3.1.1 Dataset

The Q&A dataset is based on thirteen scenes being a part of Replica [36], which is a dataset of 18 highly photo-realistic 3D indoor scene reconstructions. It includes dense 3D point clouds, high-dynamic-range textures, semantic information, object detections, and navigation meshes. Due to its compatibility with Habitat Sim [35], it is possible to easily extract additional data such as RGB-D images, videos or other sensor measurements, making it a suitable environment for development of AI agents. Despite Replica's limited number of scenes, the diverse range of available modalities played a key role in the dataset selection process.

The selected Replica environments include six multi-room scenes – a 2-floor house (*apartment 0*), three multi-room apartments (*apartment 1, 2, hotel 0*), two different setups of the Facebook Reality Lab (FRL) apartment (*FRL apartment 0, 1*); and seven single-room scenes – three apartment rooms (*room 0, 1, 2*), and four office rooms (*office 0, 2, 3, 4*). As all FRL apartments vary only in the object distribution, it was deemed unnecessary to use more than two similar scenes. Questions for *office 1* were not created due to the small size of the room, the lack of visible details, and the overexposure of available frames.

The distribution of detected objects with respect to their classes and corresponding scenes is presented in Fig. 3.1. Certain objects, like ceiling lamps, wall plugs, books, or blinds, frequently appear in scenes and can be challenging for the system to distinguish due to their intra-class visual similarities. As a result, questions that would require identifying specific instances of these objects were not included in the dataset.

Space3D-Bench contains 100 questions with answers for all multi-room scenes, 60 questions for apartment rooms and two offices, and 50 questions for the remaining two offices, summing up to a total of 1000 questions with ground-truth answers. These answers may have one of the two forms: ground truth information for factual data, such as the number of objects in a room, or an illustrative image of the objects / rooms of interest for questions that involve descriptions, predictions or identifying similarities. This distinction



Figure 3.1: **Distribution of detected objects across selected scenes in Replica.** Only the object classes that appear more than 6 times are included in this figure.

allows to not penalize the answering system's creativity, as the answer in the image-based evaluation does not need to match a specific language-limited pattern, but just needs to correspond to the scene state in the image.

Following the taxonomy described in Sec. 3.1.2, both questions and answers were manually crafted. Their generation was not automated to limit ambiguities and to ensure the suitability of questions to the scene. To further validate the developed questions, a set of 20 questions related to a specific scene was presented to approximately 50 Microsoft team members. These questions along with the most relevant comments are attached in Appendix C. The participants' feedback was used to refine and finalize the dataset. For example, the questions regarding the *neighboring* rooms were reformulated to instead use the word *accessible*, and the queries asking for the room with least number of objects were removed, as it was deemed to be ambiguous.



Figure 3.2: **Question distribution based on their lengths.** The questions are overall concise with an average length of around 8 words, but some longer ones are also present.



(a) Question distribution based on their first (b) Question distribution based on their catethree words. gories.

Figure 3.3: **Statistics of questions in Space3D-Bench.** The dataset has a large variety of phrasings (a) and is well distributed across the question categories (b).

The distribution of questions with respect to the first words used to formulate them and their assigned categories is presented in Fig. 3.3, while the variety in the lengths of questions is depicted in Fig. 3.2. In the distribution of questions, pattern- and distance-related categories are grouped together, as particularly in single-room scenarios relevant data is limited.

3.1.2 Question Taxonomy

To achieve a balanced distribution of questions with respect to their types and functions, the taxonomy of spatial questions primarily proposed for Geographic Information Systems by M. Schmidts and N. Giner [26] was investigated. Based on that, a taxonomy of spatial questions for rooms- and objects-related applications in indoor spaces is proposed, whose detailed categories descriptions follow in the corresponding sections.

Location. These questions focus on understanding where objects are located, either with respect to the coordinate system or to the associated rooms. To this category belong the following types of questions, accompanied by the objective of the question and an example:

- *Where can you find X?* identifying where specific objects are located, either with respect to the coordinate system or to the associated rooms. Example: "Where can you find chairs?";
- *What is the 3D position of X?* determining the location of objects with respect to the coordinate system. Example: "What is the 3D position of the dining table?";
- *Which rooms have X*? identifying the rooms where specific objects are located. Example: "Which rooms have beds?";
- *Are there any X in Y?* defining whether the specified objects are located in the given room (or a set of rooms). Examples: "Are there any sofas in the living room?", "Are there any tables in the ground floor?";

• *Which rooms have no X?* – determining the locations of the specified objects, to later compare it with all available rooms. Example: "Which rooms have no plants?".

Measurement. The questions in this category concern providing the information related to the size, shape and distribution of both individual objects and the rooms. This category includes:

- *Which room / object is the biggest / smallest?* measuring and comparing rooms' or objects' dimension. Example: "Which room is the smallest?", "Which rooms are two biggest ones?", "What is the largest furniture?";
- List all the rooms / objects, sorted by their area/volume in an ascending / descending order. measuring and comparing rooms' or objects' dimensions to later sort them in a specific manner. Example: "List all the rooms, sorted by their volume in a descending order.";
- *How many X are there in Y?* measuring the number of objects in a given room or in a set of rooms. Example: "How many office chairs are there in the study?", "How many standing lamps are there in all bedrooms?";
- *List all the rooms with the corresponding number of X in them.* measuring the number of objects in all the rooms. Example: "List all the rooms with the corresponding number of chairs in them.";
- *Which room has the most / least X?* measuring the number of objects in all the rooms to later compare them. Example: "Which room had the most sofas?";
- *What is the length / width / height of X?* measuring a specific dimension of an object or a room. Example: "What is the height of the living room?", "What is the length of the dining table?";
- *What is the area / volume of X*? measuring the dimensions of an object / room to calculate its area or volume. Example: "What is the area occupied by the rug?", "What is the volume of the dining room?".

Relation. The aim of these questions is to specify the spatial relationships between objects and rooms, summarize what is contained within given areas, and determine what is closest and nearby. Although the original GIS taxonomy proposed to also include the definition of what is visible from a given location, in this study's case it fits more the prediction category, as also a specification of a person's pose is included and based on that a prediction is made. Instead, this questions' type is replaced with requests on the descriptions of room's objects, to transfer the original taxonomy's idea on determining what can be seen in the room.

The spatial relationships determinations can have the following forms:

- Which objects are within X meters from Y? defining which objects are contained within a specific area from a given object. Example: "Which objects are within 2 meters from the sofa?";
- *Which object in X is closest to Y*? determining which object is closest to a given object. The question can specify which subset of objects (*e.g.* corresponding to a particular room) should be considered. Example: "Which object is closest to the bed?", "Which object in the living room is closest to the bed?";
- What objects are there on / under / behind / ... X? determining objects having a specific spatial relationship with a given object. Example: "What objects are there on the desk?";
- *How are X and Y positioned with respect to each other?* identifying the spatial relationship between two given objects. Example: "How are the armchair and the sofa positioned with respect to each other?";

- *How is X placed with respect to the surrounding objects?* identifying spatial relationships between a given object and its surrounding. Example: "How is the sofa placed with respect to the surrounding objects?";
- *Describe the objects in X.* providing descriptions of available objects and their spatial relationships in the given area. Example: "Describe the objects in the corridor.";
- *Which rooms are directly accessible from X?* defining the relationships between the given room and other rooms. Example: "Which rooms are directly accessible from the kitchen?".

Navigation. This group is associated with the objective of finding the shortest navigable and straight-line paths (or more specifically, their distances) in the apartment between given rooms or objects.

The questions have the following structure:

- What is the distance between X and Y (considering the obstacles)? finding the shortest navigable distance between objects or rooms. Example: "What is the distance between the sofa and the fridge (considering the obstacles)?";
- *What is the distance in straight line between X and Y*? finding the Euclidean distance between two objects or rooms. Example: "What is the distance in straight line between the cabinet and the mirror?";
- What distance would you have to walk to get from X to Y? determining the shortest navigable distance between objects or rooms. Example: "What distance would you have to walk to get from the kitchen to the dining room?".

Patterns. Another aspect of spatial questions covers similarities and patterns identification. Grouping objects or rooms with common features, recognizing spatial and visual trends, and determining consistent or uniform distribution of objects are just some of the topics tackled by the questions in this category. Additionally, this category is extended with the questions on the wall colors, as it requires the feature reasoning based on components (walls) located in different parts of the room or of the apartment.

The questions consist of the following types, accompanied by their objective:

- Which rooms have the same number of X? counting the number of given objects in all rooms and grouping the rooms based on the given distribution. Example: "Which rooms have the same number of chairs?";
- *Is there a consistent X across Y*? determining whether there is a pattern across a set of rooms with respect to a given factor. Example: "Is there a consistent wall coloring across the apartment?", "Is there a consistent bed distribution across the bedrooms?";
- *What are the similarities between X and Y?, What do X and Y have in common?* detecting similarities between objects or rooms. Example: "What are the similarities between the two sofas in the living room?", "What do the study and the dining room have in common?";
- What colors are the walls in X? identifying the walls and their colors in the given room or a set of rooms. Example: "What colors are the walls in the kitchen?", "What colors are the walls in the ground floor?".

Predictions. This category of questions requires making predictions based on the room sizes, and objects' presence and layout. The types of predictions vary, depending on the steps contributing to formulating the answer:

- *Where could you X?* identifying the rooms having the capacity and containing suitable objects for a specific activity. Example: "Where could you have a casual meeting with three friends?";
- *How many people could X?* identifying the room's or set of rooms' capacity (in terms of both the size, available objects or their layout) for a specific activity. Example: "How many people could comfortably sit in the dining room?"
- *What can a person X see in front of them?* predicting what is visible for a person with a specific pose. Example: "What can a person sitting in the chair in the bedroom see in front of them?", "What can a person with their back facing the bed see in front of them?";
- *Which room is best adjusted for X?* identifying a room having the best capacity (in terms of both the size, available objects or their layout) for a specific activity. Example: "Which room is best adjusted for working on a project alone?";
- What kind of activities is X best suited for? making predictions on which activities could be performed in a specific room given its capacity (in terms of both the size, available objects and their layout). Example: "What kind of activities is the living room best suited for?".

3.1.3 Replica Dataset Preprocessing

Some adjustments to the original Replica Dataset were necessary to make it compatible with this thesis' question answering system. Certain detected objects are incorrectly labelled, the navigation meshes contain artifacts, and the coordinates of the detections are expressed in Habitat Sim's coordinate systems instead of the one of Replica. Additionally, the answering system required assigning each object to the corresponding room. Moreover, to open a potential path of improvement of the system, a method for individual object descriptions is proposed, which could be integrated into structured object detection format, and be used in *e.g.* an SQL database. In this section, a high-level explanation of the applied modifications and preprocessing is presented, with a detailed list of the changes provided in Appendix D, and the corrected object detections and navigation mesh files present in the repository: https://github.com/Space3D-Bench/Space3D-Bench.

Class Names Modifications

Since the systems proposed in this thesis use Large Language Models, tokenization of the class names could potentially affect their performance. Replica's detected objects have class names, which – if composed of more than one word – include a dash. The dashes were replaced with a whitespace, followed by a simplification of some of the class names which seemed unnecessarily specific in the case of the available scenes *e.g.indoor-plant* was changed to *plant*.

Removing and Replacing Detected Objects

Replica's detections include objects which are irrelevant for the project's use-case, such as *anonymize-text*, *anonymize-picture*, *other-leaf*. Walls, ceilings or floors are not considered as actual objects of interest for a user. Therefore, detections corresponding to application-irrelevant classes were omitted when preparing the final set of detected objects. Additionally, it was noticed that some objects were incorrectly labelled, especially when it concerned furniture such as chairs, tables or beds. Moreover, at one point the user of the spatial Q&A system may want to query information about significantly big objects (*e.g.*mirrors, toys) to which Replica assigned an *undefined* label. Therefore, the detections were manually evaluated and adjusted in terms of labels where necessary.

Coordinate System Transform

The provided object detections are expressed in Habitat Sim's coordinate system, while point clouds – useful for visualizations and room dimensions definition – are expressed in Replica's coordinate system. In this project, the data was unified by transforming the detections' 3D positions and navigation meshes to match Replica's coordinate frame. This may prove valuable not only for the users of Space3D-Bench, but also for other researchers who find the Replica's and Habitat's vague explanations on how to move from one system to the other confusing.

The transforms were inferred from the way that Habitat Sim loads Replica scenes into the its environment. Regarding object detections, axis-aligned bounding boxes's data is loaded from the detections JSON file, and each object's translation and rotation are saved. The translation and rotation are then combined into a transform T. Having each detection's center p_c , the Replica-compatible object center p is obtained by premultiplying it by the transform T: $p = T \cdot p_c$.

To transform navigation meshes from Habitat Sim to Replica, a rotation around x-axis by 90° is required, so effectively each original point $p_o = (x_o, y_o, z_o)$ needs to be expressed as $p_n = (x_o, -z_o, y_o)$.

Modifications To the Navigation Meshes

Navigation meshes, used for determining the navigable distances, similarly to the object detections are expressed in Habitat Sim's coordinate system. The provided files have a binary representation, which requires using Habitat's framework to read them. Moreover, the meshes contain artifacts in a form of navigable islands separated from the main navigable areas, corresponding to the objects with considerable flat surfaces such as beds, sofas or tables with an example presented in Fig. 3.4. Complementary to Space3D-Bench, the users are provided with an improved version of Replica's navigation meshes. They are expressed in the coordinate system of Replica to achieve a unified set of data with respect to the point clouds and detections. Navigation meshes are saved as text files with the definitions of vertices and formed triangles, which makes it easy to parse and compatible with different frameworks regardless of the programming language used. Additionally, by using Breadth-First Search algorithm on the graph derived from the mesh, the largest connected component in the mesh is identified and the remaining artifacts are then removed.



Figure 3.4: **Navigation mesh overlaid on the apartment.** Green mesh represents the navigable area of interest, while the red one presents an artifact removed in the release.

Room Assignment

As the user queries to the Q&A system would naturally contain room names, the provided dataset proposes room labels, their centers and dimensions, and objects-to-rooms assignment. The objects-to-rooms assignment procedure in multi-room scenes assumed rectangular shapes of rooms and their axis-alignment as presented in Fig. 3.5, which – when complemented with particular assignments corrections – proved to be a sufficient approach with a sufficient simplicity-to-correctness ratio, with space for potential improvements in the future works. If rooms exhibited irregular geometries or significant deviations in axis alignment, a post-processing algorithm was employed to reassign incorrectly assigned objects to their corresponding rooms.



(a) Room bounding boxes

(b) Object-room associations

Figure 3.5: **Object-to-room assignments.** Although the defined room bounding boxes are not perfectly aligned with the rooms (a), they proved sufficient for assigning objects to different rooms (b).

Objects Descriptions Generation

The detections in Replica do not contain any semantic information about objects *e.g.* whether a sofa is a twoseater, or if a chair is an armchair. However, for each scene, Replica provides point clouds corresponding to individual object detections. Having extracted those with their corresponding id and class label, it was possible to automatically generate object descriptions in terms of their functional and visual properties such as sizes, shapes, and colors. A vision LLM (GPT4-V [29]) was prompted with a corresponding image for that purpose. Some sections of the individual point clouds were missing because they were identified as parts of other objects, with an example depicted in Fig. 3.6a. Therefore, to avoid VLM's confusion, all point clouds within a certain distance from the center of the considered object were displayed when generating images, to provide a full context to the VLM as presented in Fig. 3.6b. In the final implementation, the distance acting as a threshold for displaying individual objects' point clouds was defined as double the maximum dimension of the object of interest. Additionally, the floor and all the walls were by default rendered, so that the VLM would not by accident consider other rooms' context.

As the orientation of detected objects within the same class was inconsistent, with the 'front' of the objects varying between instances, it was not sufficient to place the camera facing the object's front, as the view could be occluded by an irrelevant point cloud. Therefore, eight images were saved from around the

object to ensure that at least in one of them the object is unoccluded. The camera's positions were apart by 45° , away from the center of the object by the distance 1.5 times greater than object's largest dimension, and at an angle of 30° with respect to the XY plane, with an example in Fig. 3.7. Then, the images were composed into one figure with eight sub-images, and passed to the VLM with a prompt specifying the class of the object it was supposed to describe. However, in the cases where there were several similar objects of the same class present and close to each other (*e.g.*multiple cushions on the bed), the VLM tended to describe the wrong object. To address it, the bounding box of the object of interest is displayed when saving the camera frames. An example of an image passed to the VLM is presented in Fig. 3.8.

This approach could be improved by using additional capabilities of Open3D [49] library. When generating an image from a specific viewpoint, it is possible to access the list of points projected onto the image plane. If none of these points fall within the bounding box of the object of interest, the image can be discarded because the object is occluded from this viewpoint. This approach would reduce the risk of the VLM describing the wrong object, and prevent the transfer of irrelevant data, which can occur in some images, as shown in Fig. 3.8.

Although at this stage of the research the baseline does not use the descriptions of the individual objects, integrating the individual object descriptions into a structured object detections storage such as a database could prove useful for future work.



(a) Point cloud of a sofa



(b) Point clouds of a sofa and nearby objects

Figure 3.6: An issue with the semantic segmentation of Replica's point clouds when generating object descriptions. Displaying a point cloud associated only with the to-be-described object (a) was not sufficient, neighboring point clouds needed to be displayed too (b) to provide more context to the VLM.



Figure 3.7: **Camera positions for saving images of a point cloud.** The 3D transforms corresponding to the camera positions, with the Z axes (in blue) specifying their principal axes.



Figure 3.8: **Example of an image passed to the VLM when generating object descriptions.** It includes eight perspectives of the object (i.e. the sofa) highlighted by a blue bounding box. The object is unoccluded in 4 out of 8 views. Only the point clouds of neighboring objects, located at a distance equal to the doubled object's largest dimension, as well as the walls and floors, are displayed.

3.1.4 Automatic Assessment

Space3D-Bench complements the provided dataset with an automatic assessment system, whose goal is to evaluate the responses from an answering system with respect to the actual state of the corresponding scene in the dataset. The assessment is performed by using LLMs from the GPT4 family as state-of-the-art. The assessment is divided into two cases: Ground Truth Check – when the ground truth is indisputable (e.g. number of objects in the room), Answer Cross-check – when the definition of the ground truth would either need to exceed context length or would unnecessarily limit the answering system's creativity (e.g. finding similarities between rooms). In both scenarios, the VLM is provided with the question, the system's answer, and the acceptance criterion, which varies based on the question type. In the case of the Ground Truth Check, the message to the LLM is extended with information on the actual state of the scene with respect to the given question. Answer Cross-check, however, provides a VLM with an image presenting the corresponding scene(s) in question, accompanied by an example answer. This way, a VLM can decide whether the actual system's answer matches the reality, and not necessarily matching the example, reducing the bias of the assessment system. Then, a language model makes a decision and outputs two components: the acceptance/rejection decision, and its justification, which gives the user of the dataset an insight on what can be addressed in the next iterations of their solution. The examples of the workflows of the assessment system are presented in Fig. 3.9. A detailed list of acceptance criteria corresponding to different tasks and the system prompts are provided in Appendix B.

The generation of ground truth is based on certain assumptions. First of all, when evaluating the distances between objects, their centers are used as reference points. It is reasonable to argue that, for example, if an object is located on top of a table, the distance between the object and the table could be perceived as equal to zero. However, in this method, the distance varies based on the relative positioning of their centers. Despite this, the centers were deemed a satisfactory approximation, extending this rule as well to roomsrelated distance. In cases involving navigable distances, the centers of objects are snapped to the nearest point on the navigation mesh, and the distance between these points is what is considered. Future works may propose a more refined definition of distances. However, given the current limitations posed by Replica's inaccurate 3D bounding box predictions of objects, this approach was chosen as the most pragmatic one under the given circumstances.



Figure 3.9: Automatic assessment procedure. The left chart presents the scenario of *Ground Truth Check*, the right one depicts *Answer Cross-Check*, used respectively for indisputable data and more creative answers.

Additionally, in questions concerning distances, the system's answer should include the distance in numerical form, as well as the id or the 3D coordinates of the objects being considered. This gives a guarantee to the automatic assessment system that the correct objects were used in calculations. It also allows for some flexibility, so systems that calculate navigable distances differently won't be unfairly penalized, as long as they stay within a certain tolerance level.

3.2 RAG3D-Chat

To generate answers for the assessment system to evaluate, RAG3D-Chat is proposed. This spatial Q&A system is based on two main components: Semantic Kernel (SK) [25] and Retrieval Augmented Generation (RAG) [18] within LlamaIndex framework [20]. Semantic Kernel, being an open-source framework for LLM-based implementation of agents, allowed for integrating four complementary modules – each with different applications and limitations – into one system.

Providing prior knowledge of the environment in question manually by the user would be both tedious and potentially resulting in exceeding the context length. To address this, RAG pipelines were implemented for three modules with different sources of data: texts of rooms descriptions, images from the rooms, and an SQL database of detected objects and rooms. The fourth module (navigation) was based on traditional navigation meshes.

For all the Large Language Model applications, off-the-shelf models from GPT4 family were used, specifically *GPT4 0613* [30] for text-related tasks and *GPT4-V* [29] for image analysis.

The overview of the system is presented in Fig. 3.10, with the in-detail explanation of modules present in the following sections.



Figure 3.10: **Overview of RAG3D-Chat.** Based on the received question, the Semantic Kernel library orchestrates the calls of four different modules having different specializations and types of input.

3.2.1 Image Module

Input Data Preparation

The Image Module required the collection of RGB images from the dataset. It was achieved by modifying the Habitat Sim's viewer script, so that a user could navigate through the available apartments and save the desired frames as PNG images. Afterwards, the images needed to be loaded into a specific directories structure – the name of each directory corresponded to the label of the room depicted by the images it contained. It allowed for the module to assign the room labels in metadata efficiently when uploading the images' embedding vectors into the vector store.

Pipeline

The detailed overview of the Image Module is presented in Fig. 3.11. The input images are first embedded with CLIP [34] and loaded into a vector store, with metadata containing the room label. Once a question is received by the module, it is embedded with CLIP and passed to a retriever. If an LLM decides that a specific room of the available ones is related to the query, then a metadata filter is added to the retriever to perform a search over the relevant documents only. Then, the image closest to the question in the embedding space is forwarded to a VLM along with the original question for it to provide the final answer.

As only one image at a time can be provided to the VLM, this module can only be used in the case of single room data. It can answer questions regarding spatial relationships between objects and their visual appearances, provide room descriptions, or count the number of visible objects.



Figure 3.11: **Overview of the Image Module.** After the input images with their room labels are embedded with CLIP in the vector store, the closest image to the input question in the embedding space is retrieved, additionally prefiltered if needed. This closest image is then forwarded to a VLM along with the input question, and the final answer is generated.

3.2.2 Text Module

Input Data Preparation

The rooms' descriptions required by the Text Module were generated by a VLM. The model was provided with images from two views of each room – collected similarly to the case of the Image Module – and asked to describe them in great detail, including the visible objects, their visual appearances, spatial relations between them, and the layout of the room. These descriptions were saved as text files, with names corresponding to the labels of the rooms whose descriptions they contained. The module was then able to assign the room labels in metadata efficiently when uploading the texts' embedding vectors into the vector store.

Pipeline

The pipeline is depicted in Fig. 3.12. At the very beginning, room descriptions are read from the input text files, to be embedded with ADA-002 [31] and loaded into a vector store, with metadata (extracted from the file name) containing the room label. As in the case of the Image Module, the question is embedded with the same function as the data in the vector store, and a metadata filter is added if specific rooms are mentioned in the query. Then, a query engine performs a search of most relevant data and returns an answer formulated by an LLM.

The Text Module addresses questions similar to those handled by the Image Module, but as more data can be retrieved and passed to the LLM at a time, it is suitable for queries involving multiple rooms. However, the generated room descriptions were constrained by the context length. Conversely, the Image Module, which leverages context-rich images, delivers more accurate answers, but is restricted to a single room analysis at a time. Therefore, the two modules needed to be implemented, to complement each other's limitations.



Figure 3.12: **Overview of the Text Module.** Once the room descriptions are embedded into the vector store and the input question is received, a query engine searches for the data closest to the input question in the ADA-002 embedding space, prefiltered if needed. The query engine then composes an answer with a use of an LLM.

3.2.3 Navigation Module

Input Data Preparation

The navigation meshes used as an input for the module are the Replica's navmesh files preprocessed as described in Section 3.1.3. Initially, the original files were used and and the shortest distances were determined with Habitat Sim's existing interfaces. However, this geodesic algorithm implementation encountered difficulties in finding shortest paths on highly irregular surfaces. Considering this aspect and the existence of artifacts, the preprocessed navigation meshes were used in the final implementation.

Pipeline

The Navigation Module in the current state focuses on providing information on both straight-line and shortest navigable distances between rooms and objects. In future work, this could be extended to provide the actual navigation, such as describing the paths between two points.

The question forwarded to the module must contain explicitly mentioned 3D positions of points between which the distance is to be calculated. An LLM extracts the 3D positions from the question, and outputs the start and goal position in a format specifying only the start and goal positions with their corresponding positions in brackets. This predefined output format allowed for applying regular expression matching, with which the text was parsed into a *numpy* object. In the case of the straight-line distance, the Euclidean norm of the difference of the two positions is computed. In the case of the navigable distance, the navigation mesh is considered. The two points are snapped to the closest points contained by the navigation mesh, and then to the closest vertices. The geodesic distance between those two vertices, if possible to be determined with the use of the algorithm implemented in *pygeodesic* library, is output by the module. The overview of the module is depicted in Fig. 3.13.



Figure 3.13: **Overview of the Navigation Module.** First, the 3D positions are extracted from the question. Depending on the question, the positions are passed either to the L2 distance function, or to the function calculating the geodesic distance with respect to the navigation mesh.

3.2.4 SQL Module

Input Data Preparation

The SQL database used in the SQL Module contains two tables – one with room-specific data (Table 3.2), and the other one focusing on available objects in terms of their classes, corresponding rooms, 3D positions and sizes (Table 3.1). The data for both tables can be extracted from the curated object detections prepared as described in Section 3.1.3.

id (PK)	class name	room (FK)	position x	position y	position z	size x	size y	size z
integer	varchar(16)	varchar(16)	float	float	float	float	float	float

Table 3.1: **SQL table schema for** *detected_objects* **table.** The first row presents the names of the attributes, the second row corresponds to their types. The id of the object from Replica's detections serves as a primary key, and the room name is a foreign key from the *rooms* table.

room (PK)	center x	center y	center z	size x	size y	size z
varchar(16)	float	float	float	float	float	float

Table 3.2: **SQL table schema for** *rooms* **table.** The first row presents the names of the attributes, the second row corresponds to their types. The room name serves as a primary key.

Pipeline

The integration of RAG into the SQL query creation process was essential. With a limited set of available class and room names in the database, and with the language capabilities of using words from an unconstrained set of synonyms and descriptive phrases, it was necessary to identify both objects and rooms mentioned in the question, and then perform the search of most similar available class and room names in the ADA-002 embedding space. Once these are determined for a specific prompt, the list of top names matches along with the SQL database context (schema, columns explanations etc.) is forwarded to an LLM, which creates an SQL query corresponding to the objective of the question. Such a query is then executed and its result is used to formulate an answer with an LLM. This approach is presented in Fig. 3.14.

Compared to the other modules, the SQL Module is the exclusive source of quantitative data of the apartment, such as objects count across different rooms, or rooms and objects sizes and centers.



Figure 3.14: **Overview of the SQL Module.** Input room data and object detections are first transformed into an SQL database. At the same time, the names of the rooms and object classes are embedded into a vector store with ADA-002. Once an input question is received, an LLM extracts the objects / rooms mentioned in the query and passes them to a retriever, which searches for the most similar available names / classes in the embedding space. Having those and the SQL context, an LLM creates an SQL query, which is then executed on the database, and its result is finally interpreted by an answer-formulating LLM.

3.2.5 Semantic Kernel

Semantic Kernel (SK) integrates Large Language Models with programming languages like Python, Java, or C#, enabling the chaining of available functions to execute user instructions. Developers can provide their functions to SK's core component, the kernel, with each function described using natural language to detail its purpose, inputs, and outputs. This allows the kernel to execute functions based on user commands expressed in natural language at the application level.

One key feature of SK is its planners, which construct and execute complex plans involving multiple function calls to address user needs. Without SK, developers would need to create an entire pipeline for managing interactions between users, the LLM, and code. This would involve translating LLM responses into function calls, returning results to the LLM, and conducting prompt engineering for the planning LLM. SK's planning framework simplifies this process, making implementation straightforward and efficient.

In RAG3D-Chat, a GPT4 model is integrated into the Semantic Kernel's *Function Calling Stepwise Planner*, which for each question prepares an initial plan of module calls and is able to adjust the plan execution if it encounters a failure. After the modules' implementation, prompts describing functionality, input and output of each needed to be crafted, to later be passed to the SK planner – these prompt-like descriptions of each module are detailed in Appendix E. As a result, the planner is able to propose a chain of function calls for input questions, with examples presented in Fig. 3.15.



Figure 3.15: **Example of modules chained together by the planner.** Semantic Kernel's planner divides the user question into subtasks and calls the modules in a sequence to finally formulate a response based on the retrieved information.
Chapter 4

Experiments

This section is divided into two parts: the user study evaluating the correctness of the automatic answerassessing system, and the baseline's results on the dataset, including the ablation study.

4.1 Evaluation of the Automatic Assessment

4.1.1 User Study

To ensure that the automatic assessment correctly accepts or rejects the answers, a user study was conducted. Sixty people were exposed to a random sample of 40 questions drawn from a 100-questions scene (namely *apartment 2*), and asked to assess the correctness of the system's answers. Additionally, 10 abstracted questions were added to get an insight on people's reasoning in case of ambiguous answers. The questionnaire along with the users' answers is included in Appendix A and the project's website.

As presented in Fig. 4.1, the automatic assessment agreed with the majority of participants' responses in 39 out of 40 cases, reaching the absolute agreement rate of 97.5%. To account for the distribution of participants' decisions with respect to each question, a weighted agreement score was calculated, where the weight for each question was based on the number of users choosing the system-selected option. This metric shows an agreement of 86.4%, demonstrating a strong alignment with human intuition in the evaluation.

The one question resulting in the disagreement between the system and the majority of people is "*List all the rooms with the corresponding number of chairs in them.*". The to-be-evaluated answer lists all rooms apart from the corridor, which does not contain any chairs. The automatic assessment does not approve such a response, as the instruction explicitly specifies *all the rooms*. However, participants almost unanimously agree to accept the answer, justifying that the rooms without the objects are not relevant or that even the corridor should not be considered a room.



Figure 4.1: **Results of the user study with respect to the GPT4-based assessment system.** The chart presents the agreement of the participants of the survey with the system. The users agree with the assessment system on 97.5% of cases, with a weighted agreement score of 86.4%.

Certain questions in the user study resulted in the high level of disagreement between the participants themselves. It concerned the cases in which the answering system gave a mostly correct response, but included a minor misalignment with the ground truth. Although the automatic assessment agreed with the majority's decisions, these situations remain ambiguous and should be addressed by a detailed specification of acceptance criteria.

In addition to the 40 questions on specific answers from the answering system, 10 abstracted ones were included to draw conclusions on how to address ambiguities. For three cases, the participants were almost equally divided between the provided choices, including the decisions on:

- whether an answer about the rooms with *the least* number of objects should consider the rooms without these objects,
- if rooms without specified objects have the same number of these objects,
- if an answer to a question on the number of objects having the wording *at least* instead of the exact number is correct.

These differing views provide valuable insight into the complexity of assessing natural language responses. To address these ambiguities and establish clear acceptance rules, a comprehensive, large-scale user study should be conducted. Moreover, gathering input from professional linguists would be beneficial.

4.1.2 LLM Replacement in the Automatic Assessment System

To further analyze the behavior of the system, Phi3 [1] was chosen as a replacement of GPT4 models in the automatic assessment system. The tests against the questions used in the user study were run. Phi3 scored 77.5% of weighted agreement with the participants, failing to align on 7 out of 40 questions, as presented in Fig. 4.2. Interestingly, the question that led to disagreement among users with GPT-4 resulted in alignment in the user study for Phi3. However, the overall performance of Phi3 was much poorer, therefore it was decided to keep GPT4 models as the core of the assessment system until better models are available.



Figure 4.2: **Results of the user study with respect to the Phi3-based assessment system.** The chart presents the agreement of the participants of the survey with the system. The users agree with the assessment system on 82.5% of cases, with a weighted agreement score of 77.5%.

4.2 RAG3D-Chat

4.2.1 Evaluation on Space3D-Bench

The developed system was tested using questions from Space3D-Bench, and its responses were evaluated by the automatic assessment system. RAG3D-Chat achieved an accuracy of 66.8% on the 1000 questions of the proposed dataset. As presented in Fig. 4.3, the system correctly addresses most of the questions in each category. Predictions pose the biggest challenge for the system, accounting for 76 failed cases. When making predictions, the system tends to consider the areas, or sometimes even volumes of the rooms rather than the size and number of the available furniture. It also lacks the understanding of the types of the objects present – for example, when asked about the number of people that could comfortably sit in the room, it counts the number of seating objects such as chairs, sofas, stools, instead of accounting for the capacity of each. It was additionally noticed that the LLM within the SQL module at times falsely interprets the answers from the SQL database. It is especially manifested in the case of counting the number of objects which are not present in the scene – occasionally the returned zero is interpreted as an object existence. The reason for that particular phenomenon remains unexplained, to be investigated in the future work.



Figure 4.3: **Results of the baseline on each category of questions in Space3D-Bench.** Predictions pose the biggest challenge for RAG3D-Chat due to lack of understanding of objects types. Although the highest percentage of correct answers belongs to navigation-related questions, it is the location category with the highest number of correct responses because of the larger number of questions in this group.

Out of 1000 questions, the SK planner failed to create a plan only 10 times, claiming the lack of appropriate modules. These failures occurred in prediction-related tasks. In the remaining 990 cases, the planner successfully composed a question-focused module-calling plan. Notably, 341 of these plans involved more than one function-calling step. However, as illustrated by Fig. 4.4a, the kernel determined that a single step was sufficient for the majority of cases. The longest plan consisted of 4 steps. It was created for the question *"How many people could play board games here?"* for *room 0* scene, as presented in Fig. 4.5. The generated plan shows both the promising capabilities of the Semantic Kernel, and its gaps in understanding the scene context and making predictions.

For each plan, the modules that were called to derive the final answer were additionally identified. As depicted in Fig. 4.4b, the SQL Module demonstrated the highest usefulness, with the Image Module being least frequently invoked.



Figure 4.4: **Analysis of the plans generated by SK.** Most plans include only one function call (a), and the SQL Module had the highest impact in providing the answers.



Figure 4.5: The longest plan generated by the SK planner. The plan did not provide the correct answer, as it did not consider the full context of the scene (*e.g.* presence of the sofas or existence of multiple objects classified as tables).

4.2.2 Ablation Study

To further analyze the system, an ablation study was conducted, generating answers and evaluating them using Space3D-Bench, with only one module enabled at a time. The results, categorized by the Space3D-Bench question types, are presented in Table 4.1.

The SQL Module demonstrated the highest overall accuracy, outperforming other modules in Location, Measurement, and Path questions. Conversely, the Text Module displayed its dominance in Pattern and Prediction tasks. The Image Module showed superior performance in Relation-focused queries, while the Navigation Module proved useful only when integrated with other modules, not being able to answer any question on its own.

These results highlight the strengths and limitations of each module, providing insights into their specific applications and potential areas for improvement. In contrary to what was expected, the Image Module excelled in Relation-focused tasks, and the data in the SQL database was insufficient for the LLM to be able to calculate Euclidean distances. In the future work, it would be valuable to explore how the modules' implementations would need to be adjusted to address these limitations.

Enabled Module	Location	Measurement	Relation	Path	Pattern	Prediction	Total
Image	32.26%	18.28%	37.23%	0.00%	28.57%	42.55%	28.10%
Text	48.92%	35.48%	36.17%	0.00%	42.06%	53.72%	37.80%
SQL	89.78%	68.28%	18.62%	1.59%	23.81%	43.62%	44.40%
Navigation	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 4.1: **Results of the ablation study.** The SQL module achieved the highest results in the Location, Measurement and Path questions. The Image Module handled the Relation-related tasks best, while the Text Module provided the best results for the Pattern- and Prediction-focused queries.

Chapter 5

Discussion

Space3D-Bench presented in this thesis is a valuable asset for evaluating spatial Q&A systems, addressing gaps in question types found in other datasets. It offers curated data for the Replica dataset and introduces an automated answer assessment system that does not penalize creativity of the answering system. However, various challenges remain open.

For instance, the distance-related questions in the dataset could be extended with navigation aspects, such as describing paths to be followed by a person to get from one place to another. Additionally, the definition of distances between objects could be improved. Instead of measuring the distances between the objects' centers, the nearest points between these objects could be considered as references. This approach would align more closely with human intuition.

It would also be valuable to get an insight on whether answer formulation impacts the assessment system. For example, when predicting the most suitable rooms for a specific activity, providing a description of each room's relevant properties, rather than listing arbitrarily selected room names, could allow the system to evaluate answers with greater flexibility and potentially improve the accuracy of its assessments.

Moreover, as Replica offers a relatively small number of scenes, it would be beneficial to extend the Q&A to other datasets, ideally with multi-floor, multi-room scenes. However, it could require automating the process of questions and answers generation, which could result in a decrease of the dataset's quality.

As natural language tends to be ambiguous, a more detailed user study should be conducted, to determine how human intuition perceives vague or partially-correct answers. Although abstracted questions were included in this project's survey, more research in the area could positively affect future Q&A datasets.

In terms of the provided auxiliary data, the curated Replica object detections could be improved. Objectto-room assignments, for instance, assume rectangular, axis-aligned properties of each room, which proved to be an acceptable approximation. For the sake of higher accuracy and reliability, a better assignment procedure should be employed.

This thesis complements the benchmark with a RAG-based question answering system – RAG3D-Chat. To the best of our knowledge, it is the first spatial Q&A system that does not require prior training, therefore avoiding high costs associated with the GPU usage. RAG3D-Chat yielded promising results on Space3D-Bench, encouraging the usage of RAG-based approaches in similar applications.

There are a couple of known limitations of the baseline that should be addressed in the future work. The most evident one is the lack of a bridge between the objects' semantics (their colors, shapes, types), and their quantitative data (3D positions, sizes, number of objects). It would be insightful to experiment with an integration of object descriptions into the SQL database, potentially with RAG implementation on relevant objects' row-retrieval. It would also be valuable to investigate if and how rephrasing the questions would affect the performance of the answering system. Finally, addressing the poor standalone module performance would be a potential path for the future research.

Chapter 6

Conclusion

In summary, this thesis presents a dataset of 1000 question-answer pairs for spatial Q&A, based on different indoor scenes with a variety of modalities available. The dataset is balanced with respect to the proposed spatial question taxonomy, initially employed in geography-related research. The benchmark is complemented with an answer-assessing system, which leverages a VLM's text and image understanding. The assessment's correctness is justified by the results of a user study. Additionally, this work proposes a question answering system, that applies Retrieval-Augmented Generation with multiple chained VLMs to address spatial Q&A. The system achieves 67% accuracy on the proposed dataset. Although the baseline efficiently handles a variety of question types, there is room for improvement in its 3D reasoning and scene understanding.

Appendix A

Results of the User Study

This section details the results of the user study. Sixty participants filled in the survey, which was divided into two major sections:

• 40 evaluation questions

This part consisted of 40 questions drawn randomly from 100-questions scene. Participants were presented with a question, response from the answering system, and the ground truth information or an image of the relevant scene followed by an example answer (being only a suggestion). In each case, participants were asked to decide whether they accept the system's answer as correct or not. Pie charts corresponding to each question present the distribution of the responses from participants. We additionally state the decision from the automatic assessment system, whose correctness we aimed to evaluate through the user study. The icons next to the questions (a green check and a red cross) indicate whether the automatic assessment system matched the decision of the majority of participants. Additional conditions participants were provided with: the tolerance for the navigable distance was 0.5m, for straight-line distance 0.2m, and for specific coordinates 0.1m; in the case of the questions on viewpoints, the provided images presented the viewpoint of a person described in the question.

• 10 abstracted questions

It was additionally desired to draw conclusions on how to address ambiguities. Ten questions were abstracted and asked to the participants. The divided opinions on some of the cases give a valuable insight on the ambiguities of natural language answers assessment.

The results are available on https://space3d-bench.github.io/user_study.html, and are additionally attached in the following pages.



Q6. Which room has the most chairs?
System's Answer: The dining room has the most chairs with a total of 5 chairs.
Ground Truth: The dining room.
Automatic Assessment Decision: Correct
Correct
Correct
Incorrect

System's Answer: No, there are not any plants in the dining room. Ground Truth: No, there aren't. Automatic Assessment Decision: Correct Correct Incorrect

Correct

Q8. Are there any plants in the corridor?
System's Answer: No, there are no plants in the corridor.
Ground Truth: No, there aren't.
Automatic Assessment Decision: Correct

Q9. Are there any chairs in the apartment?
System's Answer: Yes, there are 9 chairs in the apartment.
Ground Truth: Yes, there are.
Automatic Assessment Decision: Correct

Q10. Are there any armchairs in the dining room?

Ground Truth: No, there aren't.

Automatic Assessment Decision: Incorrect



Incorrect



Q11. What distance would you have to walk to get from the dining room to the living room?

System's Answer: Yes, there are armchairs in the dining room.

System's Answer: The distance you would have to walk to get from the dining room to the living room is approximately 7.87 meters. The distance is calculated between (4.625, 6.83, -0.28) and (4.625, 3.555, -0.28).

Ground Truth: From the dining room at (4.625, 6.83, -0.28) to the living room at (4.625, 3.555, -0.28) the navigable distance is 7.903 meters.

Automatic Assessment Decision: Correct

Q12. What is the distance between the mirror in bedroom and the desk in the study (considering the obstacles)?

System's Answer: The distance between the mirror in the bedroom and the desk in the study, considering the obstacles, is approximately 4.41 meters. The distance is calculated between (0.767, 0.006, -0.624) and (-0.787, 0.059, -1.227).

Ground Truth: From object with id=5 at (0.767, 0.006, -0.624) to object with id=72 at (-0.787, 0.059, -1.227) the navigable distance is 4.414 meters.

Automatic Assessment Decision: Correct



System's Answer: The distance between the sofa in the living room and the table in the dining area, considering obstacles, is approximately 8.12 meters. The distance is calculated between (4.72, 2.382, -1.223) and (4.598, 6.825, -1.195).

Ground Truth: From object with id=23 at (4.72, 2.382, -1.223) to object with id=8 at (4.598, 6.825, -1.195) the navigable distance is 8.011 meters.

Automatic Assessment Decision: Correct

Q14. What is the distance between the desk in the study and the bed in the bedroom (considering the obstacles)?

System's Answer: The distance between the desk in the study and the bed in the bedroom, considering obstacles, is approximately 2.37 meters. The distance is calculated between (0.79, 0.06, -1.23) and (4.22, 0.77, -0.87).

Ground Truth: From object with id=72 at (-0.787, 0.059, -1.227) to object with id=198 at (4.217, 0.766, -0.87) the navigable distance is 4.842 meters.

Automatic Assessment Decision: Incorrect









Q15. What distance would you have to walk to get from the dining room to the study?

System's Answer: The distance you would have to walk to get from the dining room to the study is approximately 11.09 meters. The distance is calculated between (4.625, 6.83, -0.28) and (-0.875, 0.135, -0.28).

Ground Truth: From the dining room at (4.625, 6.83, -0.28) to the study at (-0.875, 0.135, -0.28) the navigable distance is 11.088 meters.

Automatic Assessment Decision: Correct



Q16. What is the distance in straight line between the desk in the study and the bed in the bedroom?

System's Answer: The straight-line distance between the desk in the study and the bed in the bedroom is approximately 5.07 meters. The distance is calculated between (-0.79, 0.06, -1.23) and (4.22, 0.77, -0.87).

Ground Truth: Straight line distance from object with id=72 at (-0.787, 0.059, -1.227) to object with id=198 at (4.217, 0.766, -0.87) is 5.066 meters.



Automatic Assessment Decision: Correct

Q17. What are the similarities between the dining room and the bedroom?

System's Answer: I'm sorry, but I was unable to find the similarities between the dining room and the bedroom because the information about the dining room is not available.

Example Answer: Both have artworks on the walls, windows with blinds, large-area rug.

Automatic Assessment Decision: Incorrect



Correct Incorrect

Q18. Which rooms have a sauna?
System's Answer: There are no rooms with a sauna.
Ground Truth: No room has a sauna.
Automatic Assessment Decision: Correct

✓ Q19. What can a person sitting in an armchair in the bedroom see in front of them?

System's Answer: From the perspective of a person sitting in the armchair in the bedroom, they would be facing a wall with two framed pictures hanging above a small table or shelf, which has a decorative vase or plant on it. The wall appears to be the foot of the room, and the floor is covered with a large area rug. The edge of a bed can be seen to one side, suggesting that the armchair is positioned near the corner of the bedroom. The overall color scheme is neutral, with light walls and furnishings, creating a calm and minimalist atmosphere.

Example Answer: You can see a bed with cushions on it, a window, two nightstands with lamps on the opposite sides of the bed, rug on the floor, open door leading to another room.

Automatic Assessment Decision: Incorrect

✓ Q20. What can a person sitting on the sofa see in front of them? System's Answer: In the image, the sofa is positioned between two lamps that are placed on identical side tables. Directly above the sofa, there is a large framed artwork hanging on the wall. To the side of the sofa, there is a window with blinds partially drawn, allowing some natural light into the room. In front of the sofa, there is a small round coffee table with a few items on it, which appears to be closer to the viewer than the sofa. The room has a neutral color scheme with white walls and a light-colored floor, creating a bright and airy atmosphere.





Example Answer: You can see a coffee table with books, a plant and a decorative item right in front of you. On the opposite wall you can see a hanged painting, and a table with vases, boxes and white flowers beneath the it. To the right you could see a basket on the floow, to the left an open door leading to another room. **Automatic Assessment Decision:** Incorrect





Q21. Which rooms do not contain a sauna?

System's Answer: All rooms in the apartment, which are the bedroom, corridor, dining room, living room, and study, do not contain a sauna.

Ground Truth: Corridor, living room, dining room, study, bedroom do not contain a sauna.

Automatic Assessment Decision: Correct



System's Answer: The rooms that do not contain a bed are the corridor, dining room, living room, and study.

Ground Truth: Corridor, living room, dining room, study do not contain a bed.

Automatic Assessment Decision: Correct

Q23. Which rooms have no paintings?
System's Answer: I'm sorry, but I couldn't find specific information about which rooms contain paintings. Therefore, I can't accurately determine which rooms don't have paintings.
Ground Truth: Corridor has no paintings.

Automatic Assessment Decision: Incorrect





Q24. How many people could comfortably sleep in the bedroom's bed?
System's Answer: The bed in the bedroom can comfortably accommodate 2 people.
Ground Truth: The bed is king-sized.

Q25. How many people could comfortably work on a project in

System's Answer: One person can comfortably work in the study.

Automatic Assessment Decision: Correct

Ground Truth: There is one desk and one chair. **Automatic Assessment Decision:** Correct

the study?

Correct



Q26. In which room could most people sit together?
System's Answer: The living room could accommodate the most people, as it has 9 seating objects.
Ground Truth: The dining room has 5 chairs. The living room has two chairs and a small sofa.
Automatic Assessment Decision: Incorrect



Incorrect

Correct

imes Q27. List all the rooms with the corresponding number of chairs in them.

System's Answer: The bedroom has 1 chair, the dining room has 5 chairs, the living room has 2 chairs, and the study has 1 chair. **Ground Truth:** bedroom: 1, living room: 2, dining room: 5, study: 1,

corridor: 0

Automatic Assessment Decision: Incorrect

Q28. How is the couch in the living room placed with respect to the surrounding objects?

System's Answer: The couch in the living room is centrally positioned against a wall, beneath a large framed artwork. Directly in front of the couch is a small, round coffee table. Flanking the couch on both sides are matching table lamps, each placed on a small table or stand. There is a window with blinds to one side of the couch, providing natural light to the space. The overall arrangement creates a symmetrical and balanced setup in the room.



Example Answer: The couch has small tables with lamps on both its sides, its back against the wall, under a painting. The couch is facing a coffee table with a book, vase and a plant, and in the further side of the room a table with decorative items on it. It is flanked by armchairs on both its sides, and it holds some cushions. **Automatic Assessment Decision:** Correct





System's Answer: The sofa is centrally placed within the room, facing towards a large artwork that hangs on the wall. There is a coffee table positioned directly in front of the sofa, with a small decorative item on top. Flanking the sofa are two matching table lamps, each placed on a surface that appears to be either a side table or a console, which are not fully visible in the image. Behind the sofa, there is a window with blinds partially drawn, allowing some natural light into the room. The overall arrangement suggests a conversational or lounge area designed for relaxation or socializing. **Example Answer:** The sofa has small tables with lamps on both its sides, its back against the wall, under a painting. The sofa is facing a coffee table with a book, vase and a plant, and in the further side of the room a table with decorative items on it. It is flanked by armchairs on both its sides, and it holds some cushions. **Automatic Assessment Decision:** Correct







Q30. What is placed on top of the table in the dining room?

System's Answer: A delicate, tiered centerpiece featuring what seems to be a pair of ornamental dishes or shallow bowls, possibly made of glass or ceramic, with a design that complements the minimalist aesthetic of the room is placed on top of the table in the dining room.

Example Answer: There is a tiered serving tray. **Automatic Assessment Decision:** Correct





✓ Q31. What objects are there on the desk in the study? System's Answer: There is a small potted plant on the desk in the study.

Example Answer: There is a lamp and a potted plant. (Assume that the object on the left on the desk is a desk lamp) **Automatic Assessment Decision:** Incorrect





Q32. Which rooms are best adjusted for a casual conversation with two friends?

System's Answer: The dining room and the living room are the rooms best adjusted for a casual conversation with two friends. However, the dining room, with its 5 seats, might be more comfortable for such a gathering. The living room, despite its comfortable ambiance, might be a bit cramped as it only has 2 seats. The study and the bedroom, while having a comfortable ambiance,

lack the necessary seating arrangements. The corridor, despite being spacious, does not provide a suitable ambiance for a casual conversation.

Ground Truth: The living room has two armchairs and a sofa. The dining room has 5 chairs next to a table. The corridor has no furniture. The study has a desk and a chair. The bedroom has a bed and an armchair.

Automatic Assessment Decision: Correct

Q33. In the dining room, how are the chairs positioned with respect to each other?

System's Answer: In the dining room, there is a round dining table with chairs positioned around it. The chairs are evenly spaced, encircling the table, and all are oriented to face towards the center of the table. There is no clear head of the table due to its round shape, which suggests a more collaborative or egalitarian seating arrangement. The chairs are close enough to the table to be used for dining but also have enough space to be pulled out for sitting. A decorative centerpiece is placed in the middle of the table, and the chairs appear to be upholstered, suggesting a comfortable seating arrangement.

Example Answer: The chairs surround the central round table from all sides.

Automatic Assessment Decision: Correct



Q34. How many artworks are there in the apartment?
System's Answer: There are 12 artworks in the apartment.
Ground Truth: There are 8 paintings, 1 sculpture and 3 vases.
Automatic Assessment Decision: Correct







Q35. How many chairs are there in the apartment?
System's Answer: There are 9 chairs in the apartment.
Ground Truth: 9
Automatic Assessment Decision: Correct



Incorrect

Q36. Which rooms are directly accessible from the study?
System's Answer: The rooms directly accessible from the study are a waiting or sitting area, an office or meeting room, and a relaxation or informal meeting space.
Ground Truth: The corridor.

Automatic Assessment Decision: Incorrect

Q37. What kind of activities is the largest room best suited for?

System's Answer: The largest room in the apartment is the corridor. It contains a lamp and a vent. The corridor has a modern and functional aesthetic with a neutral color palette and clean lines. Given its size and the objects present, the corridor is best suited for movement and transit within the apartment. It may also serve as a good place for standing conversations, quick catch-ups, or even as a space for pets to roam, if any. However, it might not be suitable for activities that require furniture or specific equipment, such as dining, working, or entertainment.

Example Answer: The corridor is the largest room. In there, you could do any indoor activity that requires empty space e.g. yoga, dancing.

Automatic Assessment Decision: Correct







Correct

Q38. Which rooms have a bed?
System's Answer: The room that has a bed is the bedroom.
Ground Truth: Only the bedroom has a bed.
Automatic Assessment Decision: Correct



Q39. Which room would be best for having a meal for 3 people? System's Answer: The best room for having a meal for 3 people is the dining room. It features a round dining table with high-back chairs, a delicate centerpiece, and a large white area rug. The room is bright and airy, with a minimalist and sophisticated aesthetic. Ground Truth: The dining room has a table and 5 chairs. The living room has two armchairs and a sofa, but no dining table. The bedroom has a bed and an armchair, but no dining table. The study has one chair and one desk. The corridor has no furniture. Automatic Assessment Decision: Correct



Q40. Which rooms would be best for having a casual meeting for 3 people?

System's Answer: The best rooms for having a casual meeting for 3 people would be the dining room and the living room.

Ground Truth: The dining room has a table and 5 chairs. The living room has two armchairs and a sofa. The bedroom has a bed and an armchair, but no table. The study has one chair and one desk. The corridor has no furniture.

Automatic Assessment Decision: Correct



Acknowledgements

We thank all the participants of the user study for their time and effort. Special appreciation goes to our colleagues from Microsoft, the ETH CVG lab, and those who joined us through the authors' social media channels.

Q1. Imagine there are two rooms A, B with a sofa in each and three rooms C, D, E without any sofa. You are asked a question 'Which rooms have the same number of sofas?'. A system says 'A and B' without mentioning that rooms C, D, E have the same (zero) number of sofas. Would you consider the answer as a correct one?

Q2. If there are three rooms - a bedroom with two beds, a living room with one bed, a bathroom with no beds - and you are asked 'Which rooms have least beds?', would you choose the living room (LR) or the bathroom (B)?

Q3. If you ask 'How many beds are there?' and you get the response 'Several', 'A few', 'Multiple', would you accept it as a correct response?

Q4. If you ask 'Where are chairs located?' and get the response 'There are chairs in the living room and the dining room.' (which is not false, but a study also has chairs), would you accept the answer as a correct one?

Q5. If there are in total 6 chairs in the room, and a system answers the question 'How many chairs are there in the room?' with 'There are 2 chairs.' (but does not use the words 'in total'), would you accept the system's response?



Yes No







Q6. If there are in total 6 chairs in the room, and a system answers the question 'How many chairs are there in the room?' with 'There are at least 2 chairs.', would you accept the system's response?

Q7. If a system is supposed to describe a scene and uses the words 'to the left' or 'to the right', which are true for a specific viewpoint, but not for all, would you accept the answer?

Q8. If you ask a question 'How are the chair and the bed positioned with respect to each other?', and you get a correct response followed by question-irrelevant yet false description of other objects, would you accept the answer?

Q9. If you ask 'Which rooms are best adjusted for XYZ?' and in the answer one room is mentioned, would you accept the answer?

Q10. If you ask 'Which room is best adjusted for XYZ?' and in the answer a couple of rooms are mentioned (and correctly justified), would you accept the answer?









📕 Yes 📒 No



Appendix B

Automatic Assessment Prompts

This appendix details the system prompts passed to the Large Language Model in the automatic assessment system. Additionally, it presents the answer acceptance criteria divided into Location (Tab. B.1), Pattern (Tab. B.2), Measurement (Tab. B.3), Relation (Tab. B.4), Prediction(Tab. B.5) and Path(Tab. B.6) categories.

System Prompt for Text Ground Truth Check

You are an answer evaluation system. You are provided with a question, a ground truth answer, a to-beevaluated answer and a correctness criterion in the following format:

Question: How many people can sit in the room?

Ground truth answer: 4

Answer: Four people can sit in the room

Criterion: The numbers mentioned in the ground truth and the actual answer should match.

Your task is to evaluate the correctness of the answer based on the provided information. Return it in a format of JSON, where under the key 'result' should be an evaluation result, in a form of a value '1' (correct) or '0' (incorrect), and under the key 'justification' there should be your text description of why the result is correct or not. Do not return anything that could not be directly parsed to a JSON. Example of the output for the previous example:

{'result': '1', 'justification': 'Both the answer and the ground truth mention the same number of people that can sit in the room.'}

System Prompt for Image-based Answer Cross-Check

You are an answer evaluation system. You are provided with a question, a to-be-evaluated answer, an image, an example answer, and a correctness criterion in the following format:

Question: What can a person sitting in the chair see in front of them?

Example answer: A small coffee table.

Answer: The person sitting in the chair can see a small coffee table with a potted plant on top.

Criterion: Based on the image and the example answer decide, whether the actual answer is correct (if no objects are hallucinated etc.).

Your task is to evaluate the correctness of the answer based on the provided information and the image. Return it in a format of JSON, where under the key 'result' should be an evaluation result, in a form of a value '1' (correct) or '0' (incorrect), and under the key 'justification' there should be your text description of why the result is correct or not. Do not return anything that could not be directly parsed to a JSON. Example of the output for the previous example:

{'result': '1', 'justification': 'In the provided image there is a coffee table with a plant and books on it.'}

Task's identification focus	Acceptance criterion		
Object locations (in terms of	The answer should contain 3D positions of the objects or the names of		
corresponding rooms or 3D	the rooms in which they can be found. The ground truth provides a list		
positions)	of all 3D positions and names of all rooms with the specified object. For		
	3D positions, assume the tolerance of 0.1 meters for each coordinate.		
	The order is irrelevant. If the answer contains even one 3D positions or		
	room which is not mentioned in the ground truth, the answer is incorrect.		
Object locations (in terms of	f The answer should contain all the rooms that are mentioned in the		
corresponding rooms)	ground truth. The order is irrelevant. If the answer contains even or		
	room which is not mentioned in the ground truth, it is incorrect.		
Object locations (in terms of	The answer should contain 3D positions of the objects. Assume the tol-		
corresponding 3D positions)	erance of 0.1 meters for each coordinate. The order of objects is irrele-		
	vant. If the answer contains any 3D positions which are not mentioned		
	in the ground truth, it is incorrect.		
Existence of objects	The answer should match the ground truth in terms of yes/no statement.		
Rooms without specified ob-	- The answer should contain all rooms mentioned in the ground truth. The		
jects	order is irrelevant. If only some are mentioned, it is incorrect.		

Task's identification focus	Acceptance criterion		
Rooms with the same num-	The answer should contain names of rooms matching the ground truth,		
ber of the specified objects	clustered into groups if there exist groups of rooms with different objective		
	count. If the answer contains synonyms of the room names, it is correct.		
	The order within a group is irrelevant. If the ground truth contains the		
	rooms corresponding to the 0 object count, and the answer does not		
	explicitly mention the corresponding rooms without the objects, but the		
	rest is mentioned correctly, accept the answer.		
Similarities	You are provided with the RGB image, divided into two parts by a ver-		
	tical red line, corresponding to the objects in question. Decide whether		
	the answer on the similarities between the objects is correct based on		
	the provided image. An example of the correct answer is provided as		
	well, however, the actual answer does not need to fully match the exam-		
	ple answer. As long as it is reasonable in accordance to the image, the		
	example answer and the question, it is correct.		
Wall colors	You are provided with the RGB image, divided into two views of the		
	room by a vertical red line. Decide whether the answer on the room		
	wall colors is correct based on the provided image. An example of the		
	correct answer is provided as well, however, the actual answer does not		
	need to fully match the example answer. As long as it is reasonable		
	in accordance to the image, the example answer and the question, it is		
	correct.		
Consistency	The answer should contain a clear 'yes' or 'no' response, matching the		
	ground truth. If the answer is not clear, it is incorrect.		

Task's identification focus	Acceptance criterion		
Largest room	The answer should contain exactly one room name, matching the one		
	from the ground truth.		
Smallest room	The answer should contain exactly one room name, matching the one		
	from the ground truth.		
Smallest rooms	The answer should contain all the rooms that are mentioned in the		
	ground truth. The order is irrelevant. If the answer contains any rooms		
	which are not mentioned in the ground truth, it is incorrect.		
Size-sorted rooms	The answer should contain all the rooms that are mentioned in the		
	ground truth, in the exact same order. If the order does not match the		
	one of the ground truth, it is incorrect.		
Number of objects	The answer should contain a number of objects matching the ground		
	truth.		
List of rooms with corre-	The answer should contain a list of all the rooms with the corresponding		
sponding object numbers	number of specified objects. The order is not relevant, but the room-to-		
	number association should be the same as in the ground truth. If a room		
	is missing in the answer, it is incorrect.		
Room with most specified	d The answer should contain a name or names of rooms that match the		
objects	ground truth. In case of more than one room mentioned, the order is no		
	relevant. The number of rooms in the answer and in the ground truth		
	should match.		
Room height	The answer should contain a number in meters, matching the ground		
-	truth with a tolerance of 0.1 meters.		
Room area	The answer should contain a number in square meters, matching the		
	ground truth with a tolerance of 0.5 square meters.		
Room volume	The answer should contain a number in cubic meters, matching the		
	ground truth with a tolerance of 0.5 cubic meters.		
Largest furniture	The answer should contain a name of one object, matching the ground		
	truth.		
Object area	The answer should contain a number in square meters, matching the		
	ground truth with a tolerance of 0.1 square meters.		
Dimensions of rooms or ob-	s or ob- The answer should contain three numbers in meters, matching the		
jects	ground truth. The order is relevant, unless the answer specifies which		
	number corresponds to which coordinate (by default assume XYZ or-		
	der).		

Table B.3: Acceptance criteria for measurement-related answers

Task's identification focus	Acceptance criterion		
Objects within a given dis-	The answer should contain a list of objects matching the ground truth.		
tance	The order is irrelevant. If the answer aggregates the objects of the same		
	class into one mention, but with the correct number of objects or using		
	a quantifier (like many, some, a few), it is correct.		
Object closest to a given ob-	The answer should contain the name of one object, matching the ground		
ject	truth.		
Spatial relationship between	You are provided with an image, presenting the object(s) described in		
objects	the question. If the image is divided into two parts with a red vertical		
	line, the two parts correspond to the different views of the same scene.		
	Decide whether the answer on the objects spatial relationship is correct		
	based on the provided image. An example of the correct answer is pro-		
	vided as well, however, the actual answer does not need to fully match		
	the example answer. As long as it is reasonable in accordance to the		
	image, the example answer and the question, it is correct. If the answer		
	is vague, but still correct in a sense, accept it.		
Rooms accessible from a	The answer should contain the names of rooms matching the ground		
given room	truth. If the answer contains synonyms of the room names, it is correct.		
	If the answer contains any room names (or their synonyms) not men-		
	tioned in the ground truth, it is incorrect. The order is irrelevant.		
Description of a room	You are provided with the RGB image, divided into two views of the		
	same room by a vertical red line. Decide whether the answer correctly		
	describes the objects in the room based on the provided image. An		
	example of the correct answer is provided as well, however, the actual		
	answer does not need to fully match the example answer. As long as it		
	is reasonable in accordance to the image, the example answer and the		
	question, it is correct.		

Table B.4: Acceptance criteria for relation-related answers

Task's identification focus	Acceptance criterion		
Rooms where an activity can	The answer should contain at least one name of the rooms mentioned in		
be performed	the ground truth. If the answer contains synonyms of the room names,		
	it is correct. The order is irrelevant. If the answer contains any rooms		
	which are not mentioned in the ground truth, it is incorrect.		
Number of people who can	The answer should contain a number of people matching the ground		
perform a specific activity	truth.		
Visible objects from a given	You are provided with a single image, presenting a viewpoint of the per-		
viewpoint	son described in the question. Decide whether the answer on the view is		
	correct based on the provided image. An example of the correct answer		
	is provided as well, however, the actual answer does not need to fully		
	match the example answer. As long as it is reasonable in accordance to		
	the image, the example answer and the question (e.g. the hallucinated		
	objects are not mentioned), it is correct.		
Room best suited for a given	The answer should contain at least one name of the rooms matching the		
activity	ground truth. If the answer contains synonyms of the room names, it is		
	correct. The order is irrelevant in case multiple rooms are mentioned. If		
	the answer contains any rooms which are not mentioned in the ground		
	truth, it is incorrect.		
Possible activities that a	You are provided with the RGB image, divided into two views of the		
given room is suited for	room by a vertical red line. Decide whether the answer on the poten-		
	tial activities in the room is correct based on the provided image. An		
	example of the correct answer is provided as well, however, the actual		
	answer does not need to fully match the example answer. As long as it		
	is reasonable in accordance to the image, the example answer and the		
	question, it is correct.		
Possibility of performing a	The answer should contain a clear 'yes' or 'no' response, matching the		
given activity	ground truth. If the answer is not clear, it is incorrect.		

Table B.5: Acceptance criteria for prediction-related answers

Task's identification focus	Acceptance criterion		
Navigable distance between	The answer should specify the distance in meters, matching the ground		
objects	truth with a maximum tolerance of the difference being 0.5 meter. The		
	answer should also specify either the objects ids or the positions (with		
	a tolerance of 0.1 meter per coordinate) for which the distance is calcu-		
	lated. If neither ids nor positions are mentioned, or if they do not match		
	the ground truth, the answer is incorrect.		
Navigable distance between	The answer should specify the distance in meters, matching the ground		
an object and a room	truth with a maximum tolerance of the difference being 0.5 meter. The		
	answer should also specify either the object id and room name, or the		
	positions (with a tolerance of 0.1 meter per coordinate) for which the		
	distance is calculated. If neither id and name, nor positions are men-		
	tioned, or if they do not match the ground truth, the answer is incorrect.		
Navigable distance between	The answer should specify the distance in meters, matching the ground		
rooms	truth with a maximum tolerance of the difference being 0.5 meter. The		
	answer should also specify the room names or their positions (with a		
	tolerance of 0.1 meter per coordinate) for which the distance is calcu-		
	lated. If neither room names nor the positions are not mentioned, or if		
	they do not match the ground truth, the answer is incorrect.		
Straight-line distance be-	The answer should specify the distance in meters, matching the ground		
tween objects	truth with a maximum tolerance of the difference being 0.2 meter. The		
	answer should also specify either the objects ids or the positions (with		
	a tolerance of 0.1 meter per coordinate) for which the distance is calcu-		
	lated. If neither ids nor positions are mentioned, or if they do not match		
	the ground truth, the answer is incorrect.		
Straight-line distance be-	The answer should specify the distance in meters, matching the ground		
tween an object and a room	truth with a maximum tolerance of the difference being 0.2 meter. The		
	answer should also specify either the object id and room name, or the		
	positions (with a tolerance of 0.1 meter per coordinate) for which the		
	distance is calculated. If neither id and name, nor positions are men-		
	tioned, or if they do not match the ground truth, the answer is incorrect.		
Straight-line distance be-	The answer should specify the distance in meters, matching the ground		
tween rooms	truth with a maximum tolerance of the difference being 0.2 meter. The		
	answer should also specify the room names or their positions (with a talenance of 0.1 meter and coordinate) for which the distance is also		
	loterance of 0.1 meter per coordinate) for which the distance is calcu-		
	lated. If neutrer room names nor the positions are not mentioned, or if		
	they do not match the ground truth, the answer is incorrect.		

Table B.6: Acceptance criteria for path-related answers

Appendix C

Bias Study

The task of the participants in the bias study was to provide feedback on the formulation on the questions, *e.g.* whether they considered them as ambiguous or unnaturally phrased. In this section, the questions that the study participants were exposed to, along with the considered scene, depicted in Fig. C.1, are attached. In the bullet points corresponding to the individual questions the most relevant comments, which were considered when crafting Space3D-Bench, are listed.



Figure C.1: **The apartment considered in the bias study.** The participants of the bias study were presented with the top view of the apartment along with views of the individual rooms.

- 1. In which rooms can you find chairs?
 - I wonder if "in which room" part is actually needed here. Maybe "Where can you find chairs?" would be better.
- 2. In the dining room, where is the mirror placed?
 - "Where is the mirror placed?" should be sufficient.

- 3. Are there any plants in the apartment?
- 4. Which room is the biggest?
- 5. How many paintings are there in the apartment?
- 6. Which room has the least sofas?
 - I guess there are two answers to that one? Is 0 sofas the least sofas?
- 7. Which rooms have the same number of sofas?
- 8. What objects are within 1 meter from the center of the sofa in the living room?
 - This seems like a strange question. No one would formulate a question like this. Maybe "1 meter from the sofa" instead of "1 meter from the center of the sofa" would be better.
- 9. Which object in the dining room is closest to the table?
- 10. Which objects are on the small table in the living room?
- 11. How are the sofa and the TV screen in the living room positioned with respect to each other?
- 12. Which rooms are neighboring to the corridor?
 - This question is unclear to me. Are we asking which rooms share a physical wall with the corridor, or which rooms can be accessed from the corridor, i.e. there is a door from the corridor to the room?
- 13. What is the distance between the sofa in the living room and the table in the dining room (considering the obstacles)?
- 14. What is the distance in straight line between center of the living room and the center of the corridor?
 - I can't imagine a scenario where anyone would want to know this arbitrary measurement. A more practical, but similar query in terms of obtaining metric information might be "What is the length of the corridor?".
 - Mentioning the centers of the rooms sounds weird.
- 15. What distance would you have to walk to get from the living room to the center of the dining room?
 - Perhaps it is more practical to ground the question with affordances offered by objects in the scene, like walking from the sofa to a chair at the dining table.
- 16. In which room could most people eat together?
- 17. In which room could you relax, while drinking a coffee and reading a book?
- 18. How many people could comfortably rest in the living room?
 - It is too vague. By "rest" do you mean "sit", "have lunch" or "hang out on the couch"?
- 19. What can a person lying on the sofa see?
- 20. Which room would be best for having a meeting for two people?
 - This is maybe too subjective. I suppose most people would prefer the living room, but the kitchen is not a bad option either if you want to have a snack, make coffee, etc.

Appendix D

Modifications to the Replica Dataset

Semantic segmentation provided by Habitat Sim can sometimes result in incorrect object-to-class assignments. In Tables D.2 and D.1, the fixes for the Replica dataset's object detections are suggested. For each scene, the objects that are considered to be incorrectly labeled are listed, along with their object ID, original class, and the suggested class assignment.

object id	original class	updated class		
apartment 0				
222	shower stall	undefined		
101	bed	undefined		
34, 407	undefined	wall decoration		
135	undefined	teddy bear		
68, 199, 268, 349	undefined	mirror		
54	undefined	fireplace		
apartment 1				
34	anonymize picture	picture		
53	undefined	mirror		
33	cushion	plant stand		
apartment 2				
30, 32, 38, 47, 68,	anonymize picture	picture		
153, 162, 169				
5	undefined	mirror		
FRL apartment 0				
13, 65, 72, 224	anonymize picture	picture		
35	undefined	basket		
FRL apartment 1				
80, 128, 140, 202	anonymize picture	picture		
200	undefined	basket		
hotel 0				
55, 100	anonymize picture	picture		
78	undefined	mirror		
130	undefined	hanger		

Table D.1: Changes of class names in multi-room Replica scenes

object id	original class	updated class		
	room 0			
47	undefined	mirror		
	room 1			
50	50 undefined			
	room 2			
58	anonymize picture	picture		
office 0				
42	undefined	whiteboard		
49	undefined	sculpture		
	office 2			
27, 34	anonymize picture	picture		
46	anonymize text	picture		
office 3				
20	table	undefined		
37	chair	undefined		
office 4				
10	non-plane	table		
14	floor	chair		
21	panel	tablet		

Table D.2: Changes of class names in one-room Replica scenes
Appendix E

Semantic Kernel Modules' Descriptions

For the Semantic Kernel's planner to be able to schedule a plan with the available modules, each module needed to be described in terms of its input, output and functionality. In the following sections, these descriptions corresponding to the modules available in RAG3D-Chat are provided.

The Image and Text Modules contain only one SK-related function each. The Navigation Module includes two functions: one for the determination of navigable distances, one for the straight-line distance. The SQL Module is also divided into two parts. The first part handles general queries, while the other one is tailored for questions that focus on Euclidean-distance-based comparisons of a given object with others in the scene (*e.g.* finding objects closest / furthest / within certain distance from the given one).

E.0.1 Image Module

Input: Query asking for a visual data of an object, room or scene.

Output: String being the visual data of the object, room or scene.

Functionality: If the query concerns exactly one room, it can answer questions regarding the spatial relationships between objects within this room (such as A is on B, A is under B), count the number of visible objects or describe the surroundings in terms of visual appearance. It does not provide any 3D positions of objects nor the cumulative data about more than one room.

E.0.2 Text Module

Input: Query asking for a visual data of an objects, rooms or scenes.

Output: String being the visual data of the objects, rooms or scenes.

Functionality: If the query concerns two rooms or more, it can answer questions regarding visual data about the scenes, such as appearances of rooms and objects, and about spatial relations between the objects (such as A is on B, A is next to B etc.). It does not provide any quantitative or positional data, such as 3D positions of objects.

E.0.3 Navigation Module

Navigable Distance Function

Input: Natural language query specifying the positions of the start and end, including X, Y and Z components.

Output: Information on the straight-line distance between the points (NOT considering the obstacles). **Functionality:** Having the description of the 3D positions of start and goal, returns the distance between

them in straight line, NOT considering any obstacles. It does not independently determine the positions of objects or points in space, the positions need to be included in the input query.

Straight-line Distance Function

Input: Natural language query specifying the positions of the start and end, including X, Y and Z components.

Output: Information on the actual distance between the points (considering the obstacles and non-navigable areas). If the input was in correct format, it can still happen that the path between objects is not navigable. **Functionality:** Calculates the distance between two points, considering obstacles and non-navigable areas, when provided with the 3D positions of those points. It does not independently determine the positions of objects or points in space, the positions need to be included in the input query. It is the default distance measurement when the query implies walking or getting from one place to another, since it considers walls separating the rooms.

E.0.4 SQL Module

Distance-related SQL Function

Input: Natural language query containing the 3D position of an object to compare with others and the description of the task objective.

Output: Natural language answer to the query.

Functionality: Returns the data related to Euclidean distance between multiple objects, provided that the 3D position of an object which we want to compare with others is in the query. It can e.g. get the names of the objects closest/furthest/within certain distance from the given 3D position.

General SQL Function

Input: Natural language query (NOT an SQL query).

Output: Natural language answer to the query.

Functionality: Answers natural language questions regarding quantitative information about the apartment, such as the categories and number of objects in rooms, their 3D positions, sizes, and the rooms they are associated with. It is the exclusive source for 3D positional data of objects within the apartment. It does not contain data about the visual appearance of the objects or rooms, nor the spatial relations between objects (like A is next to B, A is under B). It can create SQL queries involving the calculation of Euclidean distances between objects.

Bibliography

- [1] Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. CoRR, abs/2404.14219, 2024. URL: https://doi.org/10.48550/arXiv.2404.14219, arXiv:2404.14219, doi:10.48550/ARXIV.2404.14219.
- [2] 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 visuallygrounded navigation instructions in real environments. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3674–3683, 2018. doi:10.1109/CVPR.2018.00387.
- [3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision* and Pattern Recognition (CVPR), 2022.
- [4] Efe Bozkir, Süleyman Özdel, Ka Hei Carrie Lau, Mengdi Wang, Hong Gao, and Enkelejda Kasneci. Embedding large language models into extended reality: Opportunities and challenges for inclusion, engagement, and privacy. In ACM Conversational User Interfaces 2024, CUI '24. ACM, July 2024. URL: http://dx.doi.org/10.1145/3640794.3665563, doi:10.1145/3640794.3665563.
- [5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag

Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. URL: https://arxiv.org/abs/2307.15818, arXiv:2307.15818.

- [6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yuxin Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11618–11628, 2019. URL: https://api.semanticscholar.org/CorpusID:85517967.
- [7] Haonan Chang, Kowndinya Boyalakuntla, Shiyang Lu, Siwei Cai, Eric Pu Jing, Shreesh Keskar, Shijie Geng, Adeeb Abbas, Lifeng Zhou, Kostas Bekris, and Abdeslam Boularious. Context-aware entity grounding with open-vocabulary 3d scene graphs. In 7th Annual Conference on Robot Learning, 2023. URL: https://openreview.net/forum?id=cjEI5qXoT0.
- [8] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *16th European Conference on Computer Vision (ECCV)*, 2020.
- [9] Jang Hyun Cho, Boris Ivanovic, Yulong Cao, Edward Schmerling, Yue Wang, Xinshuo Weng, Boyi Li, Yurong You, Philipp Krähenbühl, Yan Wang, and Marco Pavone. Language-image models with 3d understanding, 2024. URL: https://arxiv.org/abs/2405.03685, arXiv:2405.03685.
- [10] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision* and Pattern Recognition (CVPR), IEEE, 2017.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association* for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL: https://aclanthology.org/N19–1423, doi:10.18653/v1/N19–1423.
- [12] Cathy Mengying Fang, Krzysztof Zieliński, Pattie Maes, Joe Paradiso, Bruce Blumberg, and Mikkel Baun Kjærgaard. Enabling waypoint generation for collaborative robots using llms and mixed reality, 2024. URL: https://arxiv.org/abs/2403.09308, arXiv:2403.09308.
- [13] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022. URL: http://dx.doi.org/10.18653/v1/2022.acl-long.524, doi: 10.18653/v1/2022.acl-long.524.
- [14] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning, 2023. arXiv:2309.16650.

- [15] Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 3d concept learning and reasoning from multi-view images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id= nZeVKeeFYf9.
- [17] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023. URL: https://www.sciencedirect.com/science/ article/pii/S1041608023000195, doi:10.1016/j.lindif.2023.102274.
- [18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/ 6b493230205f780elbc26945df7481e5-Paper.pdf.
- [19] Mingsheng Li, Xin Chen, Chi Zhang, Sijin Chen, Hongyuan Zhu, Fukun Yin, Gang Yu, and Tao Chen. M3dbench: Let's instruct large models with multi-modal 3d prompts, 2023. arXiv:2312.10763.
- [20] Jerry Liu. LlamaIndex, 11 2022. URL: https://github.com/jerryjliu/llama_index, doi:10.5281/zenodo.1234.
- [21] X Ma, Y Bhalgat, B Smart, S Chen, X Li, J Ding, J Gu, DZ Chen, S Peng, JW Bian, P Torr, M Pollefeys, M Nießner, ID Reid, AX Chang, I Laina, and VA Prisacariu. When llms step into the 3d world: a survey and meta-analysis of 3d tasks via multi-modal large language models. IEEE, 2024.
- [22] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. URL: https://openreview.net/forum?id=IDJx97BC38.
- [23] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt, 2023. URL: https://arxiv.org/abs/2310.01415, arXiv:2310.01415.
- [24] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving, 2024. URL: https://arxiv.org/abs/2311.10813, arXiv:2311.10813.
- [25] Microsoft. Semantic kernel, 2024. Accessed: 2024-08-01. URL: https://github.com/ microsoft/semantic-kernel.
- [26] Nicholas M. Giner Miriam Schmidts. Understanding the basics: Introduction to the language of spatial analysis. https://proceedings.esri.com/library/userconf/proc18/ tech-workshops/tw_1593-380.pdf. Accessed: 2024-07-29.

- [27] Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. SPARTQA: A textual question answering benchmark for spatial reasoning. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online, June 2021. Association for Computational Linguistics. URL: https://aclanthology.org/2021.naacl-main.364, doi:10.18653/ v1/2021.naacl-main.364.
- [28] Huan Ning, Zhenlong Li, Temitope Akinboyewa, and M. Naser Lessani. Llm-find: An autonomous gis agent framework for geospatial data retrieval, 2024. URL: https://arxiv.org/abs/2407. 21024, arXiv:2407.21024.
- [29] OpenAI. Gpt-4v(ision) system card, 2023. Accessed: 2024-08-01. URL: https://api. semanticscholar.org/CorpusID:263218031.
- [30] OpenAI. Gpt-4 technical report, 2024. Accessed: 2024-08-01. URL: https://arxiv.org/abs/ 2303.08774, arXiv:2303.08774.
- [31] OpenAI. New and improved embedding model, 2024. Accessed: 2024-08-01. URL: https://openai.com/index/new-and-improved-embedding-model/.
- [32] Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point: A unified framework for point-language understanding and generation. In *CVPR*, 2024.
- [33] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yueping Jiang. Nuscenes-qa: A multimodal visual question answering benchmark for autonomous driving scenario. In AAAI Conference on Artificial Intelligence, 2023. URL: https://api.semanticscholar.org/CorpusID: 258866014.
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL: https://proceedings.mlr.press/v139/radford21a.html.
- [35] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- [36] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797, 2019.

- [37] Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. Llmr: Real-time prompting of interactive worlds using large language models, 2024. URL: https://arxiv.org/abs/2309.12276, arXiv:2309.12276.
- [38] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Niessner. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, 2019.
- [39] Jiaqi Wang, Zihao Wu, Yiwei Li, Hanqi Jiang, Peng Shu, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Huaqin Zhao, Zhengliang Liu, Haixing Dai, Lin Zhao, Bao Ge, Xiang Li, Tianming Liu, and Shu Zhang. Large language models for robotics: Opportunities, challenges, and perspectives, 2024. URL: https://arxiv.org/abs/2401.04334, arXiv: 2401.04334.
- [40] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. arXiv preprint arXiv:2308.16911, 2023.
- [41] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, Alexander William Clegg, and Devendra Singh Chaplot. Habitat-matterport 3d semantics dataset, 2023. URL: https: //arxiv.org/abs/2210.05633, arXiv:2210.05633.
- [42] Xu Yan, Zhihao Yuan, Yuhao Du, Yinghong Liao, Yao Guo, Shuguang Cui, and Zhen Li. Comprehensive visual question answering on point clouds through compositional scene manipulation. *IEEE Transactions on Visualization & Computer Graphics*, (01):1–13, 2023.
- [43] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. *IEEE Transactions on Visualization and Computer Graphics*, 30(3):1772–1786, 2024. doi:10.1109/TVCG. 2022.3225327.
- [44] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Jing Shao, and Wanli Ouyang. Lamm: Language-assisted multimodal instruction-tuning dataset, framework, and benchmark, 2023. URL: https://arxiv.org/ abs/2306.06687, arXiv:2306.06687.
- [45] Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20623–20633, June 2024.
- [46] Sha Zhang, Di Huang, Jiajun Deng, Shixiang Tang, Wanli Ouyang, Tong He, and Yanyong Zhang. Agent3d-zero: An agent for zero-shot 3d understanding, 2024. URL: https://arxiv.org/abs/ 2403.11835, arXiv:2403.11835.
- [47] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences, 2020. URL: https://arxiv.org/ abs/2001.06891, arXiv:2001.06891.
- [48] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. arXiv preprint arXiv:2305.16986, 2023.
- [49] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.

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