SAT: Spatial Aptitude Training for Multimodal Language Models

Arijit Ray[1] Jiafei Duan[2] Reuben Tan[1,4] Dina Bashkirova[1] Rose Hendrix[3]
Kiana Ehsani[3] Aniruddha Kembhavi[2,3] Bryan A. Plummer[1] Ranjay Krishna[2,3*]
Kuo-Hao Zeng[3*] Kate Saenko[1*]
[1]Boston University, [2]University of Washington, [3]Allen AI, [4]Microsoft Research

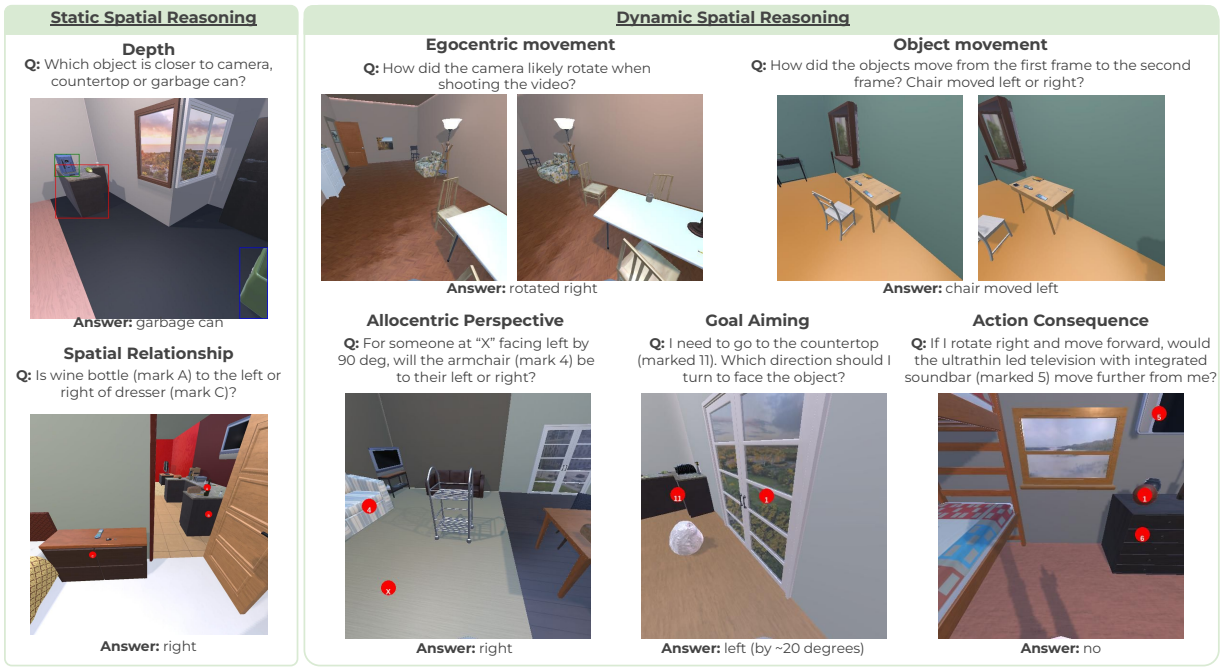

Figure 1. We propose Spatial Aptitude Training (SAT), an approach to improving spatial reasoning capabilities in Multimodal Language Models (MLMs). SAT generates both *static* spatial questions covered by existing benchmarks, and, inspired by cognitive science, the more challenging *dynamic* spatial questions involving egocentric actions, object movement, and perspective-taking. Training with SAT data improves both MLMs' static spatial reasoning on existing benchmarks and dynamic reasoning on our new benchmark.

Abstract

Spatial perception is a fundamental component of intelligence. While many studies highlight that large multimodal language models (MLMs) struggle to reason about space, they only test for static spatial reasoning, such as categorizing the relative positions of objects. Meanwhile, real-world deployment requires dynamic capabilities like perspective-taking and egocentric action recognition. As a roadmap to improving spatial intelligence, we introduce SAT, Spatial Aptitude Training, which goes beyond static relative object position questions to the more dynamic tasks. SAT contains 218K question-answer pairs for 22K synthetic scenes across a training and testing set. Generated using a photo-realistic physics engine, our dataset can be arbitrarily scaled and easily extended to new actions, scenes, and 3D assets. We find that even MLMs that perform relatively well on static questions struggle to accurately answer dynamic spatial questions. Further, we show that SAT instruction-tuning data improves not only dynamic spatial reasoning on SAT, but also zero-shot performance on existing real-image spatial benchmarks: 23% on CVBench, 9% on the harder BLINK benchmark, and 18% on VSR. When instruction-tuned on SAT, LLaVA-13B matches larger proprietary MLMs like GPT4-V and Gemini-3-1.0 in spatial reasoning.

*equal advising

1. Introduction

Cognitive scientists posit that spatial reasoning is not merely a subset of human cognitive abilities but rather the fundamental underpinnings of most intellectual processes [82]; spatial reasoning in school children improves their aptitude in geometry, physics, and even linguistic reasoning [61, 78]. Hence, it is perhaps not surprising that many idioms utilize space to explain concepts: "hitting a wall," "a step in the right direction," and "thinking outside the box."

Despite their widespread adoption and promise as human-level intelligent agents, multimodal language models (MLMs) [2, 21, 58, 100] still struggle to reason spatially [14, 19, 37, 80]. Most of these studies focus on simple spatial questions in *static* scenarios, such as the relative positions of static objects in a fixed scene. Meanwhile, many downstream applications, such as smart glasses and embodied AI, could benefit from dynamic capabilities involving movement, like perspective-taking and egocentric action recognition [31, 56, 94]. Such dynamic capabilities are well-studied in human developmental cognition [9, 83] and argued to be fundamental to intelligence; children understand the consequences of their movement (the moving room test [6]), understand how entities in the scene move [6], and take the perspectives of other humans when planning [12].

To promote progress towards dynamic spatial understanding for MLMs, we propose Spatial Aptitude Training (SAT), an approach for generating spatial question-answer (QA) data without any human supervision to train and evaluate MLMs. Annotating scenes with 3D information is expensive; instead, SAT leverages 22K ProcTHOR [24] scenes composed of 1K assets to generate 218K QA pairs. With perfect 3D information and control of the assets, SAT goes beyond static object relationships to questions that require reasoning based on objects moving and egocentric actions in the scene. Since our data is generated procedurally by composing assets, it can be scaled up without human annotation. Hence, SAT is more flexible than 3D datasets [7, 10], which are not composable and have fewer object classes (~98) [10].

With SAT, we analyze what kinds of training data improve spatial reasoning in MLMs. We focus on two kinds of spatial reasoning data. First, spatial-QA about simpler object relations in static scenes, shown in Fig. 1-left, to impart reasoning about the relative locations of objects in the scene (*e.g.* where is object X with respect to object Y? behind, above, left, or right?) as well as counting. Next, we evaluate the effect of dynamic spatial tasks, Fig. 1-right. This includes egocentric movement, object movement, allocentric perspective, goal aiming, and action consequences. These complex spatial tasks go beyond static object relationships, assessing the MLM's capability to reason about spatial movements, perspective changes, and certain degrees of spatial causality.

We use LLaVA-1.5-13B [58], a widely adopted open-source MLM, as our base model for evaluations. To test static spatial reasoning, we use three contemporary real image benchmarks: CV-Bench [80], BLINK [37], and Visual Spatial Relations (VSR) dataset [57]. Since no real dataset exists for dynamic spatial reasoning, we use SAT test set.

First, our results find that both open and closed MLMs struggle to reason spatially, performing near random chance on our dynamic spatial QAs, including MLMs that perform well on static QAs. We find that instruction tuning data in our SAT dataset improves spatial performance on both our dynamic SAT test set as well as real image spatial benchmarks- notably, by 23% on CVBench [80], 9% on harder relationships on BLINK [37], and 18% on Visual Spatial Relations (VSR) [57], without using any training data from these sources. Interestingly, the synergy of dynamic tasks added to static spatial reasoning tasks improves performance over tuning with static QAs alone, especially on test sets requiring 3D estimation. Notably, our instruction tuning makes a 13B parameter LLaVA model match or outperform some large closed-source models [2, 79] on zero-shot performance on CVBench [80] and BLINK [37].

Adjacent to concurrent work that finds spatial cognition to not emerge in frontier models [69] trained on disembodied web data, our work shows promise that training with data generated using embodied movements and interactions in photo-realistic simulators can indeed help instill spatial intelligence in MLMs.

2. Related work

Our work draws inspiration from fundamental schools of thought in neuroscience that suggest spatial intelligence is a core foundation for most cognitive abilities [30, 61, 82].

3D Spatial Understanding Benchmarks 3D and spatial understanding in vision have been extensively studied. Holistic 3D scene understanding [22, 40, 62, 71] mostly focuses on estimating the layout of an indoor scene as opposed to grounding the 3D nature of the objects of the scene with spatial language words. This is also true for models for localization with coordinates for both 3D and 2D, such as predicting segmentation maps [41, 46, 50, 52, 65, 85], object localization [48, 66, 72], and tracking [4, 53, 55]. Various works in the joint 3D and language understanding space like fine-grained scene captioning [18, 76], open-vocabulary classification and localization [3, 15, 41, 73], question answering [7, 92], and language models [42] deal with full 3D scans. In contrast, our framework emphasizes understanding the 3D spatial configuration of the objects from 2D images since most open MLMs operate on RGB images and 3D scans are expensive to compute, especially in dynamically changing environments. While there are works that deal with the 3D structure of a 2D scene [8, 75, 99], most of them do not have high-resolution images (since they are rendered from point clouds) and diverse object annotations since they are expensive to annotate and collect. In contrast, SAT does

Table 1. Comparison of existing datasets to ours. Unlike ours, most benchmarks do not have pipelines to generate captions and question-answer pairs on new scenes. Our dataset is synthetic and interactive, allowing us to collect large-scale 3D spatial reasoning data for free. We focus on 3D spatial reasoning on 2D images since most open MLMs are tuned to accept 2D images, as opposed to taking a 3D scan input.

	SAT (Ours)	2D Vision-Language GQA, VG, Obj365	3D Vision-Langauge Omni3D, ScanQA	Spatial Rel VSR, 2.5VRD	Spatial QA CVBench, BLINK, Sp VLM, Sp RGPT
2D Annotations	✓	✓	✓	✓	✓
3D Annotations	✓	✗	✓	✓*	✗
Basic Spatial QA	✓	✓*	✓*	✗	✓
Complex Spatial QA	✓	✗	✗	✗	✗
New Scene/Task Gen	✓	✗	✗	✗	✗
Object Interaction	✓	✗	✗	✗	✗

not require any human annotations and can be arbitrarily scaled up to include more assets and tasks. Further, being built on the AI2Thor engine enables the ability to move objects and alignment with multiple interactive applications [23, 25, 88]. In light of potential applications in embodied AI [30], we align our dataset to be compatible with physics simulators, unlike existing 3D understanding datasets with object-attribute annotations [13]. We differ from recent existing works [14, 29, 37, 38, 49, 57, 76, 89], which only address basic spatial reasoning like the relationships between objects in a static scene since we include more complex tasks that require reasoning over dynamic frames with egocentric movement and allocentric perspectives. We also differ from 3D LLMs [20, 42], which do large-scale pretraining using 3D inputs, by focusing on low-resource adaptations using only 2D images without changing encoders. Adjacent to a concurrent work that finds spatial cognition to not emerge in frontier models [69] by testing on cognitive science-based graphical tests, we formulate similar tests in realistic environments aligned with embodied applications [33]. Tab. 1 overviews the comparison between SAT and existing benchmarks.

Synthetic-to-Real Data Training An age-old question in computer vision has been whether perfect synthetic information can boost reasoning in real environments. Works in this domain have studied the effect of synthetic images in classification [16], semantic understanding [64], correcting biases [67], and recently embodied AI [33, 74]. Closing the syn-to-real gap is a well studied problem with various domain adaptation works [17, 33] proposing techniques to close the gap to generalize to real domains, especially in embodied AI. Inspired by this, our work explores if synthetic data with perfect 2D/3D information can improve spatial reasoning in MLMs.

Vision and Language Models. Our task is heavily influenced by the emergence of multimodal foundation models [36, 47, 68, 84, 86, 91, 93]. To leverage the real world knowledge in LLMs and create unified model for real worlds applications, recent approaches have proposed to adapt pretrained visual encoders with LLMs for a wide range of downstream image [5, 21, 28, 55, 81, 100] and video [54, 59, 60, 77, 87, 97] understanding tasks along with taking actions that require spatial understanding of the scene [11, 98]. While these MLMS have demonstrated impressive zero-shot results, recent work [20, 35, 41] has noted their weakness in 3D pose and location estimation. Adjacent to works like [43, 70] that point to deficiencies in understanding compositions of spatial relationships with objects and attributes, we explore if perfect 3D information in synthetic images can improve spatial understanding.

3. SAT: Spatial Aptitude Training

Our goal is to improve the 3D spatial reasoning capabilities of MLMs in situations involving both static and dynamic scenes. Existing 3D datasets have few object annotations and are not controllable/interactive [7, 10]. Since obtaining varied 3D annotations with interactive movements of the camera and objects on real images is expensive and tedious, we propose to teach the model such spatial reasoning using data from procedurally generated photo-realistic environments. The resulting data generation pipeline, SAT, serves as both instruction-tuning data for MLMs and as a benchmark to test the dynamic spatial reasoning capabilities not present in existing benchmarks. While it may be possible to pseudo-annotate 3D spatial information on real images, we show later that this might require extensive cleaning.

In total, our dataset contains 218K questions across 22K procedurally generated scenes from ProcTHOR-10K dataset of indoor apartment buildings. We generate template-based static as well as dynamic spatial questions-answer pairs (QAs) that go beyond the kinds of questions in existing benchmarks. We first collect attribute descriptions (*e.g.*, brown wooden chair) for assets with multiple variations (*e.g.* chairs) by instructing humans to describe the asset in a few words. We then use GPT4-o [2] refine the descriptions into compact phrases. Only some object/asset descriptions required human annotation as a one-time cost. As scenes can be composed arbitrarily from these assets, we can scale up the scenes and QA generation without any human annotation. Next, we outline the generation process for the two kinds of spatial questions- static and dynamic.

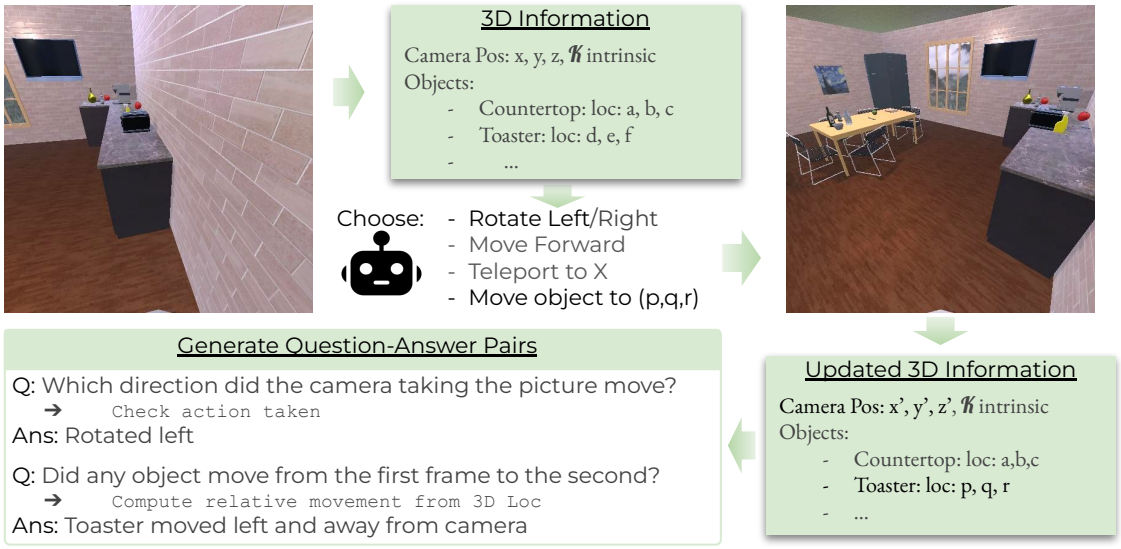

Figure 2. Method of generating our SAT dynamic data: we take actions in a 3D simulator and check the 3D locations of assets. We generate natural language descriptions of the assets and make QA pairs based on how the 3D nature of the scene changes with actions taken.

3.1. Generating Static Spatial QAs.

Aligning with existing contemporary benchmarks for spatial reasoning, we first generate instruction-tuning data for static spatial reasoning that deal with relative relations of objects. Overall, we generate 127K static spatial QA pairs across 8K images across the following types:

Relative spatial relations. We generate questions about the relative location of one object in the scene to other objects. We form two kinds of questions - (1) judging if object X is to the left, right, above, or below object Y. For example, *Is the wine bottle to the left or right of the plate?* (2) judging if object A or B is closer to another object C. For example, *Which object is closer to the wine bottle- the cup, or the plate?* Given the camera parameters, we first project the objects' poses into the camera coordinate system and then generate the corresponding answers. More details about the camera coordinate normalization are in the appendix.

Relative Depth. We generate questions about judging whether object X is closer to the camera than object Y. For example, *Is the wine bottle closer to the camera than the plate?* In our simulator, we calculate the distance between each object and the camera to generate the answer.

Count. Since it is known [39] that many MLMs struggle with counting, we include counting questions. For example, *How many cups are visible in the image?* Since the metadata from our simulator provides the number of object instances with their attributes in a scene, we can automatically obtain answers to this type of question for free.

3.2. Generating Dynamic Spatial QA

Grounded in spatial cognitive tests [6, 12, 83], we outline five different complex tasks that require reasoning about egocentric actions taken, object movements, and allocentric perspectives. We generate 86K QAs on 13K images. The high-level idea behind generating such QAs is illustrated in Figure 2. Given a frame in a simulated environment, we take an action and formulate QAs based on how the 3D orientation of objects changes based on the action taken. Below, we outline the specific approach for each of the question types.

Egocentric Movement. This is based on the "moving room test" [6], a fundamental test designed to assess and improve spatial cognitive development in children. This test aims to measure if an agent can judge how they moved given two frames. This is useful beyond just measuring spatial cognition since high accuracy on this task can help pseudo-annotating navigation data from just egocentric videos. We take a random action from the choices of rotating left or right by an angle sampled randomly. We also randomly choose to move forward by a random distance. We take the first frame and the frame after taking this movement action. Based on the actions taken, we formulate a question of the type: *How did the camera taking the video likely move?* with the answer being the action sequence taken. We only take at most two steps (rotating and moving forward) since we want to ensure only one correct answer of what movement happened from the first frame to the end frame. We have 6.9K training image-QA pairs of this type. The test performance on this task is denoted as **EgoM** in the tables.

Object Movement. Similar to above, we randomly choose

an object and move it by a randomly chosen distance and direction ensuring that the object is still in the frame of view. Next, we compare the updated 3D position of the object with the original position normalized by the camera coordinates to decide if the object got closer, further, more left or more right from the original position. Based on that, we form QA pairs of the type: *Did any of the objects move from the first frame to the second frame?* with the answer being the way the object moved. Note that sometimes objects may move in conjunction with camera movement. To answer the questions accurately, the agent needs to learn to distinguish between egocentric movement and objects moving. We have 6.9K training image-QA pairs of this type. The test performance is denoted as **ObjM** in the tables.

Allocentric Perspective. Inspired by a spatial cognitive test for humans and animals [12], this test checks if the agent is able to take the perspective of another viewer and judge the relative locations of objects according to the other viewer. To make such reasoning QAs, we first choose a 2D point in the scene and mark it as "X". Next, we teleport that agent to the 3D location corresponding to "X" (determined by ray tracing). We check the relative positions of objects according to the camera view from "X" (similarly as described in 3.1- specifically if something is to the left or right of the viewer, and if something got closer or further. We make questions of the type: *For someone at the mark 'X' facing left/right by 90 degrees, would the <object> be to their left or right?* We have 50K training image-QA pairs of this type. The test performance is denoted as **Pers** in the tables.

Goal Aiming. Aiming is a prerequisite for efficient navigation to objects [34], a fundamental spatial cognitive capability. Hence, we design QAs that check how well agents can aim to the desired object. We pick a random object and calculate the angle of the object to the camera using the 3D location of the object and camera assuming looking forward is 0 degrees (exact equations in the supplementary). Based on the angle, we formulate questions of the type: *I need to go to the countertop. Which direction should I turn to face it?*. Since precise angles are hard to judge from a single image, we give the agent choices of rough angles to turn towards the left or right. We have 6.8K training image-QA pairs of this type. The test performance is denoted as **Aim** in the tables.

Action Consequence. This can be thought of as a corollary of the egocentric movement test. Here, the agent needs to reason about how the scene changes when it takes a certain action, inspired by how humans can reason about the consequence of the actions in an environment [34]. Here, we show the first frame and ask the agent to judge if we would move closer/further, or look towards or away from an object if it took that action. *e.g. If I rotate left and move forward, would I be move further from the sofa?*. Note that in most cases moving forward would get us closer to an object. To make the distribution of answers even, we rephrase the question sometimes as to whether we would be facing the object or not. We have 15K training image-QA pairs of this type. The test performance is denoted as **EgoAct** in the tables.

Precise 3D QAs. Finally, using the perfect 3D information in the simulator, we generate question-answer pairs about the precise 3D locations and poses of the objects in the 2D scene. The questions simply ask to estimate the 3D location of an object: *Imagine you are at origin looking at positive z-axis. Given the camera parameters, K, what is the estimated 3D location of the bottle?* We have 98K object descriptions and corresponding 3D annotations on 5K images. While not the main focus of the paper, this can also help drive further research in semantic-based 3D grounding.

4. Experiments

Using our SAT data generation pipeline, we investigate the effects of kinds of instruction-tuning data on the spatial performance of MLMs. Specifically, we wish to compare our synthetic SAT over pseudo-annotated questions on real images using off-the-shelf depth models and ablate the effect of static and dynamic spatial QAs.

4.1. Experimental setup

Evaluation benchmarks. We use 4 spatial benchmarks for our evaluation. CVBench [80], BLINK [37], and Visual Spatial Relations (VSR) [57] measure static spatial understanding on real images. These constitute contain 7K image-QA pairs (around 2.6K for CV Bench, 400 for BLINK, and a 4K set for VSR). We use three spatial splits of BLINK - Multiview reasoning (MV), Relative Depth (RelDep), and Spatial Relations (SpRel). Finally, our SAT test set measures dynamic spatial understanding on 4K QAs on 805 images. Since perspective is overrepresented in our data, we subsample to keep all tasks roughly balanced. Hence, we have 647 object movement, 647 egocentric movement, 592 goal aim, 1336 action consequence, and 778 perspective questions on 805 images.

Instruction-tuning training datasets. Next, we outline the various instruction-tuning sources we use for imparting spatial reasoning:

SAT Static: Using SAT, we first analyze the effect of tuning with only static spatial questions. These tasks are most aligned with the types of questions in existing benchmarks. This set contains 8K images and 127K QAs. This is denoted in row (d) in Table 5.

SAT Dynamic: This set contains dynamic spatial questions. We note that we have proportionally more perspective questions than other tasks and we empirically find perspective easier to overfit to. Hence, we subsample only 2.6K perspective image-QAs and keep other splits the same as described

Table 2. Both closed and open-source models struggle on our dynamic SAT spatial tasks. Table showing the effect of various kinds of spatial data on our SAT Complex Spatial QA test set.

	SAT **Dynamic Test**					
	EgoM	ObjM	EgoAct	Aim	Pers	Avg
Random	47.9	50.6	49.1	50.6	49.8	50
Closed source						
GPT4-o	**61.5**	**33.2**	53.3	**67.5**	34.0	49.7
Gemini-1.5-pro	57.6	29.8	**55.5**	56.9	**49.4**	**51.4**
Open source, zero-shot on SAT						
RoboPoint-13B	50.2	69.4	48.8	**72.6**	25.5	**51.4**
LLaVA-1.5-13B	46.6	**73.8**	**49.7**	45.6	**39.9**	50.6
+ GQA/VG	**51.3**	50.1	46.6	57.2	8.21	42.0
+ VSR/2.5V	46.1	64.6	48.9	38.5	26.2	45.0
Open source, fine-tuned on SAT						
+ SAT-13B Static	45.7	71.5	47.2	72.1	35.2	52.2
+Dynamic	**61.7**	**90.2**	**91.4**	**96.8**	**98.5**	**88.6**

Table 3. Performance with other state-of-the-art closed source or spatially tuned models on CVBench [80].

	2D Avg	**3D Avg**
GPT4-V	64.3	73.8
Cambrian-13B	72.5	71.8
RoboPoint-13B	65.2	73.0
SAT-13B Static+Dynamic	**73.2**	**74.1**

Table 4. Performance with other state-of-the-art closed-source or spatially tuned models on BLINK [37].

	MV	**RelDep**	**SpRel**
GPT4-V	55.6	59.7	72.7
GPT4-o	**59.4**	**74.2**	69.2
Gemini-3-1.0	44.4	40.3	74.8
Claude3 Opus	56.4	47.6	58.0
RoboPoint-13B	48.1	51.6	**75.5**
SAT-13B Static	55.6	66.9	66.4
+Dynamic	55.6	**74.2**	65.7

in Section 3.2 during training. Hence, we use 7.5K images with 40K QAs. We mix both SAT Static and Dynamic when tuning in this case. This is denoted in row (e) in Table 5.

Real GQA/VG + Depth: In contrast to our synthetic scenes, this set is a strong baseline containing real images. Ideally, we would use the instruction tuning data used in recent work on spatial understanding; however, their datasets have not been made public [14, 19]. Hence, we reproduce similar instruction-tuning sets. Specifically, we first infer the depth on real images from GQA [45] and VisualGenome [51] using the DepthAnything [90] model. We also use the 2D bounding box annotations in these datasets to create 2D spatial relationships. Since the annotations tend to be noisy, we filter out potential incorrect relationships using simple heuristics (more details in the appendix). Using the 2D relationships and the 3D depth estimates, we formulate questions similar to SAT static QAs. Generating dynamic QAs with real images is difficult since we cannot take actions on real images. We create 225K static spatial image-QA tuples. This is denoted in row (b) in Table 5.

Real VSR/VRD: We also use available data in spatial relationship datasets like VSR [57] and 2.5VRD [76] to produce more real spatial data. These datasets contain relationships such as "touching", or "behind", allowing us to formulate QAs such as "Is the cat touching the sofa?" Due to human annotation, the diversity of relationship words used here are higher than SAT. Since each of the datasets here are small, we combine VSR and 2.5VRD to create 107K image-QA tuples. This is denoted in row (c) in Table 5.

Spatially-tuned models: Robopoint [94] is an existing work that contains instruction tuning data for finetuning MLMs for robotics applications. We use their fine-tuned model as a strong baseline for a spatially-tuned model. For CVBench [80], we use their spatially-tuned model, Cambrian-13B, as another strong baseline.

Metrics. We report the standard accuracy metric used for question-answering evaluations, by checking if the predicted answer matches the GT answer. We find that off-the-shelf MLMs are sensitive to prompt formats. For instance, we notice better performance when we provide the options in text (*e.g.* Choose between *right* or *left*) as opposed to option numbers (*e.g.* Choose between option *A* or *B*). Hence, we sometimes report higher performances for the baselines than those reported by the original papers.

Real-data mixed tuning details. We base our experiments on a widely used open-source MLM, LLaVA-1.5-13B [58] for fine-tuning experiments. We LoRA-tune [44] with rank 128 and alpha 256. We also find adding precise 3D QAs to not help performance, and hence, we exclude it during tuning. To prevent catastrophic forgetting, we randomly sample examples from the LLaVA Instruct Tuning dataset [58] with 40% probability while tuning with our synthetic QA pairs. We train until the model has converged based on training loss (full details in supplementary). Due to memory constraints, we use a batch size of 8 (using gradient accumulation). We set a small learning rate of $5e^{-6}$ (due to our small batch size) with cosine annealing with 1K warm-up steps and a weight decay of 0. Training requires two 48GB NVIDIA GPUs, while inference is possible with one GPU. We provide a prompt (with the same format as LLaVA [58] instruction tuning) that includes a question, the 2D image, and the possible answer choices following the standard convention of existing MLM benchmarks. Please find our exact prompts in the supplementary.

Table 5. Table showing the effect of instruction-tuning with various kinds of spatial data on the zero-shot accuracy on existing benchmarks. We see that synthetic dynamic spatial reasoning data improves over static spatial data in improving the spatial reasoning of LLaVA.

	CV-Bench [80]					BLINK [37]				VSR [57]
	Count	**2DRel**	**3DDep**	**3D Dist**	**Avg**	**MV**	**RelDep**	**SpRel**	**Avg**	**Avg**
a. LLaVA-1.5-13B	58.2	46.6	53.0	47.8	51.4	45.1	56.4	**69.9**	57.0	48.6
b. + GQA/VG + Depth	61.9	69.4	35.5	51.8	54.6	54.9	48.4	55.9	53.0	56.3
c. + VSR / 2.5V	47.1	67.2	31.5	31.2	44.2	33.1	52.4	44.7	43.4	65.8
d. + SAT Static	59.5	81.7	72.5	54.2	66.9	**55.6**	66.9	66.4	63.0	61.1
e. +Dynamic	**62.9**	**85.8**	**76.6**	**71.6**	**74.3**	**55.6**	**74.2**	65.7	**66.7**	**66.5**

Table 6. Performance on other VQA benchmarks. Our spatial tuning performs at par with the baseline LLaVA, suggesting that it remembers pre-training commonsense after our instruction tuning.

	GQA [45]	OK-VQA [63]	VQA-v2 [39]
LLaVA-13B	78.6	30.7	60.5
+ SAT Dynamic	**79.8**	**36.6**	**63.0**

5. Results

Stronger closed-source and spatially-tuned models struggle on SAT dynamic QAs despite performing well on static. We note that closed-source models (GPT4-o [2], Gemini-1.5-pro [79]) and spatially-tuned Robopoint [95] struggle on complex QAs (in Table 2. For GPT4-o [2] and Robopoint [94], they perform well on static QAs (in Tables 3, 4), but not on SAT. We see that perspective is challenging when tested zero-shot since whether something is to the left or right often flips when taking the perspective of another viewer. Aiming to goal is easier for spatially stronger models like RoboPoint and GPT4-o.

Tuning on SAT improves performance across both dynamic and static spatial QAs. Tuning on SAT QAs improves performance on the test set for dynamic spatial reasoning on SAT as shown in Table 2. While this is not surprising, this also improves performance on static spatial questions on BLINK [37] (by 9%), CV-Bench [80] (by 23%) and VSR [57] (by 18%) benchmarks as shown in Table 5 compared to off-the-shelf LLaVA (rows a vs d, e). BLINK is harder - for instance, spatial relations that have abstract relationships (especially with people) not present in our synthetic data (*e.g.* "looking away"). An example is shown in Figure 3 (bottom row, 3rd image). Multiview Reasoning gains are also modest, and the same issue is observed on the related Egocentric Movement split on SAT data.

Our 13B tuned model matches/outperforms some larger proprietary or spatially-tuned models We compare our performance with that of closed-source models reported by the official benchmark papers. In Table 3 and Table 4, we observe that instruction tuning with synthetic data on SAT can make a LLaVA1.5-13B model match closed-source models on zero-shot performance for real images on CVBench and BLINK. Compared to spatially-tuned baselines, we outperform RoboPoint [95] on multiview reasoning and relative depth on BLINK in Table 4, and overall for CVBench in Table 3. We also outperform Cambrian-1 [80], another strong model on CVBench. This shows promise that SAT instruction-tuning may push performance further for the some for these stronger models.

Our SAT-tuned model remembers pre-training commonsense- we slightly improve on other VQA benchmarks. We run an evaluation on some standard VQA benchmarks- namely, GQA [45], VQAv2 [39] and OK-VQA [63] (9K image-QA pairs, 3K from each). We slightly improve performance on them compared to the off-the-shelf LLaVA [58] as shown in Table 6. This suggests that we remember pre-trained vision-language commonsense while adding stronger spatial capabilities.

Adding dynamic QAs further improves static QA performance over just static QAs. We notice improvements on all splits of CVBench when we add complex spatial data in the training as noted in Table 5 (row d vs e). For BLINK [37], we see an improvement in relative depth and a minor improvement in multiview reasoning (MV). MV seems especially challenging since we observe our model heavily biased to "rotated right." The accuracy on the related EgoM split on SAT also remains low despite fine-tuning (Table 2)

SAT synthetic data is competitive over using pseudo-annotated spatial QAs on real images In Table 5, we can also improve over using spatial questions generated using GQA and VSR/2.5VRD annotations as described in Section 4.1. We see that while real data performs well in the 2D splits like Count and 2DRel on CVBench, it lacks 3D reasoning. We observe that grounding annotations may be noisy in these real datasets, with bounding boxes often not accounting for the entire object, leading to errors in judging relations. Depth estimation further adds noise, resulting in errors in 3D perception. However, following more careful curation similar to [19] may improve this performance further, which we leave to future work since their data is not

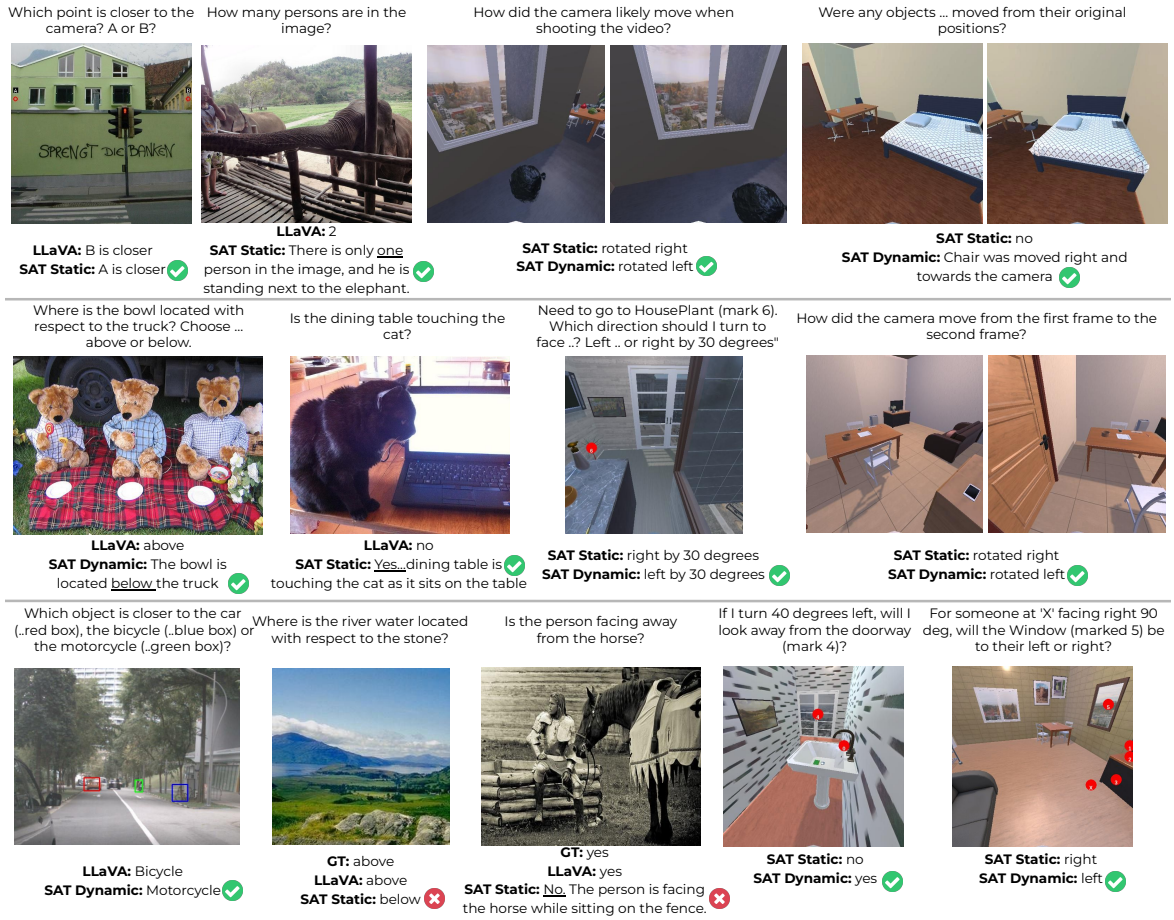

Figure 3. Some qualitative results of spatial question answering comparing baseline LLaVA on real benchmarks and difference between static and dynamic tuning on SAT. More examples in supplementary.

yet easily available.

Human annotated static spatial relations tend to perform well only in-domain Human-annotated spatial relations (using datasets like VSR/2.5VRD as described in 4.1) on real images only tend to perform well in-domain in Table 5 - VSR/2.5V train performs well on VSR test. This is because such datasets are static with a finite amount of relations annotated. We cannot easily generate varied instruction data with other kinds of 3D spatial reasoning using such annotations.

Despite perfect counting data in synthetic images, models continue to struggle with counting. We see minor gains in counting accuracy from the zero-shot models for LLaVA in Table 5. We observe real data (GQA/VG) tends to perform better on counting along with our SAT Dynamic model. We observe that failure cases are due to focusing only on salient, visible objects while often neglecting background objects. We include examples in the supplementary.

6. Conclusion

Limitations. We instruction-tune an MLM for spatial reasoning, LLaVA [58]. While we do remember pretraining commonsense as noted in Table 6, we haven't explored improving other capabilities like math and science reasoning [96]. The scope of the paper, however, is to analyze what kinds of data improve spatial performance, and not a large-scale training of a new MLM. Additionally, further analysis is necessary on more recent MLMs [26, 32].

Future work. Although our study focuses on evaluating the spatial reasoning capabilities of MLMs, it can be extended in various avenues. For instance, to determine the kinds of embodied applications that benefit from improved complex spatial reasoning. As a preliminary study, we checked the action prediction accuracy (from a choice of going left/right/forward) for a given frame on the SPOC EasyObjectNav benchmark [33]. Our model (SAT Dynamic) scores an accuracy of 51% compared to 40% for a model trained only on basic spatial questions. This suggests that

embodied navigation might benefit from improved dynamic spatial reasoning. We leave a more thorough evaluation in this direction as a future work. Further, leveraging the interactive nature of our scenes could facilitate explorations in dynamic and causal reasoning. Another exciting avenue would be to explore how to make a more realistic dynamic spatial set instead of synthetic images.

Conclusion. We propose a dataset of dynamic spatial tasks that go beyond simpler static reasoning on existing datasets. This improves spatial reasoning of MLMs on various benchmarks, while maintaining pre-trained commonsense. We hope that SAT paves the way for developing strategies to improve the spatial reasoning of MLMs, making them more suitable for deployment in real-life applications.

References

[1] AI2-THOR: An Interactive 3D Environment for Visual AI. *ArXiv*, abs/1712.05474, 2017. 1, 2

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3, 7

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

[4] Christopher Agia, Krishna Murthy Jatavallabhula, Mohamed Khodeir, Ondrej Miksik, Vibhav Vineet, Mustafa Mukadam, Liam Paull, and Florian Shkurti. Taskography: Evaluating robot task planning over large 3d scene graphs. In *Conference on Robot Learning*, pages 46–58. PMLR, 2022. 2

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[6] David I Anderson, Joseph J Campos, David C Witherington, Audun Dahl, Monica Rivera, Minxuan He, Ichiro Uchiyama, and Marianne Barbu-Roth. The role of locomotion in psychological development. *Frontiers in psychology*, 4:440, 2013. 2, 4

[7] 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*, pages 19129–19139, 2022. 2, 3

[8] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 2

[9] Mark Blades and Christopher Spencer. The development of children's ability to use spatial representations. *Advances in child development and behavior*, 25:157–199, 1994. 2

[10] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13154–13164, 2023. 2, 3

[11] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 3

[12] Maria Brucato, Andrea Frick, Stefan Pichelmann, Alina Nazareth, and Nora S Newcombe. Measuring spatial perspective taking: Analysis of four measures using item response theory. *Topics in Cognitive Science*, 15(1):46–74, 2023. 2, 4, 5

[13] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving, 2020. 3

[14] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024. 2, 3, 6

[15] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020. 2

[16] Wuyang Chen, Zhiding Yu, Shalini De Mello, Sifei Liu, Jose M. Alvarez, Zhangyang Wang, and Anima Anandkumar. Contrastive syn-to-real generalization, 2021. 3

[17] Xiaoyu Chen, Jiachen Hu, Chi Jin, Lihong Li, and Liwei Wang. Understanding domain randomization for sim-to-real transfer, 2022. 3

[18] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203, 2021. 2

[19] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model, 2024. 2, 6, 7

[20] Jang Hyun Cho, Boris Ivanovic, Yulong Cao, Edward Schmerling, Yue Wang, Xinshuo Weng, Boyi Li, Yurong You, Philipp Krähenbühl, Yan Wang, et al. Language-image models with 3d understanding. *arXiv preprint arXiv:2405.03685*, 2024. 3

[21] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[22] Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese. Delay: Robust spatial layout estimation for cluttered indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 616–624, 2016. 2

[23] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3164–3174, 2020. 3

[24] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022. 2, 1

[25] Matt Deitke, Rose Hendrix, Ali Farhadi, Kiana Ehsani, and Aniruddha Kembhavi. Phone2proc: Bringing robust robots into our chaotic world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9665–9675, 2023. 3

[26] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, Yen-Sung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models, 2024. 8

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 3

[28] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 3

[29] Jiafei Duan, Samson Yu, Soujanya Poria, Bihan Wen, and Cheston Tan. Pip: Physical interaction prediction via mental simulation with span selection. In *European Conference on Computer Vision*, pages 405–421. Springer, 2022. 3

[30] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022. 2, 3

[31] Jiafei Duan, Wentao Yuan, Wilbert Pumacay, Yi Ru Wang, Kiana Ehsani, Dieter Fox, and Ranjay Krishna. Manipulate-anything: Automating real-world robots using vision-language models. In *Conference on Robot Learning*, 2024. 2

[32] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 8

[33] Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, et al. Spoc: Imitating shortest paths in simulation enables effective navigation and manipulation in the real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16238–16250, 2024. 3, 8

[34] Matthias O Franz and Hanspeter A Mallot. Biomimetic robot navigation. *Robotics and autonomous Systems*, 30 (1-2):133–153, 2000. 5

[35] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 3

[36] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 3

[37] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 2, 3, 5, 6, 7, 4

[38] Ankit Goyal, Kaiyu Yang, Dawei Yang, and Jia Deng. Rel3d: A minimally contrastive benchmark for grounding spatial relations in 3d. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[39] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. 4, 7

[40] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *2009 IEEE 12th international conference on computer vision*, pages 1849–1856. IEEE, 2009. 2

[41] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023. 2, 3

[42] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models, 2023. 2, 3

[43] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality, 2023. 3

[44] 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, 2021. 6, 3

[45] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. 6, 7

[46] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping, 2023. 2

[47] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3

[48] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr – modulated detection for end-to-end multi-modal understanding, 2021. 2

[49] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning, 2023. 3

[50] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 2

[51] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 6

[52] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022. 2

[53] Andrey Kurenkov, Michael Lingelbach, Tanmay Agarwal, Emily Jin, Chengshu Li, Ruohan Zhang, Li Fei-Fei, Jiajun Wu, Silvio Savarese, and Roberto Martın-Martın. Modeling dynamic environments with scene graph memory. In *International Conference on Machine Learning*, pages 17976–17993. PMLR, 2023. 2

[54] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3

[55] Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, et al. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35:31199–31212, 2022. 2, 3

[56] Benlin Liu, Yuhao Dong, Yiqin Wang, Yongming Rao, Yansong Tang, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondence elicit 3d spacetime understanding in multimodal language model. *arXiv preprint arXiv:2408.00754*, 2024. 2

[57] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning, 2023. 2, 3, 5, 6, 7

[58] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 6, 7, 8, 3

[59] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 3

[60] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 3

[61] Hanspeter A Mallot. *From geometry to behavior: An introduction to spatial cognition*. MIT Press, 2024. 2

[62] Arun Mallya and Svetlana Lazebnik. Learning informative edge maps for indoor scene layout prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 936–944, 2015. 2

[63] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019. 7

[64] Samarth Mishra, Rameswar Panda, Cheng Perng Phoo, Chun-Fu Richard Chen, Leonid Karlinsky, Kate Saenko, Venkatesh Saligrama, and Rogerio S Feris. Task2sim: Towards effective pre-training and transfer from synthetic data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9194–9204, 2022. 3

[65] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 2

[66] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2

[67] Maan Qraitem, Kate Saenko, and Bryan A Plummer. From fake to real (ffr): A two-stage training pipeline for mitigating spurious correlations with synthetic data. *arXiv preprint arXiv:2308.04553*, 2023. 3

[68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[69] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models?, 2024. 2, 3

[70] Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A. Plummer, Ranjay Krishna, and Kate Saenko. Cola: A benchmark for compositional text-to-image retrieval, 2023. 3

[71] Yuzhuo Ren, Shangwen Li, Chen Chen, and C-C Jay Kuo. A coarse-to-fine indoor layout estimation (cfile) method. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V 13*, pages 36–51. Springer, 2017. 2

[72] Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd Van Steenkiste, Filip Pavetic, Mario Lucic, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. *Advances in Neural Information Processing Systems*, 35:9512–9524, 2022. 2

[73] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 2

[74] Sneha Silwal, Karmesh Yadav, Tingfan Wu, Jay Vakil, Arjun Majumdar, Sergio Arnaud, Claire Chen, Vincent-Pierre Berges, Dhruv Batra, Aravind Rajeswaran, et al. What do we learn from a large-scale study of pre-trained visual representations in sim and real environments? In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 17515–17521. IEEE, 2024. 3

[75] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 2

[76] Yu-Chuan Su, Soravit Changpinyo, Xiangning Chen, Sathish Thoppay, Cho-Jui Hsieh, Lior Shapira, Radu Soricut, Hartwig Adam, Matthew Brown, Ming-Hsuan Yang, and Boqing Gong. 2.5d visual relationship detection, 2021. 2, 3, 6

[77] Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A Plummer, Bryan Russell, and Kate Saenko. Koala: Key frame-conditioned long video-llm. *arXiv preprint arXiv:2404.04346*, 2024. 3

[78] Holly A Taylor and Barbara Tversky. Spatial mental models derived from survey and route descriptions. *Journal of Memory and language*, 31(2):261–292, 1992. 2

[79] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 7

[80] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024. 2, 5, 6, 7, 4

[81] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 3

[82] Barbara Tversky and Masaki Suwa. Thinking with sketches. 2009. 2

[83] Marina Vasilyeva and Stella F Lourenco. Development of spatial cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(3):349–362, 2012. 2, 4

[84] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35:5696–5710, 2022. 3

[85] Peng-Shuai Wang. Octformer: Octree-based transformers for 3d point clouds. *ACM Transactions on Graphics (TOG)*, 42(4):1–11, 2023. 2

[86] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 3

[87] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 3

[88] Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. Visual room rearrangement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5922–5931, 2021. 3

[89] Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition, 2019. 3

[90] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 6

[91] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 3

[92] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. *arXiv preprint arXiv:2112.08359*, 2021. 2

[93] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3

[94] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. In *Conference on Robot Learning*, 2024. 2, 6, 7

[95] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics, 2024. 7

[96] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024. 8

[97] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3

[98] Jimuyang Zhang, Zanming Huang, Arijit Ray, and Eshed Ohn-Bar. Feedback-guided autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15000–15011, 2024. 3

[99] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 519–535. Springer, 2020. 2

[100] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2, 3

SAT: Spatial Aptitude Training for Multimodal Language Models

Supplementary Material

In this supplementary, we provide more details on the dataset, training, experiments, and some more qualitative examples highlighting failure cases. We also include an HTML data viewer in the zip file. We will be releasing all code and data.

7. Dataset Details

Human Performance on our SAT We conduct a human study with experts to measure the quality of our dataset. We observe that spatial awareness demands more mental power since one has to pay more attention and reason about how the orientation of the scene changed or would change based on an action. We conduct an expert human study, where we ask anonymized graduate students to answer 200 randomly sampled questions from our test set using the interface showed in Figure 4 We see that humans are 92.8% accurate on our SAT dataset. This is still a significant gap compared to the performance of best existing MLM (around 51%). We will release the dataset on Huggingface.

7.1. More details on dataset creation

We first take an apartment from ProcTHOR-10K and place the camera at a position where many objects are visible. We do this by randomly choosing 20 points to place the camera and then choosing the point with max objects visible.

Normalizing the camera coordinates In ProcTHOR [24], in the camera view, the y coordinate is the height coordinate, which means the y increases pointing upwards (*e.g.* the ceiling has a greater y than the floor). Hence, from the bird's eye view, the coordinates are x and z. The rotation of the camera is such that it is always parallel to the x-z plane. Hence the rotation is described as angle clockwise around the y-axis with the camera pointing to the positive z-axis as a 0-degree rotation.

Given a camera rotation, we normalize the view by translating to $(0, 0)$ for x and z by subtracting the camera x_0 and y_0. Further, we rotate the x-z plane such that the camera points to the positive z-axis.

For rotation, we use the formula:

$$R = \begin{bmatrix} \cos(a) & -\sin(a) \\ \sin(a) & \cos(a) \end{bmatrix}$$

Hence, the normalized x', z' for any object is computed using:

$$\begin{bmatrix} x' \\ z' \end{bmatrix} = R \cdot \begin{bmatrix} x - x_0 \\ z - y_0 \end{bmatrix}$$

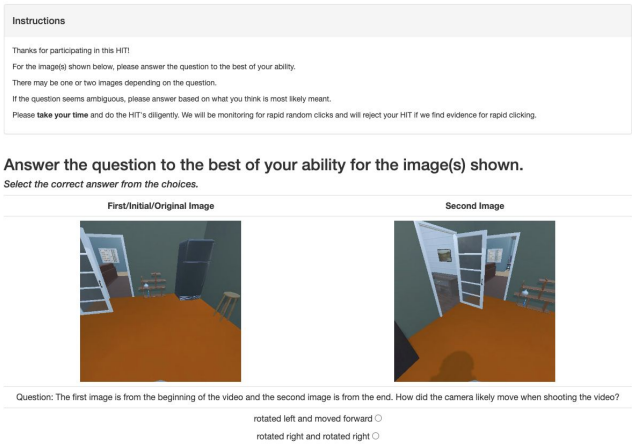

Figure 4. Interface to compute human accuracy for SAT.

The y value remains unchanged since it is the height, which is not affected since we do not change the camera height.

Hence, finally, x' goes negative to the left and positive to the right, z' goes positive towards the depth and y goes positive upwards from the floor level. We use the values of x' and z' to calculate relative relationships (left, right, in front of, and behind) as described below.

7.2. SAT Static Spatial QAs

Relative spatial relations. For instance, if the value of x' for "chair" is lower than that of "table", the chair is to the left of the table. We can also compute the distance between objects. We randomly choose 3 objects. We compute the pairwise distances using their (x, y, z) 3D coordinates. Based on whether object 1 is closer or further to object 2, we make QAs like "Is the couch closer to the lamp or the table?"

Relative Depth. Similarly, if the value of z' for, say, "lamp" is greater than that of "couch", we say the "lamp" is further away from the camera than the couch.

7.3. SAT Dynamic Spatial QAs

Egocentric Movement. We first choose an image frame. Next, we first choose to rotate left or right from angles 20, 30, 40, 50, 60 chosen randomly. We use the `controller.step(action='RotateRight', degrees=angle)` function in the AI2THOR [1] platform. Next, we move forward with probability 0.5 by a random distance from 20 to 40 centimeters (`controller.step(action='MoveAhead',`

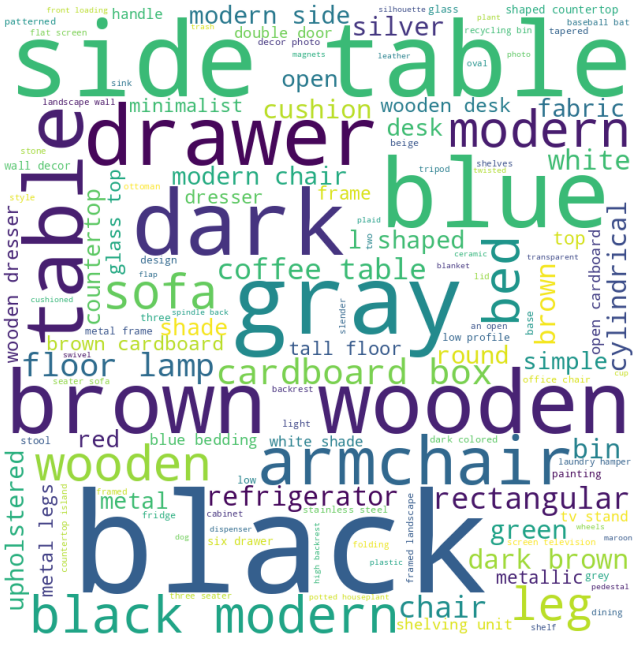

Figure 5. Word cloud of the annotated attributes for the assets.

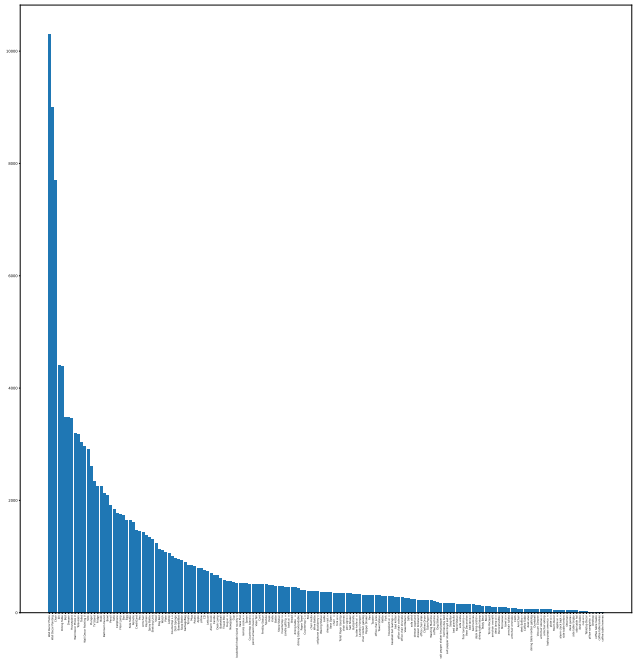

Figure 6. The histogram of object classes in the rooms in our dataset. Best viewed when zoomed.

moveMagnitude=dist). We capture the next frame from this final position.

Object Movement. We first choose an object visible in the scene with at least a certain bounding box area to make sure the object is not too small or non-salient. Next we decide to move the object by a random distance sampled from 0.25 to 0.5 meters in a random direction if possible. We use the PlaceObjectAtPoint function in the AI2THOR [1] platform.

Allocentric Perspective. We choose a point on the 2D image that is not too close to the margsns (in between 0.2-0.8 normalized width and height of the image). We use GetCoordinateFromRaycast action in AI2Thor [1] to get the 3D location of the point. We try a few points since some points may not possible to navigate to and pick one possible point. We randomly turn left or right by 90 degrees. Next, we check whether an object is left/right based on the method described in Section 7.2.

Goal Aiming. We compute the angle between a randomly chosen object with the camera assuming looking straight is 0 degrees. If the angle of an object is $-\alpha$, we say we have to "turn left by α degrees" and "right by α degrees" otherwise. When $|\alpha - 0| \leq \epsilon$, we say we have to look "roughly straight". We define $\epsilon = 10$. Note it is very hard for a human or machine to judge the exact degrees to turn from one single image, hence, we give multiple choices where the difference in choices are between left/right and hence, the machine really only has to decide between them and not the exact angles to turn. We use the following equation to calculate the angle of an object with the camera. First we normalize the object coordinates to the camera coordinates, (x_0, y_0), and then calculate the angle, α.

$$\begin{bmatrix} x' \\ z' \end{bmatrix} = R \cdot \begin{bmatrix} x - x_0 \\ z - y_0 \end{bmatrix}$$

$$\alpha = \arctan\left(\frac{x'}{z'}\right)$$

Action Consequence. Here, we just compute at the objects we got close to or further away while taking the action for the action sequence task or the perspective task. We formulate questions based on that. Note that most of the time we get close to objects in the scene and hence we random subsample such cases.

Precise 3D QAs. Here we just formulate a question where the machine has to estimate the camera-normalized 3D coordinates of an object from the camera parameters and the image. *e.g.*, what is the estimated location of cup if the scene was taken by a camera with focal length f_x, f_y and principle center point c_x, c_y and the image resolution is w, h?

7.4. Dataset statistics

We also compute some salient dataset statistics like the distribution of attributes for the assets annotated in Figure 5, the object classes (in Figure 6), and the distribution of answers for SAT Static (in Figure 7) and for SAT dynamic (in Figure 8).

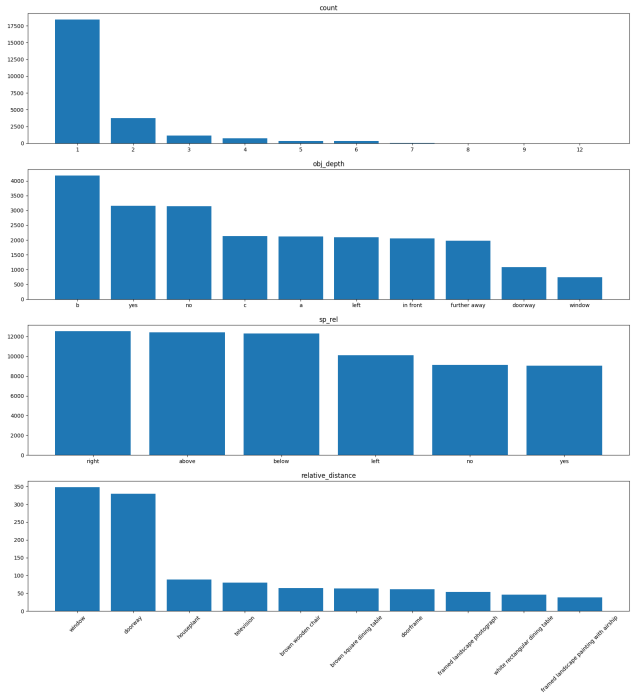

Figure 7. Histogram of the answers for SAT Static.

8. Tuning details

LLaVA All our numbers are reported after training until the learning curve plateaus at around 200K iterations. We use LORA [44] tuning with a rank of 128 and an alpha of 256. We found tuning the image encoder ViT [27] also with LORA to be important for performance. We tune the query and key projection layers for all the transformer layers with LORA. We use learning rate of $5e-6$, a batch size of 1 and gradient accumulation as 8 (effective batch size of 8), weight decay of 0, with a cosine annealing scheduler and a warm up of 1000 steps. We use standard next-token prediction loss from LLaVA official implementation. We train using two 48 GB NVIDIA A6000/RTX6000ada/L40 and we use Huggingface accelerate for the multi-GPU training. Each training takes around 48 hours. For each of the experiments we tune until the training loss plateaus. We see this requires around 200K steps for the static QAs and around 300K steps for the dynamic QAs. We notice no further improvement if we keep training for more iterations with the static QAs. During inference, we use a greedy sampling with temperature 0 following the standard hyperparameters as in the huggingface codebase.

The inference is possible using a single 48GB GPU.

Following LLaVA [58] convention, we use `<image>` tokens to represent images. This is the exact prefix we use.

```
A chat between a curious human
and an artificial intelligence
assistant. The assistant gives
helpful, detailed, and polite
answers to the human's questions.
```

Next, we add the question prompt to the prefix:

```
###Human: <im_start><image><im_end>
Human: Answer in natural language.
Is the person facing the frisbee?
Choose between the following
options: yes, or no.###Assistant:
```

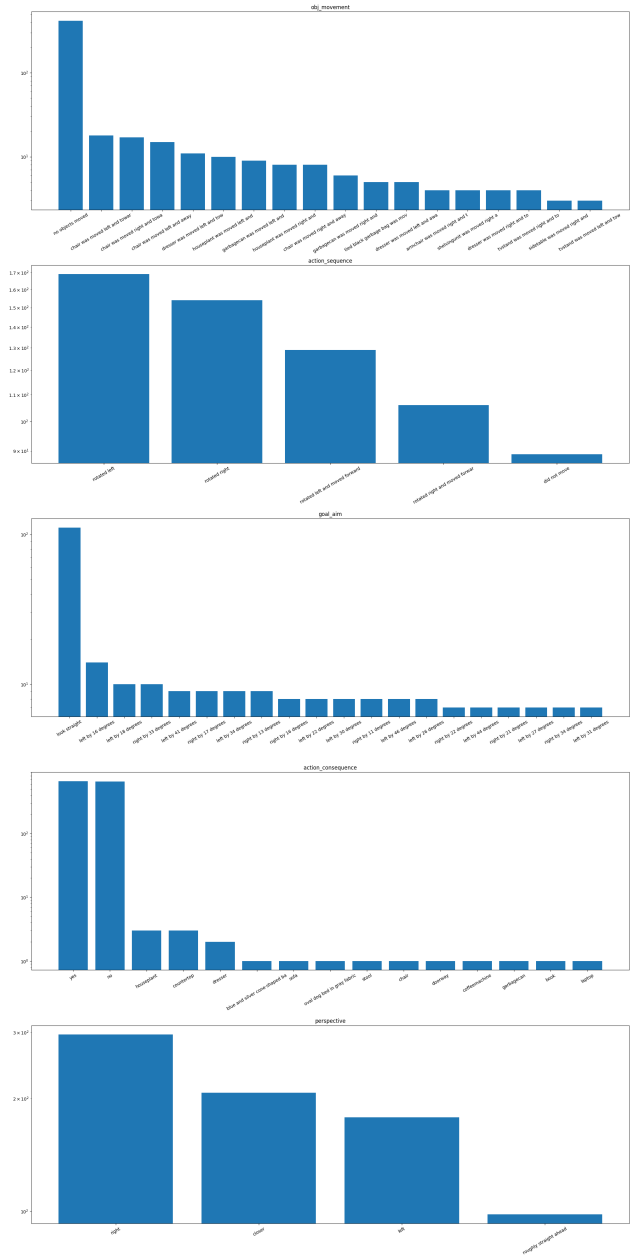

Figure 8. Histogram of the answers for SAT dynamic.

Table 7. Table showing the effect of instruction-tuning with various kinds of data on the zero-shot accuracy on existing benchmarks. Here we make the extra ablation of training with precise 3D QAs mixed in. The * represents the model we presented in the main paper.

	CV-Bench [80]					BLINK [37]			
	Count	**2D Rel**	**3D Depth**	**3D Distance**	**Avg**	**MultiView**	**RelDepth**	**SpatialRel**	**Avg**
a.LLaVA ZS	58.25	46.61	53.00	47.83	51.42	45.11	56.45	69.93	57.16
b. + InstructTune	53.55	50.46	47.33	51.00	50.59	01.50	48.38	48.95	32.94
c. + SAT Static	59.51	81.69	72.50	54.16	66.97	55.64	66.93	66.43	63.00
d. +dynamic*	62.9	85.8	76.6	71.6	74.3	55.64	74.2	65.7	66.7
e. + precise	62.56	80.77	80.33	57.83	70.37	45.11	64.52	69.93	59.85

For questions with two images, we simply have

```
<im_start><image><image><im_end>
```

in the image part of the prompt.

The prefix with "A chat between a ..." is something we found to be very important for LLaVA performance. Hence, we append this prefix to the question both, when tuning and testing. Further, we also found performance improvements when we specify the answer choices in text like "choose between 'left' or 'right"' than asking the model to choose an answer option letter (like A or B) or number (like option 1 or 2). We randomize the answer choice order during evaluation. We also note a higher variance in performance between different training seeds on BLINK due to the small size of the dataset. However, the trends remain the same. We will release all checkpoints, the training script, and the best tips and tricks in the training schedule.

9. Some extra ablations

Simply using more instruct tuning data instead of our data We wish to answer if the gains trivially come from just simply more training on data. Hence, we run a naive baseline of training on more LLaVA instruct tuning data. Unsurprisingly, this does not lead to any gains in spatial performance. Results are shown in row b in Table 7.

Adding precise QAs to the dynamic mix does not help performance Interestingly, we see no improvement when we add precise QA mixed with our full dynamic SAT data (row e). While some tasks improve slightly like depth on CVBench and spatial relations on BLINK, overall performance decreases slightly. This could be due to precise QAs being somewhat ill-defined for a single image. However, this requires further exploration and we leave this to future work.

10. More Qualitative Results

More examples and Failure cases We show more qualitative results in Figure 9. We especially want to investigate some failure cases and hence we display more cases here where our model fails. First, we note that the kinds of camera movements in SAT are sometimes different from those in BLINK (for the multiview reasoning split). Our QAs deal with camera *rotation*, while BLINK [37] QAs deal with camera *movement*. Note that these can be conflicting since the camera can be rotating right while moving left. Hence, we see most failure cases regarding our model answering the direction of camera rotation instead of movement. We also tried adding more camera movement questions to SAT, but however, we do not see any significant performance improvement on that split. Hence, further work is needed to see why egocentric movement is so challenging for MLMs. We also show some counting failures. Understandably, our model often cannot count very non-salient objects (like a tiny bench in the background), and has a heavy bias towards 1 since many objects had only one instance in our randomly generated scenes.

Figure 9. More qualitative results showing some failure cases. Not how some of the camera movement questions focus on camera movement instead of rotation. Our SAT QAs mostly focused on rotations, and hence performance improvements on this split are lower.