NUMINA: A Natural Understanding Benchmark for Multi-dimensional Intelligence and Numerical Reasoning Abilities

Changyu Zeng[1,2,4,*], Yifan Wang[2,*], Zimu Wang[1,2,*], Wei Wang[1], Zhengni Yang[1,2], Muyi Bao[1,2], Jiming Xiao[1], Anh Nguyen[2], Yutao Yue[3,4,†]

[1]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, China
[2]Department of Computer Science, University of Liverpool, United Kingdom
[3]The Hong Kong University of Science and Technology (Guangzhou), China
[4]Institute of Deep Perception Technology, JITRI, China
changyu.zeng17@student.xjtlu.edu.cn, yutaoyue@hkust-gz.edu.cn

Abstract

Recent advancements in 2D multimodal large language models (MLLMs) have significantly improved performance in vision-language tasks. However, extending these capabilities to 3D environments remains a distinct challenge due to the complexity of spatial reasoning. Nevertheless, existing 3D benchmarks often lack fine-grained numerical reasoning task annotations, limiting MLLMs' ability to perform precise spatial measurements and complex numerical reasoning. To address this gap, we introduce NUMINA, the first **N**atural **U**nderstanding benchmark for **M**ulti-dimensional **I**ntelligence and **N**umerical reasoning **A**bilities to enhance multimodal indoor perceptual understanding. NUMINA features multi-scale annotations and various question-answer pairs, generated using NUMINA-Flow, an automated annotation pipeline that integrates LLM rewriting and rule-based self-verification. We evaluate the performance of various state-of-the-art LLMs on NUMINA following the Chat-Scene framework, demonstrating that current LLMs struggle with multimodal numerical reasoning, particularly in performing precise computations such as distance and volume estimation, highlighting the need for further advancements in 3D models. The dataset and source codes can be obtained from https://github.com/fengshun124/NUMINA.

1 Introduction

Multimodal perception has gained increasing attention in recent years. Leveraging large language models (LLMs), 2D multimodal LLMs (MLLMs) have progressed rapidly, leading to groundbreaking achievements in image classification (Zeng et al., 2024, 2025; Bao et al., 2025) and reasoning (Kang et al., 2025; Yan et al., 2025). To extend the applications of MLLMs to 3D domains, researchers
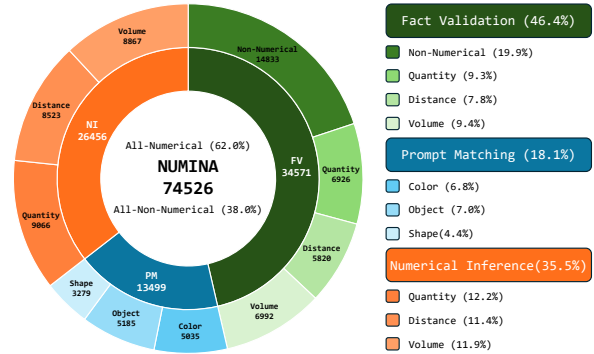

Figure 1: Statistics of the NUMINA benchmark. NUMINA is composed of non-numerical and numerical questions, where the latter are further divided into three categories with increasing difficulty: Fact Validation (FV), Prompt Matching (PM), and Numerical Inference (NI).

have strived to unlock the potential of MLLMs in spatial intelligence to enable practical functionalities, such as autonomous driving (Yang et al., 2023), indoor navigation (Coffrini et al., 2025), and route planning (Li et al., 2025). Nevertheless, understanding accurate spatial information in 3D environments remains a significant challenge due to the exponential increase in structural complexity introduced. For instance, when positioning a cabinet in a living room, one must consider not only the availability of space in the front and back but also factors such as excessive vertical space occupancy and efficient space utilization. Consequently, equipping MLLMs with spatial perception capabilities, including precise measurement and localization, is critical for improving their real-world applicability.

Achieving precise spatial reasoning necessitates well-annotated 3D vision-language datasets. However, existing benchmarks offer only coarse, global annotations, which limit their applicability for fine-grained reasoning. Datasets like ScanRefer (Chen et al., 2020), Talk2Nav (Vasudevan et al., 2021), and ScanQA (Azuma et al., 2022) advance 3D tasks by supporting object localization, navigation, and question answering, but they lack multimodal nu-

---

merical reasoning incorporating spatial precision. Unlike simpler spatial understanding tasks, numerical reasoning involves complex scene interpretation, requiring detailed annotations of object distances, sizes, and positions, which current datasets do not provide. Additionally, these benchmarks predominantly focus on textual outputs, limiting the diversity of question formats and response types, and consequently, their utility in developing models for sophisticated multimodal reasoning. This highlights the need for a more comprehensive, numerically enriched 3D vision-language benchmark.

To this end, we present NUMINA, the first **N**atural **U**nderstanding benchmark for **M**ulti-dimensional **I**ntelligence and **N**umerical reasoning **A**bilities. As shown in Figure 1, NUMINA stands out by focusing on fine-grained spatial understanding and numerical reasoning in indoor environments, featuring 74,526 question-answer (QA) pairs across three response types with increasing difficulty: Fact Validation (*judgment* questions), Prompt Matching (*multiple-choice* questions), and Numerical Inference (*value-output* questions). Each category is evaluated using tailored metrics and can be further subdivided. These tasks, spanning from simple verification to complex numerical reasoning, ensure that progress on NUMINA reflects genuine advances in 3D multimodal reasoning rather than overfitting to a single question format.

Compared with existing benchmarks, as shown in Table 1, NUMINA is the first dataset that diversely focuses on large-scale, object-level natural understanding of multi-dimensional relationships and numerical reasoning across multiple categories, offering the following key features. (1) **Comprehensive Numerical Annotations**. Building upon ScanNet (Dai et al., 2017), NUMINA introduces multidimensional labels, including object center coordinate, bounding box dimensions, and convex hull distances between objects, to enable value-based inference and facilitate the development of diverse numerical reasoning question types. (2) **Type-Rich Data**. NUMINA enhances global spatial understanding by incorporating LLM rewriting on ScanQA (Azuma et al., 2022), alongside numerical reasoning tasks. The dataset features a diverse task formats, including judgment, multiple-choice, and value-output questions, with difficulty levels ranging from basic to advanced. (3) **Automatic Annotation Pipeline**. To streamline data generation, we introduce NUMINA-Flow, an automated dataset construction pipeline. NUMINA-Flow employs LLMs

Name	#Pairs	Question Types		
		Non-Num.	Num.	Sub-Cat.
ScanRefer	41,846	✓	✗	✗
Scan2Cap	34,345	✓	✗	✗
ScanQA	30,813	✓	✗	✗
FE-3DGQA	20,215	✓	✗	✗
SQA3D	29,884	✓	✗	✗
Multi3DRefer	48,810	✓	✗	✗
CLEVER3D	171,174	✓	✗	✗
NUMINA	74,526	✓	✓	✓

Table 1: Comparison between NUMINA and existing 3D vision-language benchmarks. (Num.: Numerical Questions; Cat.: Categories)

to generate high-quality data samples with numerical ground truth. It also incorporates diverse QA templates, adopts the convex hull distance measurement that aligns with human intuition, and ensures a balanced data distribution to prevent models from exploiting shortcuts or biases.

To rigorously assess the utility of NUMINA and establish a baseline, we utilize several mainstream open-source LLMs, such as Vicuna (Chiang et al., 2023) and Qwen (Yang et al., 2024), to serve as decoders within the Chat-Scene (Huang et al., 2024) framework, enabling the simultaneous processing of 3D point clouds, 2D images, and textual inputs. Experimental results reveal substantial challenges of existing LLMs in 3D numerical inference and spatial reasoning, with accuracy falling below 3% under a 5% error Threshold Accuracy (TA@5) in distance and volume estimation tasks. While the models perform well in non-numerical evaluations, their accuracy declines considerably in distance-related tasks, achieving only about 54%, close to random selection. Notably, no single model consistently outperforms across all task categories, highlighting diverse strengths depending on the specific evaluation criteria.

2 Related Work

3D Vision-Language Understanding. Existing 3D vision-language benchmarks play a pivotal role in advancing research on vision and natural language understanding in 3D environments. ScanRefer (Chen et al., 2020) is a pioneering effort, integrating detailed textual descriptions with complex indoor scenes to facilitate 3D object localization based on natural language queries. Building on this foundation, ScanQA (Azuma et al., 2022) expands the scope by introducing diverse QA pairs to
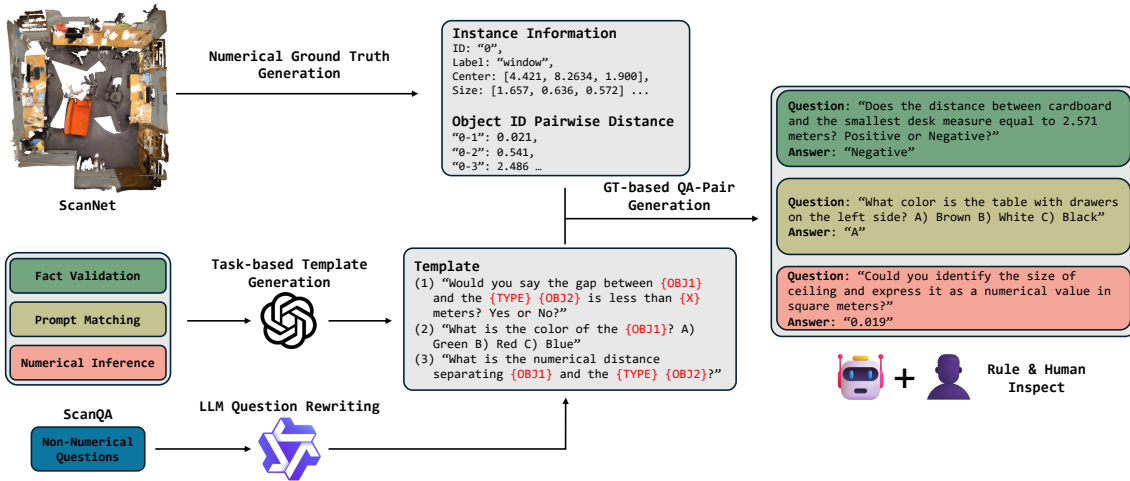
Figure 2: Overview of the NUMINA-Flow pipeline. Numerical Ground Truth (NGT) is extracted from ScanNet, including instance details and pairwise distances. GPT-4o generates diverse question templates filled with NGT, followed by rule-based and manual validation. Non-numerical questions are rewritten using Qwen2.5-72B with the ScanQA dataset for added diversity.

support various vision-language tasks, such as 3D visual grounding and 3D visual question answering (VQA). ReferIt3D (Achlioptas et al., 2020) focuses on fine-grained 3D object localization using natural language references, encompassing both real-world (Natural ReferIt3D, Nr3D) and synthetic (Synthetic ReferIt3D, Sr3D) contexts. Meanwhile, Talk2Nav (Vasudevan et al., 2021) introduces dialog-based interactions with agents, enabling the clarification of ambiguous or incomplete instructions through natural language queries, particularly for long-range navigation tasks. However, none of these benchmarks offer multi-level numerical annotations, which poses challenges for training models capable of achieving precise 3D perception. As a result, such models face limitations in applications requiring accurate measurement and localization, such as interior design and spatial planning.

Indoor Environment Perception. Indoor environment perception is an essential research field with broad applications in robotics and augmented reality (AR). Researchers have proposed various approaches leveraging diverse sensing modalities and computational methods to advance scene understanding and mapping. Vision-based techniques (Ran et al., 2021; Ruotsalainen et al., 2021), which rely on RGB or RGB-D cameras, often employ deep neural networks (DNNs) such as R-CNN (He et al., 2017) and U-Net (Ronneberger et al., 2015; Cao et al., 2022) for tasks like object detection and semantic segmentation, enabling the identification of structural components within indoor spaces (Gan et al., 2025). However, the limitations of these sys-

tems, such as visual illusions and occlusions caused by the absence of depth information, have led to the adoption of LiDAR and depth cameras. These sensors provide rich 3D spatial information, enabling methods like Cartographer (Dwijotomo et al., 2020) and KinectFusion (Izadi et al., 2011) to construct accurate geometrical representations of indoor environments. Despite notable advancements, several challenges remain. For instance, current MLLMs often lack the precise numerical reasoning capabilities required to locate objects accurately within a room. Additionally, there is a significant shortage of indoor multimodal datasets with comprehensive, multi-level numerical annotations, hindering the development of robust perception models.

3 NUMINA and NUMINA-Flow

This section describes in detail the automatic construction pipeline NUMINA-Flow for the NUMINA dataset. As illustrated in Figure 2, NUMINA-Flow generates numerical data by extracting Numerical Ground Truth (NGT) from ScanNet (Dai et al., 2017). It leverages GPT-4o (Hurst et al., 2024) to create question-answer templates, followed by rule-based and manual verification. The generated templates are validated by human evaluators to ensure clarity and ambiguity, making them comprehensible to both humans and models. Additionally, variations in expression and syntactic structure are incorporated to promote generalization across different reasoning patterns. To enhance question diversity, non-numerical questions are further generated using Qwen2.5-72B (Yang et al., 2024).
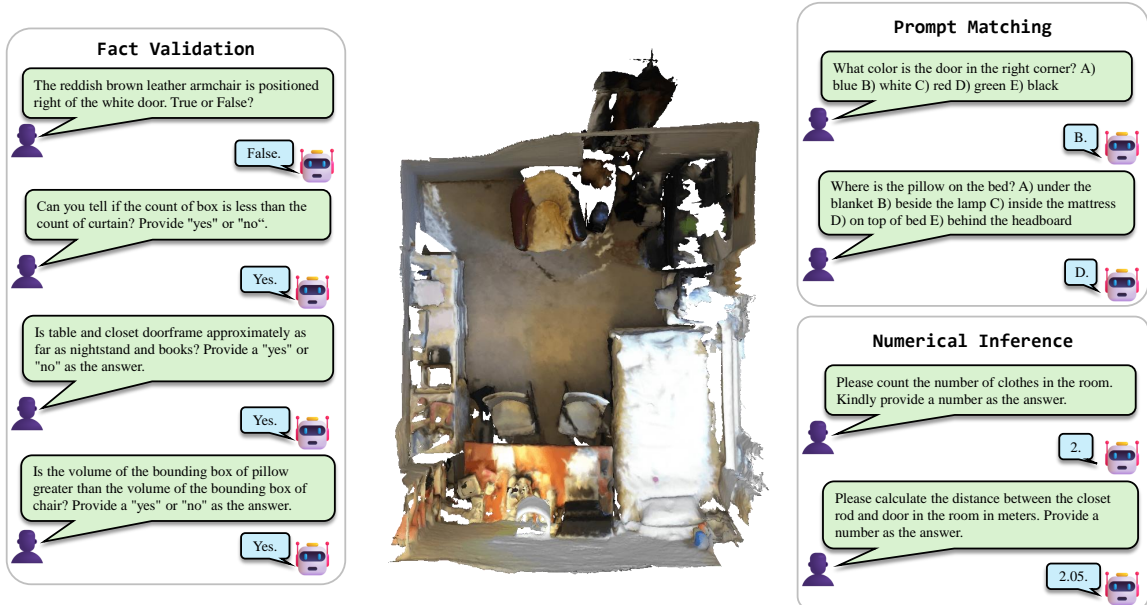
Figure 3: Example of various scene understanding and numerical reasoning tasks in NUMINA dataset. All tasks are formulated as single-turn question-answering pairs without the use of additional task-specific heads, ensuring a unified and consistent evaluation framework.

3.1 Dataset Construction

To bootstrap the dataset construction process, we select the indoor 3D scenes from the existing Scan-Net (Dai et al., 2017) dataset. Building on these complex indoor environments, NUMINA expands the scope of questions beyond text-only answers. Three distinct tasks, designed with varying levels of difficulty, are defined as follows. Examples of these tasks are shown in Figure 3.

- **Fact Validation (FV):** Binary classification questions requiring "yes" or "no" responses to evaluate factual accuracy.

- **Prompt Matching (PM):** Multiple-choice questions requiring the selection of the correct answer from five options, demanding stronger spatial comprehension.

- **Numerical Inference (NI):** The most challenging task requiring precise numerical outputs for quantities, volumes, or distances.

To enhance diversity beyond numerical reasoning, 14,000 QA pairs from ScanQA (Azuma et al., 2022) covering color, shape, and spatial relationships are integrated after rewriting with Qwen2.5-72B (Yang et al., 2024), with diverse templates in Appendix A. This yields the dataset of 74,526 total QA pairs, containing 46,194 numerical QA pairs (62.0%) across quantity, distance, and volume subcategories, plus 28,332 non-numerical pairs (38.0%) from ScanQA, as shown in Figure 1.

3.2 Numerical Ground Truth Extraction

Based on the original annotations, we calculate crucial geometric parameters for each instance in the scene, including centroid coordinates, bounding box dimensions, and the minimum and maximum values along the (x, y, z) axes. Furthermore, we calculate the convex hull distance (Bentley et al., 1982) for every pair of instances. This metric, defined as the shortest distance between the outer boundaries of two point sets determined by their convex hulls, provides a precise representation of relative positional relationships, especially for complex shapes, as it accounts for both the center points and the overall geometry. We adopt this distance due to its strong alignment with human visual perception (Bentley et al., 1982). To represent the volume of irregular point cloud objects, we utilize the smallest axis-aligned bounding box, a rectangular box aligned with the coordinate axes that fully encloses the object. The calculated pairwise convex hull distances and bounding box dimensions are systematically organized and form the basis for generating QA pairs in subsequent stages.

3.3 QA-Pair Generation

Our QA-pair generation process follows a systematic three-stage pipeline designed to ensure linguistic diversity, factual accuracy, and balanced difficulty distribution.

Step 1: Template Generation and Selection. To enhance linguistic diversity and question complexity, we employ GPT-4o (Hurst et al., 2024) to generate ten distinct question templates for each task category. These templates incorporate varied syntactic structures and expressions to promote generalization across different reasoning patterns. During question generation, templates are randomly sampled to ensure diversity, with reserved placeholders systematically replaced by NGT values and object references from the target scene.

Step 2: Task-specific Formatting. Each task type requires specific formatting to ensure clear evaluation criteria. For Fact Validation (FV) tasks, we append explicit prompts such as "*Is the statement Yes or No?*" to ensure binary output. Prompt Matching (PM) tasks include constraint-oriented instructions guiding models to select from five given options (from A to E). Numerical Inference (NI) tasks explicitly request precise numerical values with appropriate units.

Step 3: Bias Mitigation and Balance. To prevent systematic biases, we implement several balancing strategies: (1) For PM tasks, correct answers are uniformly distributed across all five options (A-E) to prevent positional bias; (2) For FV tasks, "yes" and "no" answers are equally represented; (3) Template selection follows uniform random sampling to ensure no linguistic patterns dominate the dataset.

Furthermore, to expand dataset diversity beyond purely numerical reasoning, we integrate 14,000 non-numerical QA pairs from ScanQA (Azuma et al., 2022) covering color, shape, and spatial relationships. These questions undergo systematic rewriting using Qwen2.5-72B (Yang et al., 2024) to conform to our FV and PM task formats, ensuring consistency with our evaluation framework while preserving their original semantic content.

The complete pipeline produces 74,526 QA pairs with progressive difficulty levels across all categories. Detailed templates and examples for each task type are provided in Appendix A.

3.4 Rule-based and Human Inspection

To ensure NUMINA's accuracy and consistency, outputs from the automatic NUMINA-Flow pipeline undergo a two-stage validation. First, a rule-based framework detects issues in Qwen2.5-generated questions, such as hallucinations or incorrect option counts, and regenerates faulty items up to five times. Questions must meet all inspection criteria before passing this stage. Second, five human evaluators conduct a detailed review of 20,000 randomly selected samples, checking for grammar, logic, expression, and unit consistency, achieving a correctness rate of 99.5%, validating the quality and reliability of the NUMINA dataset.

All aforementioned annotation processes, except human inspection, are consolidated and integrated into the single NUMINA-Flow annotation pipeline. This pipeline enables practitioners to replicate and further improve our annotation procedure or modify specific portions of the code to create customized datasets tailored to their needs. The detailed process of the QA-pair generation is presented in Appendix B.

4 Experiments

4.1 Model Architecture

To assess data quality and task difficulty, we conduct our 3D-language VQA experiments following the Chat-Scene (Huang et al., 2024) paradigm. As illustrated in Figure 4, a pre-trained detector segments the point cloud into individual objects, which are then projected onto multi-view images using corresponding masks. Object-centric features are extracted using tailored 3D Mask3D (Schult et al., 2023) and 2D DINO-V2 (Oquab et al., 2023) encoders, processed through projection layers, and integrated with unique object identifiers (IDs) to form a sequence of object-level embeddings. Finally, given the object-level embeddings and question prompts, the LLM generates a concise response either True or False, a multiple-choice option, or a numerical value based on the task type.

4.2 Evaluation Metrics

Since the outputs in all three tasks of NUMINA are straightforward and unambiguous, we define distinct metrics for each task separately. For the Fact Validation (FV) and Prompt Matching (PM) tasks, we use accuracy as the evaluation metric, where a response is considered correct only if the prediction exactly matches the ground truth. For the Numerical Inference (NI) task, accurately measuring the distance between two objects or the size of a specific item presents a significant challenge for the model. Therefore, we employ Threshold Accuracy (TA) as the assessment metric, which measures the model's performance within specific pre-defined thresholds. Formally, the TA metric
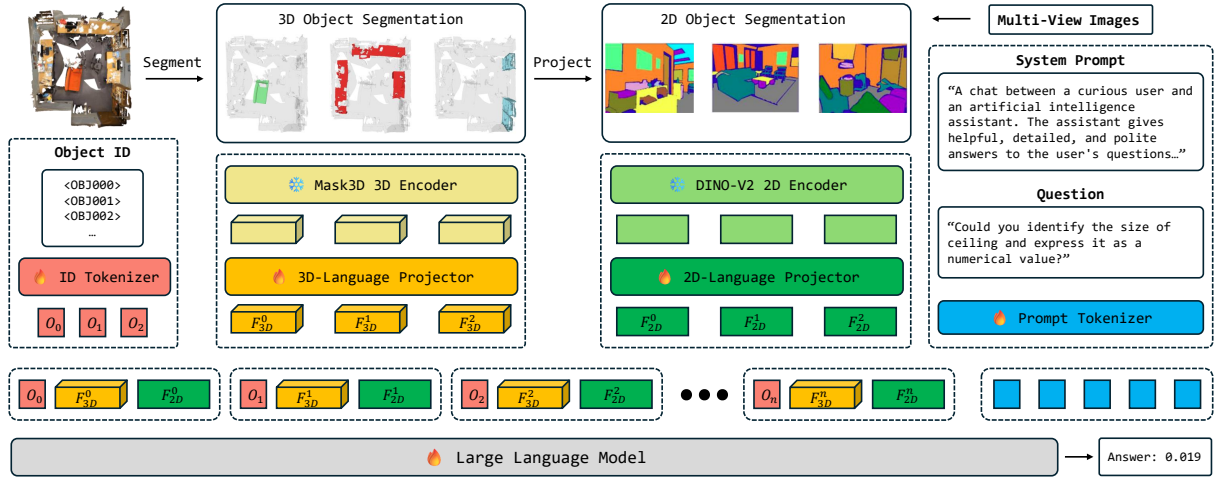
Figure 4: Overall Chat-Scene framework. The framework processes 3D scenes through a multi-stage pipeline: (1) scene decomposition into object segments, (2) mapping of segments to multi-view images via corresponding masks, (3) extraction of object-centric representations using dedicated 3D and 2D encoders, and (4) combination of these representations with object identifiers to generate scene embeddings as sequences of object-level embeddings. These embeddings serve as input to the language model component. For evaluation on the NUMINA benchmark, we substitute the original large language model with open-source alternatives including Vicuna, Mistral, Qwen, and Phi to assess their numerical reasoning capabilities.

can be expressed as follows:

$$\text{TA} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left(|d_i^{\text{pred}} - d_i^{\text{true}}| < \text{Threshold} \right),$$

(1)

where d_i^{pred} represents the model's prediction, and d_i^{true} denotes the ground truth. The indicator function $\mathbf{1}$ evaluates to 1 if the specified condition is satisfied and 0 otherwise. Threshold is a predefined value that limits the allowable error between the prediction and the ground truth. In our case, we set the threshold values as 5%, 10%, 20% corresponding to TA@5, TA@10, TA@20, respectively.

4.3 Experimental Setup

We adhere to the original ScanNet (Dai et al., 2017) split, using the training set for training and the validation set for evaluation, as the test set annotations are not publicly available. This protocol is consistent with prior works such as ScanQA (Azuma et al., 2022). All tasks are formulated within a user–assistant interaction framework. Joint training is performed using the Cross-Entropy loss function and the AdamW optimizer, enabling parameter updates across unfrozen components of the model, including the tokenizer, projector, and LLMs.

We adopt six open-source LLMs in our experiments: Vicuna 7B v1.5, Vicuna 13B v1.5 (Chiang et al., 2023), Qwen2.5-7B (Yang et al., 2024), Phi 4-mini (Abouelenin et al., 2025), Mistral-7B

(Instruct-v0.3, Jiang et al., 2023) and DeepSeek-R1-Dstill-Qwen-7B (Guo et al., 2025). To fine-tune these models, we apply LoRA (Hu et al., 2021) with a rank of 16. The learning rate is set to 2e-6, following a cosine annealing schedule. The training process spans three epochs and is completed using 4 NVIDIA A100 Tensor Core GPUs.

4.4 Experiment Results

Table 2 presents the performance of the six LLMs across various NUMINA subtasks. In Prompt Matching (PM), all models achieve high accuracy on non-numerical tasks, ranging from 82.19% to 85.70%. Similarly, in Fact Validation (FV), models perform well on non-numerical assessments but exhibit a significant drop in accuracy for distance-related tasks, with scores around 54%, indicating challenges in understanding convex hull distances. Numerical Inference (NI) proves to be the most difficult category, particularly in distance and volume estimation, where accuracy remains below 3%. Among the models, Vicuna 13B v1.5 achieves the highest average score in numerical inference, though it only slightly outperforms Vicuna 7B v1.5 by 1.26%. These results highlight the limitations of current LLMs in fine-grained 3D spatial reasoning, particularly in tasks requiring precise numerical analysis. For Qwen, Phi, Mistral, and the latest DeepSeek-R1 distilled models, their performance has been unsatisfactory. Overall, no single model

LLM	Prompt Matching	Fact Validation				Numerical Inference		
	Non-Num	Non-Num	Quantity	Distance	Volume	Quantity	Distance	Volume
Vicuna 7B v1.5	83.35	77.31	71.69	53.89	83.44	8.01	1.73	2.40
Vicuna 13B v1.5	85.12	76.17	71.30	54.35	83.66	10.15	2.71	3.06
Phi 4-mini	84.79	75.24	65.23	51.62	80.36	10.36	1.21	3.17
Mistral 7B	84.75	75.88	71.97	55.31	83.14	8.49	2.54	3.22
Qwen2.5-7B	85.70	75.34	69.76	54.29	82.95	10.63	1.21	1.80
DeepSeek-R1-Distill-Qwen-7B	82.19	73.11	69.59	52.70	81.81	9.99	2.02	2.67

Table 2: Performance comparison of different LLMs on the NUMINA benchmark for different kinds of problems. PM and FV are evaluated by Accuracy. NI is assessed by the Threshold Accuracy (TA@5), which measures the proportion of predictions falling within a 5% error margin with respect to the ground truths.

excels in all numerical sub-tasks. These six large models exhibit varying strengths and weaknesses depending on the specific tasks and problems to which they are applied. Figure 3 provides some illustrative examples of the three tasks for further clarification.

4.5 Analysis

Current MLLMs exhibit strong performance on purely linguistic and non-numerical spatial tasks (Prompt Matching and non-numeric Fact Validation, both >75% accuracy) but break down when asked to perform fine-grained numerical reasoning, especially in 3D settings. Architecturally, Transformers lack inductive biases for geometry: they treat numbers as discrete tokens without true magnitude awareness and are not designed to encode spatial relationships such as distances, angles, or volumes. Moreover, standard pretraining corpora of LLMs seldom include explicit 3D-spatial supervision or grounded numerical examples, leaving models underexposed to the types of geometric computations required by NUMINA.

Even in emerging multimodal variants, MLLMs typically fuse visual inputs only via generic embeddings, without dedicated modules for point-cloud processing or multi-view depth estimation. As a result, tasks like convex-hull distance estimation and precise volume calculation, which demand multi-step geometric reasoning and spatial continuity, fall outside the models' learned capabilities. The anomalously high "volume" accuracy (83%) likely reflects reliance on commonsense size priors (e.g., "*monitors are larger than phones*") rather than true geometric inference.

These findings suggest that improving numerical reasoning in MLLMs requires fundamental architectural changes beyond simply scaling model size. The consistent poor performance across all tested models (including recent architectures like

LLM	Type	Fact Validation		
		Quantity	Distance	Volume
Vicuna 7B v1.5	Ori	74.47	65.52	63.64
	CP	73.56	67.06	66.32
Vicuna 13B v1.5	Ori	73.61	64.69	64.88
	CP	74.47	66.23	63.84
Qwen2.5-7B	Ori	72.49	67.18	72.11
	CP	74.89	69.08	72.31

Table 3: Logical consistency ablation study. "Ori" refer to original QA-pair and "CP" indicates the corresponding contrapositive ones.

Phi4 (Abouelenin et al., 2025) and DeepSeek-R1 (Guo et al., 2025) indicates these limitations reflect deeper constraints in how current transformer architectures process spatial and numerical information, pointing toward the need for specialized geometric reasoning modules and grounded training paradigms that incorporate explicit 3D spatial supervision.

4.6 Ablation Study

Logical Consistency. The Fact Validation (FV) task includes two sets of questions with corresponding answers designed to evaluate semantic understanding, common sense reasoning, and logical coherence. For example, in one QA pair, the original question asks, "*Is it correct that the sofa is bigger than the cell phone?*", with the expected answer being "*Yes.*" In the corresponding contrapositive pair, the question states, "*The sofa is smaller or equal than the cell phone*", with the correct answer being "*No.*" By systematically altering relational statements and their corresponding answers, we test the model's ability to maintain logical consistency and accurately interpret semantic relationships. Details are in Appendix D.

Table 3 summarizes the logical consistency evaluation results of three models in the FV task. The quantitative analysis reveals that the average performance discrepancy between original and con-

LLM	Fact Validation		
	Quantity	Distance	Volume
Vicuna 7B v1.5	63.41	53.76	80.72
+ CoT	66.61	56.93	79.85
Vicuna 13B v1.5	70.14	54.55	83.12
+ CoT	72.20	55.34	79.74

Table 4: Chain-of-Thought study on fact validation.

LLM	Numerical Inference (Volume)		
	TA@5	TA@10	TA@20
Vicuna 7B v1.5	2.62	4.91	8.41
+ CoT	2.95	4.20	11.95
Vicuna 13B v1.5	2.67	4.48	8.57
+ CoT	3.88	5.46	9.39

Table 5: Chain-of-Thought study on volume numerical inference task. TA@5/10/20 indicates Threshold Accuracy at error levels of 5%, 10%, and 20%.

trapositive QA-pairs across these models is 1.71%, 1.15%, and 1.50%, respectively. Notably, Vicuna 13B v1.5 exhibits superior performance stability compared to other models, which can be reasonably attributed to its enhanced model capacity with 13 billion parameters. Experimental findings demonstrate that all evaluated LLMs exhibit robust logical reasoning capabilities, which consistently produce accurate judgments, even when the logical structure of both the questions and corresponding answers is reversed. This outcome highlights the models' resilience to logical perturbations and their capacity to generalize reasoning patterns beyond fixed, predefined formats. Furthermore, model scale appears to provide a stronger baseline for performance. However, the training methodology and the nature of the source data, such as differences between Vicuna and Qwen, play a critical role in enabling more complex forms of reasoning. These factors suggest that both architectural scale and training data composition significantly influence the depth and flexibility of logical reasoning in LLMs.

Effects of Chain-of-Thought. In many natural language tasks, the Chain-of-Thought (CoT) paradigm has been shown to enhance reasoning by guiding models through step-by-step problem-solving processes. To investigate its effectiveness in value reasoning tasks, we adapted FV and volume-related tasks of the NI problems to the CoT framework. Specifically, instead of providing short numerical answers, we incorporate structured intermediate reasoning steps leading to the final answers. For instance, in volume-related questions, the CoT paradigm first extracts the 3D bounding

box data of the specified object. It then computes the object's volume by multiplying the bounding box dimensions, ultimately deriving the final size. This structured reasoning process aims to improve the model's numerical inference ability. A concrete example can be found in Appendix E.

Tables 4 and 5 present the impact of CoT reasoning on Vicuna 1.5 models across FV and NI tasks. The results reveal mixed effectiveness of CoT across different model scales and task types. For the 7B model, CoT demonstrates consistent improvements in quantity (+3.20%) and distance (+3.17%) estimation within FV tasks, while showing minimal degradation in volume estimation (-0.87%). Conversely, the 13B model exhibits more variable performance: improvements in quantity (+2.06%) and distance (+0.79%) tasks are accompanied by a notable decline in volume estimation (-3.38%). In the NI task, both models show modest improvements with CoT. The 7B model achieves gains across all threshold levels, with the most substantial improvement at TA@20 (+3.54%). Similarly, the 13B model demonstrates improvements at TA@5 (+1.21%) and TA@10 (+0.98%), though with a slight decrease at TA@20 (-1.18%).

The varied effectiveness of CoT can be attributed to a complex interplay between task requirements and the reasoning process it induces. CoT demonstrates significant utility for tasks amenable to logical decomposition, such as quantity and distance estimation, by enforcing a structured, procedural pathway. In contrast, its efficacy diminishes for tasks that rely on holistic, commonsense heuristics, as exemplified by the performance degradation in volume fact validation. This negative impact likely stems from CoT's explicit reasoning steps interfering with the models' more accurate, pre-existing intuitive knowledge, an effect amplified in the larger 13B model, whose robust implicit knowledge is more susceptible to such disruption.

5 Conclusion

We introduce NUMINA, a Natural Understanding Benchmark for Multi-dimensional Intelligence and Numerical Reasoning Abilities, which is designed to advance multimodal indoor perceptual understanding. NUMINA provides a diverse dataset with multi-scale annotations and a variety of question-answer pairs spanning multiple task types. This dataset is generated using NUMINA-Flow, an automated annotation platform that leverages LLM

rewriting and rule-based self-verification. Following the Chat-Scene paradigm, we replace LLMs within the architecture and jointly train and evaluate all tasks. Experimental results indicate that current LLMs exhibit significant limitations in multimodal numerical reasoning, particularly in tasks requiring fine-grained computational precision, such as quantity, distance, and volume estimation. In the future, we will develop further advancements in numerical and spatial reasoning for multimodal AI systems.

Limitations

While the NUMINA benchmark represents a significant advancement in the assessment of multidimensional intelligence and numerical reasoning, several limitations warrant consideration. First, the complexity of the benchmark necessitates substantial computational resources, which may pose accessibility challenges for researchers with limited hardware capabilities. Second, although the benchmark encompasses a wide range of tasks, it may not fully capture all aspects of human cognition, particularly those dimensions that are inherently subjective or difficult to quantify, such as emotional intelligence and creativity. Finally, although the evaluation framework is designed to be comprehensive, the interpretation of outcomes may still be influenced by subjective decisions made during the task selection and formulation process. These limitations should be carefully considered, and future iterations of the benchmark could benefit from addressing these concerns to enhance its applicability and robustness.

Ethical Considerations

We discuss the following ethical considerations related to our NUMINA dataset: (1) **Intellectual Property.** The ScanQA dataset is distributed with the CC BY-NC-SA 3.0 license[1], and the ScanNet dataset is also available for research use[2]. (2) **Worker Treatments.** We hired five annotators and fairly pay them according to agreed salaries and workloads. (3) **Intended Use.** NUMINA can be utilized to develop more persuasive 3D reasoning models. Researchers can also inherit our dataset design and utilize NUMINA-Flow to develop their own datasets. (4) **Controlling Potential Risks.** Since

the documents of NUMINA do not contain private information and annotating this dataset is not necessary to make many judgements about social risks, we believe NUMINA does not introduce any additional risks. We manually verified some randomly sampled data to ensure the dataset did not contain risky issues. (5) **AI Assistance.** The writing of this paper adopts ChatGPT in refining some sentences.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

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

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. 2022. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139.

Muyi Bao, Changyu Zeng, Yifan Wang, Zhengni Yang, Zimu Wang, Guangliang Cheng, Jun Qi, and Wei Wang. 2025. Ftcformer: Fuzzy token clustering transformer for image classification. *arXiv preprint arXiv:2507.10283*.

Jon Louis Bentley, Franco P Preparata, and Mark G Faust. 1982. Approximation algorithms for convex hulls. *Communications of the ACM*, 25(1):64–68.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 202–221. Springer.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 3(5).

---

[1] https://creativecommons.org/licenses/by-nc-sa/3.0/

[2] https://github.com/ScanNet/ScanNet

Alberto Coffrini, Mohammad Amin Zadenoori, Paolo Barsocchi, Francesco Furfari, Antonino Crivello, and Alessio Ferrari. 2025. Toward a method for llm-enabled indoor navigation. *arXiv preprint arXiv:2503.11702*.

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

Abdurahman Dwijotomo, Mohd Azizi Abdul Rahman, Mohd Hatta Mohammed Ariff, Hairi Zamzuri, and Wan Muhd Hafeez Wan Azree. 2020. Cartographer slam method for optimization with an adaptive multi-distance scan scheduler. *Applied Sciences*, 10(1):347.

Hong-Seng Gan, Muhammad Hanif Ramlee, Zimu Wang, and Akinobu Shimizu. 2025. A review on medical image segmentation: Datasets, technical models, challenges and solutions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(1):e1574.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. 2024. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. 2011. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Xiaoqiang Kang, Zimu Wang, Xiaobo Jin, Wei Wang, Kaizhu Huang, and Qiufeng Wang. 2025. Template-driven llm-paraphrased framework for tabular math word problem generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24303–24311.

Kechen Li, Yaotian Tao, Ximing Wen, Quanwei Sun, Zifei Gong, Chang Xu, Xizhe Zhang, and Tianbo Ji. 2025. Gridroute: A benchmark for llm-based route planning with cardinal movement in grid environments. *arXiv preprint arXiv:2505.24306*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

T Ran, L Yuan, and JB Zhang. 2021. Scene perception based visual navigation of mobile robot in indoor environment. *ISA transactions*, 109:389–400.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Laura Ruotsalainen, Aiden Morrison, Maija Mäkelä, Jesperi Rantanen, and Nadezda Sokolova. 2021. Improving computer vision-based perception for collaborative indoor navigation. *IEEE Sensors Journal*, 22(6):4816–4826.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. 2023. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2021. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision*, 129:246–266.

Shuo Yan, Ruochen Li, Ziming Luo, Zimu Wang, Daoyang Li, Liqiang Jing, Kaiyu He, Peilin Wu, George Michalopoulos, Yue Zhang, et al. 2025. Lmr-bench: Evaluating llm agent's ability on reproducing language modeling research. *arXiv preprint arXiv:2506.17335*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. 2023. Llm4drive: A survey of large language models for autonomous driving. *arXiv preprint arXiv:2311.01043*.

Changyu Zeng, Zimu Wang, Jimin Xiao, Anh Nguyen, Kaizhu Huang, Wei Wang, and Yutao Yue. 2024. Monotcm: Semantic-depth fusion transformer for monocular 3d object detection with token clustering and merging. In *International Conference on Neural Information Processing*, pages 332–346. Springer.

Xinyi Zeng, Zimu Wang, Haiyang Zhang, Yiming Luo, and Wei Wang. 2025. Skin disease classification with lvlms: An empirical study. In *2025 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 2304–2309. IEEE.

A Templates for Each Category

Table 6 presents the QA-pair template for each task type in NUMINA. In this framework, <OBJX> within the question is substituted with actual object names present in the given scenario, while the corresponding answer is generated based on the annotated dataset. This ensures consistency and accuracy in the question-answer pairs while maintaining adaptability across different scenarios.

Each category assesses different aspects of an AI model's reasoning capabilities related to object properties, spatial relationships, and numerical estimations. Fact Validation (highlighted in red) focuses on binary (yes/no) reasoning about object comparisons, including quantity differences, relative distances, and bounding box size comparisons. Prompt Matching (highlighted in green) evaluates the model's ability to correctly interpret and match prompts to predefined categorical choices, such as identifying an object's color, spatial positioning, or orientation. Numerical Inference (highlighted in blue) measures the model's ability to provide precise numerical responses, including object counting, distance estimation, and volume computation. These templates serve as standardized benchmarks to assess LLMs' spatial reasoning, object recognition, and numerical estimation accuracy in indoor environments.

B Details of Dataset Construction

The construction of the dataset follows a dual-track generation pipeline, NUMINA-Flow, which comprises independent LLM-based and rule-based methods. Each track operates separately, contributing contrapositive strengths to ensure high-quality, factually grounded question-answer pairs. The LLM-based track focuses on rewriting Short-Answer Questions (SAQs) from ScanQA into structured formats, while the rule-based track generates questions directly from annotated scene data. Together, these methods enhance linguistic fluency, maintain logical consistency, and minimize hallucinations.

B.1 LLM-Based Generation

The LLM-based track transforms SAQs into other question types, including Prompt Matching (PM) and Fact Validation (FV) Questions. This process leverages the generative capabilities of large language models while enforcing structural constraints to ensure output consistency and reliability. Key components of this track include:

- **Structured Prompting:** The LLM is guided by carefully designed prompts that explicitly define the rewriting task, specify output constraints in JSON format, and provide illustrative examples. These constraints ensure that the reformulated questions maintain clarity and coherence while adhering to strict formatting guidelines.

- **Rigorous Validation:** Post-generation, an automated validation process verifies key structural properties, such as the correct number of options, proper answer placement, and logical distinctiveness of distractors (for PM) or alternative statements (for FV). Any outputs failing these criteria are discarded and regenerated to maintain dataset integrity.

To ensure consistency and high-quality question generation, the system employs predefined prompts tailored for different question types. Tables 7 and 8 illustrate the structured LLM prompts used for PM and FV generation, respectively.

B.2 Rule-Based Generation

Independent of the LLM-based track, the rule-based track constructs questions directly from structured scene annotations using deterministic procedures. This method ensures factual accuracy by deriving question-answer pairs solely from observable, verifiable data. The rule-based track consists of the following steps:

Fact Validation
Q1: Are there fewer <OBJ1> than <OBJ2>? Please reply with a "yes" or "no" only.
A1: yes
Q2: Is the distance between <OBJ1> and <OBJ2> greater than the distance between <OBJ3> and <OBJ4>?
A2: no
Q3: Is the size of the bounding box of <OBJ1> less than the one of <OBJ2>? Select "yes" or "no" as the answer.
A3: yes
Prompt Matching
Q1: What color is the <OBJ1>? A) Red B) Blue C) Green D) Yellow E) Black
A1: B
Q2: What is to the left of the <OBJ1>? A) Armchair B) Bookshelf C) Lamp D) Painting E) Couch
A2: E
Q3: What is the <OBJ1> facing? A) Bookshelf B) Window C) Door D) Round table E) Fireplace
A3: D
Numerical Inference
Q1: Please count the number of <OBJ1> in the room. Give a number as the answer.
A1: 2
Q2: Please estimate the distance between the <OBJ1> and <OBJ2> in the room in meters. Give a numerical response.
A2: 1.04
Q3: Can you estimate the volume of the bounding box of <OBJ1> in cubic meters? Give a numerical response.
A3: 3.83

Table 6: Templates for each category in NUMINA.

- **Candidate Selection:** The pipeline selects objects or object pairs that meet predefined criteria, such as bounding box sizes, or numerical attributes. Non-informative labels (e.g., item, object) are filtered out to ensure relevance and diversity in the generated questions.

- **Automated Question Formation:** Once appropriate candidates are identified, the pipeline applies structured templates and logical rules to generate FV or NI questions. These templates guarantee consistency and accuracy in queries while keeping them diverse and reflective of real-world data.

- **Grounding in Factual Data:** Since this track draws directly on scene metadata (e.g., object labels, spatial relationships, and numerical attributes), the resulting questions are inherently grounded in real, observable data, minimizing ambiguity and eliminating the risk of hallucinated content.

By operating as independent yet contrapositive tracks, the LLM-based and rule-based approaches provide a balanced mechanism for dataset construction. The LLM-based track enhances linguistic naturalness while adhering to strict structural constraints, whereas the rule-based track guarantees factual accuracy through deterministic question generation. This dual-track framework ensures that the resulting dataset is both diverse and reliable, making it well-suited for evaluating AI models in spatial reasoning and numerical inference tasks.

C System Prompt

Table 9 presents the system prompt provided to the LLM, which defines its role, outlines the task requirements, and specifies the meanings of "distance" and "volume" used in the subsequent questions. This ensures standardization in spatial reasoning tasks within the NUMINA dataset, promoting uniform interpretation and evaluation of distance and volume-related queries. However, the LLM struggles to accurately comprehend the concept of convex hull distance, likely due to the complexity of its calculation. As a result, the model performs poorly on all distance-related tasks, highlighting its limitations in numerical spatial reasoning.

D Logical Consistency Question-Answer Pairs

Table 10 presents a structured set of logical consistency QA-pairs in NUMINA, designed to assess the reasoning capabilities of language models. Each original question (Ori-Q) is accompanied by its corresponding original answer (Ori-A), establishing a baseline response. Additionally, a contrapositive or logically equivalent question (CP-Q) is introduced to verify coherence in reasoning, with its expected logically consistent answer (CP-A).

The QA pairs encompass comparisons involv-

Introduction
Please rewrite a Short Answer Question (SAQ) into a Prompt Matching (PM) format with `n` options.
Input
- Original SAQ: <Original Question>
- Original Answer: <Original Answer>
- Expected Correct Option: <Correct Option Label>
- Number of Options: n
Task Description
1. Convert the SAQ into a clear and concise PM format.
2. Generate n-1 incorrect options as distractors:
- Keep distractors in the same category as the correct answer (e.g., if the answer is a noun, all distractors should be nouns).
- Ensure distractors are plausible but incorrect.
- Avoid synonyms, overly obvious wrong choices, or options that reveal the correct answer.
3. Place the correct answer exactly as specified (including any spelling variations) in the designated option label.
4. Format the options as follows:
- "A) Option A B) Option B C) Option C ..."
- Ensure the correct answer appears at the predefined correct option label.
5. Include a hint instructing the user to answer using the correct option letter.
Output Format
Return only a JSON object with the following keys:
{
"question": "The rewritten PM",
"Answer": "The correct option label (e.g., 'A')"
}
Do not include any additional text or explanations outside this JSON object.
Example Formatting
Input:
- SAQ: "What is the capital of France?"
- Answer: "Parris" (note the misspelling)
- Expected Correct Option: "B"
- Number of Options: 4
Expected Output:
{
"question": "What is the capital of France? Answer using the correct option letter.
A) Berlin B) Parris C) London D) Rome",
"Answer": "B"
}

Table 7: LLM Prompt for PM question Generation.

Introduction
Please rewrite a Short Answer Question (SAQ) into a Fact Validation (FV) format and also generate its contrapositive version.
Input
- Original SAQ: <Original Question>
- Original Answer: <Original Answer>
- Boolean Indicator: <True or False>
- Answer Options: "<Affirmative Word>" (for True) / "<Negative Word>" (for False)
Task Description
1. Convert the SAQ into a clear, factual statement incorporating the given answer.
2. Rewrite the statement into a FV based on the Boolean Indicator:
- If `True` → Keep the statement affirmative (e.g., `"Bob sits next to Alice."`).
- If `False` → Negate the statement (e.g., `"Bob does not sit next to Alice."`).
3. Append the answer options at the end as:
- "Is this correct? Answer with <Affirmative Word> or <Negative Word>."
4. Generate a contrapositive question by logically inverting the original statement and answer.
5. Ensure that both original and contrapositive questions are logically and grammatically correct.
Output Format
Return only a JSON object with the following keys:
{
"question": "The rewritten FV question",
"Answer": "The answer option corresponding to the preset Boolean indicator",
"cp_question": "The contrapositive version of the FV",
"cp_answer": "The answer option corresponding to the negation of the preset Boolean indicator"
}
Do not include any additional text or explanations outside this JSON object.
Example Formatting
Input:
- SAQ: "Who sits next to Alice?"
- Answer: "Bob"
- Boolean Indicator: False
- Answer Options: "<Affirmative Word>" (True) / "<Negative Word>" (False)
Expected Output:
{
"question": "Bob does not sit next to Alice. Is this correct? Answer with <Affirmative Word> or <Negative Word>.",
"Answer": "<Negative Word>",
"cp_question": "Bob sits next to Alice. Is this correct? Answer with <Affirmative Word> or <Negative Word>.",
"cp_answer": "<Affirmative Word>"
}

Table 8: LLM Prompt for FV Generation.

You are an AI Assistant providing accurate, detailed, and polite explanations. Your goal is to give clear and helpful responses. Distances refer to convex hull distances, the shortest distance between points on the convex hulls of two objects, where a convex hull is the smallest convex shape enclosing an object. Volumes refer to the space within an object's minimum axis-aligned bounding box (AABB), a rectangular box aligned with the coordinate axes. Always use these definitions for distances and volumes.

Table 9: System Prompt of NUMINA.

Ori-Q: Are there fewer kitchen counter than table? Reply with a "yes" or "no" only.
Ori-A: yes
CP-Q: Can you tell if the count of kitchen counter is greater than or equal to the count of table? Reply with a "yes" or "no" only.
CP-A: no
Ori-Q: Is the distance between tissue box and guitar greater than the one between refrigerator and sink? Offer a "yes" or "no" as the answer.
Ori-A: no
CP-Q: Is the distance between tissue box and guitar less than or equal to the distance between refrigerator and sink? Offer a "yes" or "no" as the answer.
CP-A: yes
Ori-Q: Is the size of the bounding box of kitchen counter greater than or equal to the one of door? Select "yes" or "no" as the answer.
Ori-A: yes
CP-Q: Can you tell if the volume of the bounding box of kitchen counter is less than the volume of the bounding box of door? Select "yes" or "no" as the answer.
CP-A: no

Table 10: Templates for logical consistency QA-Pairs in NUMINA. Each original question (Ori-Q) is followed by its corresponding original answer (Ori-A), while a logically equivalent or contrapositive question (CP-Q) is provided to verify coherence in reasoning, with its expected logically consistent answer (CP-A). such contrary-logical QA-pairs aids in evaluating the reliability and robustness of language models in maintaining logical consistency.

ing quantities, distances, and bounding box sizes, where the contrapositive questions are framed to test whether the model maintains consistency when restating or reversing relational statements (e.g., "*greater than*" vs. "*less than or equal to*"). This framework ensures the reliability and robustness of language models by systematically evaluating their logical consistency across related queries.

E Chain of Thoughts Question-Answer Pairs

Table 11 presents a set of question-answer (QA) pairs designed to illustrate the application of Chain-of-Thought (CoT) reasoning in the NUMINA dataset. Each original question (Ori-Q) is accompanied by a direct answer (Ori-A), followed by a reformulated question incorporating CoT reasoning (CoT-Q) and its corresponding answer (CoT-A). The CoT-based responses provide step-by-step logical explanations to enhance interpretability and reasoning transparency.

The QA pairs cover various spatial reasoning and numerical estimation tasks, including volume computation, object counting, and comparative distance assessments. For instance, in the volume estimation task, the CoT response explicitly details the bounding box dimensions and the multiplication process to derive the final volume, reinforcing numerical reasoning. Similarly, in object counting

tasks, the CoT response explicitly states the numerical values being compared before concluding the answer. Distance-based reasoning follows a similar pattern, where spatial distances are quantitatively analyzed before providing the final decision.

Ori-Q: Can you estimate the volume of the bounding box of bed in cubic meters? Give a numerical response.
Ori-A: 3.83
CoT-Q: Can you estimate the volume of the bounding box of bed in cubic meters? Please solve the problem step by step. Show each intermediate thought process clearly and provide the final answer after completing the reasoning process.
CoT-A: Given the bounding box dimensions of the object along the X, Y, and Z axes as 1.98 m, 2.32 m, and 0.83 m respectively, the volume of the bounding box is calculated as (length x width x height) yielding approximately 3.83 cubic meters.
Ori-Q: Are there fewer stool than cabinet? Please reply with a "yes" or "no" only.
Ori-A: yes
CoT-Q: Are there fewer stool than cabinet? Please solve the problem step by step. Show each intermediate thought process clearly and provide the final answer after completing the reasoning process.
CoT-A: Given the count of stool as 4 and the count of cabinet as 5, the count of stool is less than the count of cabinet. Therefore, the answer is yes.
Ori-Q: Is backpack and toaster further than coffee table and mirror? Please provide a "yes" or "no" as the answer.
Ori-A: no
CoT-Q: Is backpack and toaster further than coffee table and mirror? Please solve the problem step by step. Show each intermediate thought process clearly and provide the final answer after completing the reasoning process.
CoT-A: The distance between backpack and toaster is approximately 5.05 meters. The distance between coffee table and mirror is approximately 5.27 meters. Since the distance between backpack and toaster is greater than the distance between coffee table and mirror, the correct answer is no.
Ori-Q: Is the size of the bounding box of curtain approximately equal to the one of cabinet? Kindly provide a "yes" or "no" as the answer.
Ori-A: no
CoT-Q: Is the size of the bounding box of curtain approximately equal to the one of cabinet? Please solve the problem step by step. Show each intermediate thought process clearly and provide the final answer after completing the reasoning process.
CoT-A: Given the volume of the bounding box of curtain as 10.41 cubic meters and the volume of the bounding box of cabinet as 4.88 cubic meters, the volume of the bounding box of curtain is approximately equal to the volume of the bounding box of cabinet. Therefore, the answer is no.

Table 11: Templates for Chain-of-Thoughts QA-Pairs in NUMINA. Each example includes an original question (Ori-Q) with its corresponding direct answer (Ori-A), followed by a CoT-augmented question (CoT-Q) that prompts a step-by-step reasoning process. The CoT-augmented answer (CoT-A) then provides a detailed breakdown of the logical steps leading to the final response.