BOP-Ask: Object-Interaction Reasoning for Vision-Language Models

Vineet Bhat[1] Sungsu Kim[1] Valts Blukis[2] Greg Heinrich[2] Prashanth Krishnamurthy[1]
Ramesh Karri[1] Stan Birchfield[2] Farshad Khorrami[1] Jonathan Tremblay[2]

[1]New York University [2]NVIDIA

Abstract

Vision–Language Models (VLMs) have achieved impressive performance on spatial reasoning benchmarks, yet these evaluations mask critical weaknesses in understanding object interactions. Current benchmarks test high-level relationships ("left of," "behind", etc.) but ignore fine-grained spatial understanding needed for real-world applications: precise 3D localization, physical compatibility between objects, object affordances and multi-step spatial planning. In this work, we present BOP-Ask, a novel large-scale dataset for object-interaction reasoning for both training and benchmarking. Our data generation pipeline leverages 6D object poses from the Benchmark for Object Pose Estimation (BOP) datasets from which we derive fine-grained annotations such as grasp poses, referred object poses, path planning trajectories, relative spatial and depth relationships, and object-to-object relationships. BOP-Ask comprises over 150k images and 33M question–answer pairs spanning six tasks (four novel), providing a rich resource for training and evaluating VLMs. We evaluate proprietary and open-sourced VLMs, and conduct human evaluations on BOP-ASK-core, a contributed test benchmark. We also release BOP-ASK-lab, an out-of-distribution benchmark with images not sourced from BOP, enabling testing of generalization. Our experiments demonstrate that models trained on BOP-Ask outperform baselines and exhibit emergent capabilities such as precise object and grasp pose estimation, trajectory planning, and fine-grained object-centric spatial reasoning in cluttered environments. Project website: https://bop-ask.github.io/

1. Introduction

Vision-language models (VLMs) have demonstrated impressive performance across robotics, augmented reality, and egocentric vision tasks, supporting capabilities such as scene description [14], or code generation for robot control [30, 48]. Despite this progress, further improvements are needed to enable embodied applications of VLMs. For example, they may correctly identify the coffee jar in a cluttered scene (see Figure 1), but struggle to determine exactly where to grasp it, how to navigate around neighboring objects, or which items must be moved first to access it, which are the very capabilities essential for real-world embodiment. We define this gap as *object-interaction reasoning*: the ability to understand and predict fine-grained physical relationships between objects, including grasp affordances, collision-aware motion paths, and manipulation sequencing in cluttered environments.

Figure 1. The BOP-Ask dataset facilitates object-interaction reasoning for robot manipulation. This illustration demonstrates how a model trained on BOP-Ask enables human and robot-aligned spatial understanding for different actions, supporting physical relationship, locating where to grasp objects, precise pose estimation, and motion planning between objects.

Recent efforts such as SpatialVLM [5] and RoboPoint [62] have taken steps toward training VLMs with explicit spatial objectives. While these approaches improve performance on synthetic or internet-scale benchmarks, they often under-perform with object-to-object relationships, manipulation or long-term reasoning ques-

Table 1. Comparison with spatial reasoning datasets including reference frames and whether they provide motions, poses, and grasping. Only BOP-Ask includes all three.

Dataset	Domain	Ref. Frames	# Images	# Spatial Q&As	Motions	Poses	Grasping
EmbSpatial-Bench [12]	Indoor	✗	2k	4k	✗	✗	✗
Visual Spatial [32]	MSCOCO	✓	10k	10k	✗	✗	✗
SpatialRGPT-Bench [8]	Indoor, AV	✗	1.4k	1.4k	✗	✗	✗
BLINK-Spatial [15]	Generic	✓	286	286	✗	✗	✗
What's up [25]	Generic	✗	5k	10k	✗	✗	✗
Spatial-MM [47]	Generic	✓	2.3k	2.3k	✗	✗	✗
RoboSpatial [50]	Indoor, tabletop	✓	1M	3M	✗	✗	✗
RoboBrain [22]	Tabletop	✗	667k	1M	✓	✗	✗
RoboRefer [65]	Mixed	✗	550k	20M	✗	✗	✗
BOP-Ask	Tabletop	✓	150k	33.8M	✓	✓	✓

tions. More recently, RoboSpatial [50] introduced a spatial reasoning benchmark, establishing an important first step toward bridging perception, spatial reasoning and spatial relationships. In parallel, other approaches such as MolmoAct [29] and RoboBrain [22] train models directly on robot datasets to predict manipulation trajectories and motions. While these directions are promising, they remain limited in scope, especially in object-interaction reasoning: they do not capture the breadth of fine-grained physical interactions needed for general reasoning, such as grasp feasibility, pose estimation, physical relationships, object-based motion planning.

To bridge this gap, we introduce BOP-Ask, a new dataset that extends BOP (Benchmark for Object Pose estimation) [38] into the domain of object-interaction reasoning. While BOP traditionally provides high-quality 3D annotated images with accompanying 3D models for object detection and pose recovery, it does not directly address reasoning about interactions between objects. In contrast to prior spatial reasoning datasets that rely on approximate methods for annotation [5]—such as monocular depth estimation—BOP-Ask inherits precise 3D ground-truth poses from BOP, enabling reasoning at a level of accuracy unseen in the spatial reasoning datasets. BOP-Ask introduces a diverse set of question-answer pairs targeting object-interaction reasoning, such as motion planning, physical relationships, pose estimation, grasping, *etc.* The stated goals for our dataset are: 1) Precision, in contrast to existing datasets that use approximate annotations that cannot support the millimeter-level accuracy required for grasping and motion planning [5, 15]; 2) Interaction completeness to capture the full chain of reasoning from perception to executable manipulation without the need of multiple choice questions, overcoming the limitations of current benchmarks [12, 50, 54] that only evaluate spatial relationships, and, 3) Scale and diversity: spatial reasoning datasets remain small, while large-scale datasets lack tasks related to object interaction [62]. Finally BOP-Ask uses pixel-level question-answers where models are asked to output pixel locations with great precision, this is a departure from previous spatial reasoning datasets that focused on multiple-choice or yes/no questions [5, 15, 50, 53].

BOP-Ask is designed to serve as both a large-scale training corpus and a benchmark for evaluating object-interaction reasoning. The training dataset consists of approximately 150k high-quality images drawn from the BOP benchmark, from which we generate over 33M question–answer pairs spanning object relationships, manipulation affordances, motion feasibility, and scene-level reasoning. This scale already surpasses prior spatial reasoning datasets, while maintaining accurate 6D pose annotations as geometric grounding (see Table 1).

Accompanying the training data, we also introduce two new evaluation benchmarks: BOP-Ask-core is formed from hand-selected held-out BOP scenes, and BOP-Ask-lab is hand-constructed from images taken in different labs, ensuring that it is fully independent of BOP, *e.g.*, camera, object, point of view. Evaluations of both open- and closed-source vision–language models on our hand-curated benchmarks (core and lab) show that existing systems struggle with object-interaction reasoning, underscoring the need for datasets that more directly probe geometric grounding. These results are also striking when we compare them to the results obtained from human evaluation of the same tasks in BOP-Ask-core as these models. We also observe clear gains in real-world robot performance when using the model fine-tuned on BOP-Ask. We highlight our contributions as follows:

- We introduce a novel object-interaction reasoning dataset built on BOP, augmenting perception data with structured queries about object interactions, manipulation affordances, motion feasibility, and scene-level reasoning.
- We show that including BOP-Ask into the training recipe improves VLM performance not only on our test sets but also on out-of-domain spatial reasoning benchmarks such as RoboSpatial-Home [50], Spatial-

Bench [12], and CV-Bench [54].
- We propose BOP-Ask-core and BOP-Ask-lab as benchmarks for embodied spatial and object-interaction reasoning, bridging the gap between pixel-level perception and high-level reasoning.

2. Related Work

Spatial reasoning. Spatial reasoning has long been an implicit and explicit component of numerous vision and question answering tasks [1, 15, 21, 23, 24, 27, 46, 51, 64]. While many benchmarks and methods have been proposed, they show limitations: some are restricted to simulations [52] or generic image datasets [5, 8, 15, 25, 32, 42, 43, 47], others are challenging to evaluate due to reliance on free-form text outputs [12, 31, 52], some require complete 3D scans [31, 35, 36, 63], and many fail to incorporate object-interaction reasoning or robotics [5, 8, 15, 31, 35, 36, 43, 63]. Moreover, several works overlook actionable, robotics-relevant spatial relationships such as spatial compatibility, grasping, motion, and contextual reasoning [12, 25, 31, 43, 44, 47, 57].

VLMs for robotics. VLMs have become pivotal in robotics, enabling systems to interpret and act upon complex visual and textual inputs [6, 48]. By integrating visual perception with natural language understanding, VLMs enable more intuitive human–robot interaction and strengthen autonomous decision-making. Recent progress has demonstrated their potential across diverse robotic applications. For example, vision-language-action models (VLAs) [26, 39, 66] allow robots to parse complex instructions and output executable actions. VLMs like GPT-4v [41] have been leveraged for high-level task planning [56], generating action sequences directly from natural language prompts. Other applications include keypoint/mask prediction [20, 37, 58, 65], error analysis [13, 49], and grasp pose estimation [19]. Recent efforts have explored grasping and motion planning: MolmoAct [29] learns from robot trajectories to generate executable actions, while SpatialPin [34] uses a planner to generate motions, though the model itself cannot output motion sequences directly.

More closely related, prior studies on spatial reasoning [25, 32] and RoboSpatial [50] introduced datasets for reasoning about object reference frames, showing strong potential for robotic applications and spatial reasoning research. Inspired by Robospatial, we further extend the set of tasks that model can now tackle such as grasping, motion planning, object rearrangement, and pose estimation. We note the concurrent work of TIGeR [17], which similarly proposes a dataset for object-interaction reasoning, encompassing tasks like localization, distance estimation, and pose computation. As their dataset and models were not publicly available at the time of writing, a direct comparison on shared tasks was not feasible.

3. Methodology

In this section, we introduce *Object-Interaction Reasoning*—a set of perception and reasoning skills essential for deploying VLMs as embodied agents in real-world robotic environments. We first outline the six core skills that constitute this reasoning framework and discuss their relevance to physical manipulation. We then describe our automated data generation pipeline for constructing BOP-Ask, a large-scale, geometrically grounded dataset for training and evaluating these capabilities.

3.1. Object-Interaction Reasoning

We identify six fundamental skills that a VLM needs to serve as a robust perceptual interface for embodied reasoning and manipulation. *(1) Object pose estimation* involves predicting the 3D bounding box of a referred object, enabling precise visual grounding in a scene. This extends beyond conventional VLMs, which typically mark only points or 2D rectangles. *(2) Object grasp estimation* trains the model to infer stable 3D grasp poses for target objects, providing actionable inputs for downstream robotic execution and manipulation tasks. *(3) Inter-object motion prediction* requires generating collision-free waypoints to move a source object toward a target object. This skill integrates spatial awareness and motion planning, allowing the model to reason about dynamic object-pair interactions in cluttered environments. *(4) Object rearrangement* assesses the ability to understand severely cluttered workspaces and identify which obstructing objects must be moved before grasping a target. This skill allows the model to declutter the workspace before attempting to grasp a target object. Finally, following prior works, we include two binary reasoning tasks: *(5) spatial reasoning* and *(6) relative depth perception*. The former focuses on identifying spatial relationships such as “left of,” “right of,” “above,” and “below,” while the latter determines relative distance attributes (“closer” and “farther”) between object pairs. Together, these six skills provide a unified framework for object-centric perception and reasoning. A VLM fine-tuned under this formulation gains general-purpose scene understanding and manipulation abilities, with its intrinsic pre-training enabling improved zero-shot transfer to unseen objects and environments.

3.2. Dataset Generation

Our data generation framework is designed to produce rich, geometrically consistent interaction data with minimal manual supervision. It grounds the dataset in accurate 3D object poses while automatically generating diverse, object-centric metadata describing interactions and spatial relationships. Formally, the framework takes as input a pose dataset D_p containing RGB-D images, 3D object poses, camera intrinsics, and outputs an object-interaction reasoning dataset where each sample $S = \langle I_k, Q_k, A_k, T_k \rangle$ comprises the input image I_k, a question Q_k, its corresponding answer A_k, and a task label T_k. Figure 2 describes our dataset pipeline.

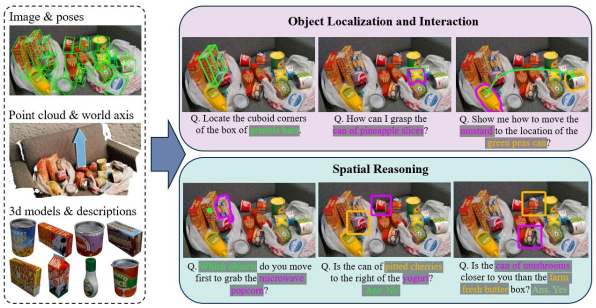

Figure 2. Overview of the BOP-Ask dataset. We automatically generate object-interaction and spatial reasoning annotations from 3D point clouds, images, object poses and 3D models with description. We create question/answer pairs covering 6 types of questions (from left to right, top to bottom), object pose estimation, grasp affordance, motion planning, physical interaction, object relationship, and depth relationship.

3.2.1. World frame construction

A typical scene in a pose estimation dataset consists of RGB-D images, camera intrinsics, and ground-truth object poses. Estimating the world-frame Z direction from object poses alone is unreliable, as the orientation of cluttered objects offers a poor proxy for gravity. To reconstruct a consistent world coordinate system, we estimate the camera-to-world transformation ${}^{cam}T_{world} \in SE(3)$. We first localize the planar support surface (*e.g.*, a table) on which objects rest using a pointing focused VLM [9] and fit a plane to the corresponding 3D points via RANSAC. The plane normal $\mathbf{n}_p$ defines the world up direction. Let $\mathbf{v}_z = [0, 0, 1]^\top$ denote the canonical world up-axis. The rotation aligning $\mathbf{v}_z$ to $\mathbf{n}_p$ is computed using the Rodrigues rotation formula [45]. Finally, the translation vector $\mathbf{t}$ is determined such that the fitted plane aligns with the world origin, ensuring all scene coordinates are expressed in the same frame.

3.2.2. Geometrical priors

Our data generation algorithm produces a physically grounded, geometry-aware dataset from a rich source of information: object 3D poses. First we compute the 3D cuboid bounding boxes from each scene using the object pose and model sizes. Motion trajectories are synthesized using a Rapidly-exploring Random Tree (RRT) planner [28] operating in 3D Cartesian space to compute collision-free pick-and-place paths between object locations. For n objects in a scene, we generate $\binom{n}{2}$ inter-object trajectories by initializing the planner at one object's centroid and applying a 10% goal-biased random sampling toward the second object. Paths that intersect with neighboring object meshes—derived from their 3D poses—are filtered out. The raw RRT paths are then refined using the Ramer–Douglas–Peucker algorithm [11] for 3D path simplification, reducing redundant waypoints and producing smooth, physically plausible trajectories.

Object grasps are computed in the reconstructed world frame using the transformer-based parallel gripper model M2T2 [61], which employs a dual-sampling strategy combining global scene points for contextual reasoning and object-centric points for accurate geometric representation. We retain the top-5 grasps per object to ensure diversity in contact configurations. If all predicted grasps for an object result in collisions with surrounding objects, these are discarded and the object is labeled as *fully cluttered*. This labeling forms the foundation of the *object rearrangement* skill, where the VLM must identify which objects should be moved first to enable a successful grasp. Since the data source we are injecting includes multiple camera viewpoints, the resulting combination of poses, grasps, and motion trajectories yields a densely annotated, geometrically consistent dataset.

3.2.3. Generating Question-Answer pairs

Once the geometrical priors are generated, we form question and answer pairs using object description and templates. First, we render the 3D models used in the scene, these renders are sent to a VLM to acquire object descriptions including: shape, color, size, and general utility. Each description is then manually verified and refined when necessary to ensure accuracy and linguistic diversity. The metadata generated in Section 3.2.2 serves as input for constructing VQA templates. We design a distinct template for each reasoning task, following the structure {TASK_TYPE} {OBJECT A} {OBJECT B}. These templates are then supplied to the LLM, together with carefully curated in-context examples, to generate linguistically diverse and human-like questions. Robotic scenes often contain multiple instances of the same object category, where visual cues such as color, shape, or texture alone cannot uniquely disambiguate the target. To distinguish multiple instances of the same object category within a scene, we compute the centers of their 3D bounding boxes and assign relative positional attributes {POS. ATTRIBUTE} such as "leftmost", "rightmost", "topmost", and "bottommost". These attributes are then concatenated with the above question generation template for objects that have multiple instances in a particular scene.

Object poses, grasps, trajectories, and rearrangements are represented as ordered 2D lists of keypoints. Spatial reasoning and relative depth perception tasks are annotated as binary ("yes"/"no") responses. The final dataset comprises approximately 33M high-quality VQA pairs spanning the six object-interaction reasoning skills.

Table 2. Distribution of questions and frequency in BOP-Ask

Question Type	# Q&As	% of Total
Object Poses	4.2M	12.4%
Grasps	5.4M	16.0%
Trajectories	5.4M	16.0%
Object rearrangement	2.6M	7.6%
Spatial Reasoning	5.4M	16.0%
Relative Depth Perception	10.8M	32.0%
Total	33.8M	100%

4. BOP-Ask: A Spatial Reasoning Benchmark for Object-Interaction Reasoning

We choose the BOP [38] family of datasets which provide both real and simulated images comprising diverse table-top scenes in indoor and outdoor environments. Our data generation pipeline results in three main datasets: a training dataset (BOP-Ask) and two test datasets: (BOP-Ask-core and BOP-Ask-lab). In this section, we present the three datasets with their accompanying task metrics.

BOP-Ask We utilize images and pose annotations from HOPE [55], HANDAL [16], YCB-V [59], and LineMOD [18] as our base source for generating rich spatio-geometric visual question-answers (VQAs). These datasets were chosen to capture sufficient scene and object diversity, variations in camera poses and clutter as observed in real-world robot environments. Our datasets span high-clutter and geometric variability (HOPE), increased object diversity (HANDAL), and lower-resolution but varied scenes (YCB-V and LineMOD), providing a broad range of visual conditions.

Our final dataset includes 104 unique household and industrial objects arranged in high degrees of clutter and dense environments. Many objects contain textual labels such as "Milk" or "BBQ Sauce" which make them easier to identify when viewed from the front. However, occlusions and randomized arrangements demand global scene understanding and referring attributes such as color and position to uniquely identify the objects. Some scenes also contain distractors (multiple instances of the same category) making visual grounding challenging which affects subsequent grasping and trajectory predictions. After rigorous manual (established from different dataset metrics) and occlusion filtering for grasps and trajectories, our dataset contains 150K unique images and 33M QA pairs. Table 2 summarizes the distribution of question types across the dataset. While the data pipeline prioritizes extracting maximum number of high-quality spatio-geometric annotations for adapting VLMs for spatial indoor tasks, our algorithm can be easily extended to novel scenes with object poses, which is trivial to obtain using simulated environments.

BOP-Ask Test sets. We release BOP-Ask-core and BOP-Ask-lab, two test corpora comprising high-quality VQA pairs. Specifically for BOP-Ask-core we use held-out scenes from BOP and each VQA pair is manually verified using a semi-automated pipeline where the annotations created by the framework are manually updated if necessary. In total, we release 688 VQA pairs as a part of our testing benchmark, with the following distribution: object poses (17.4%), object grasps (17.4%), trajectories (17.4%), spatial reasoning (17.4%), relative depth perception (23.3%) and object rearrangement (7.1%). We also release BOP-Ask-lab with images not present in BOP datasets to measure the performance on out-of-domain objects. This corpus is manually annotated and comprises 240 VQA pairs from 15 images.

Evaluation Metrics. We adopt standard 3D metrics and success rates to evaluate model performance across all spatial reasoning tasks in BOP-Ask. For *object pose estimation*, we report the Intersection-over-Union (IoU) metric, which measures the volumetric overlap between the 3D projected cuboids of the predicted and ground-truth poses. For *trajectory prediction*, we employ two complementary metrics: (i) *success rate*, defined as 1 if the initial and final points of the predicted trajectory lie on the corresponding source and target objects, and (ii) *distance error*, computed as the mean per-pixel distance between predicted and ground-truth waypoints. The success rate evaluates whether the model correctly identifies the relevant object pair for manipulation, while the distance error quantifies the geometric accuracy of the predicted trajectory. While multiple valid motion paths may exist between two objects, distance error provides a continuous measure of how closely the predicted trajectory aligns with the ground-truth path. For *object rearrangement*, we report Recall (%)—the proportion of correctly identified obstructing objects that must be moved relative to the total number of required moves. Finally, for *spatial reasoning* and *relative depth perception*, we compute the average success rate over all VQA pairs.

Samples in BOP-Ask adopt a grasp representation distinct from conventional 2D grasp datasets. We employ a five-point formulation that specifies the grasp center, left finger base, right finger base, left finger tip, and right finger tip. This representation captures both position and orientation more comprehensively than rectangle-based [3, 60] or point-based affordance formulations [10], offering a more flexible transition to 3D by allowing greater freedom in selecting the grasp plane.

Table 3. Performance of popular VLMs on BOP-Ask-core, which features cluttered object configurations requiring fine-grained spatio-geometric reasoning for robotics. Note that 'inf' indicates that the model did not produce a valid output. Bold and underlined scores indicate the best and second best performing methods for a particular task.

Method	Pose	Trajectory		Grasps	Spatial	Rel. Depth	Obj. Rearr.
	3D IOU ↑	SR ↑	Dist. Err. ↓	NCE ↓	SR ↑	SR ↑	Recall (%) ↑
Human	54.2	**67.3**	112.4	**1.1**	84.9	87.3	44.1
Proprietary Models							
GPT-5 [40]	9.0	0	inf	inf	68.3	74.6	14.8
Gemini Robotics-ER 1.5 [53]	24.4	43.0	138.4	4.2	84.2	88.0	48.9
Open-Source Models							
Molmo [9] (72B)	22.6	13.2	224.8	<u>1.4</u>	68.3	65.6	28.0
Qwen-VL 2.5 [2] (3B)	26.5	0	240.6	3.1	50.8	49.4	15.2
NVILA [33] (2B)	6.5	0	inf	8.2	65.0	52.5	8.3
NVILA [33] (15B)	27.2	6.2	206.3	5.3	75.0	65.0	25.0
RoboRefer [65]	34.3	0	inf	inf	81.7	84.0	16.6
Intern-VL [7]	18.3	0	inf	inf	72.5	76.0	8.6
Models Trained on BOP-Ask							
Qwen-VL 2.5 (3B) - SFT	48.2	22.5	116.3	1.5	92.6	<u>94.1</u>	43.4
NVILA (2B) - SFT	**77.4**	50.8	<u>78.5</u>	1.69	<u>94.2</u>	**94.6**	<u>56.4</u>
NVILA (15B) - SFT	<u>73.5</u>	<u>64.2</u>	**77.4**	1.40	**95.8**	**94.6**	**57.7**

To evaluate grasp prediction performance, we use the *normalized coordinate error* (NCE) metric, which measures the mean normalized Euclidean distance between corresponding predicted and ground-truth grasp points:

$$\mathrm{NCE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\|p_i - \hat{p}_i\|_2}{d},$$

where $N = 5$, p_i and $\hat{p}_i$ denote the predicted and ground-truth grasp points, respectively, and d is the gripper width. Each point is normalized by the image width and height to ensure scale invariance.

5. Experiments

We evaluate a diverse set of proprietary and open-source VLMs on our test suites to quantify performance gaps in the novel spatial reasoning and object interaction skills introduced in BOP-Ask. In addition, we fine-tune two open-source VLMs and demonstrate the performance gains and downstream utility of training on our dataset.

5.1. Setup and Baselines

We benchmark a range of widely adopted VLMs, including Molmo [9], Qwen-VL 2.5 [2], NVILA [33], RoboRefer [65], and Intern-VL [7]. For proprietary systems, we use GPT-5 [40] and Gemini 1.5 ER [53] as closed-source baselines. We fine-tune Qwen-VL 2.5 and NVILA using their official training codebases on BOP-Ask. Closed-source models are evaluated via their official APIs, while open-source models are tested using their publicly released implementations. All open-source model inferences are run on a single NVIDIA A100 GPU, and fine-tuning experiments are conducted on a cluster of eight A100 GPUs with default hyperparameters.

5.2. Human Benchmark

While humans can typically answer simple yes/no questions on spatial or depth reasoning with high accuracy, the images in BOP-Ask present substantial challenges due to heavy clutter, occlusions, and complex scene layouts. Moreover, tasks such as object pose estimation, grasp prediction, and trajectory reasoning require an understanding of *semantic 3D stability*. For example, the appropriate grasp pose for a knife depends on task intent—grasping by the blade for a handover versus by the handle for cutting. To establish a comprehensive human performance baseline on BOP-Ask-core, we developed a simple annotation interface for the images in the test set and distributed it to anonymous contributors. Participants were asked to complete tasks such as marking precise 3D object poses, annotating grasp positions, drawing inter-object motion trajectories, identifying which objects must be moved to grasp a target object in a cluttered scene, and answering binary yes/no questions for spatial reasoning. Human annotations were evaluated using the same metrics applied to the VLM baselines, enabling direct comparison across all six object-interaction reasoning tasks. 1037 responses were gathered from 40+ contributors, covering all 688 VQAs in BOP-Ask-core. Each participant annotated between 10 to 25 questions spanning different task types to ensure balanced coverage. Roughly 30% of the questions were answered by multiple annotators, and their scores were averaged to

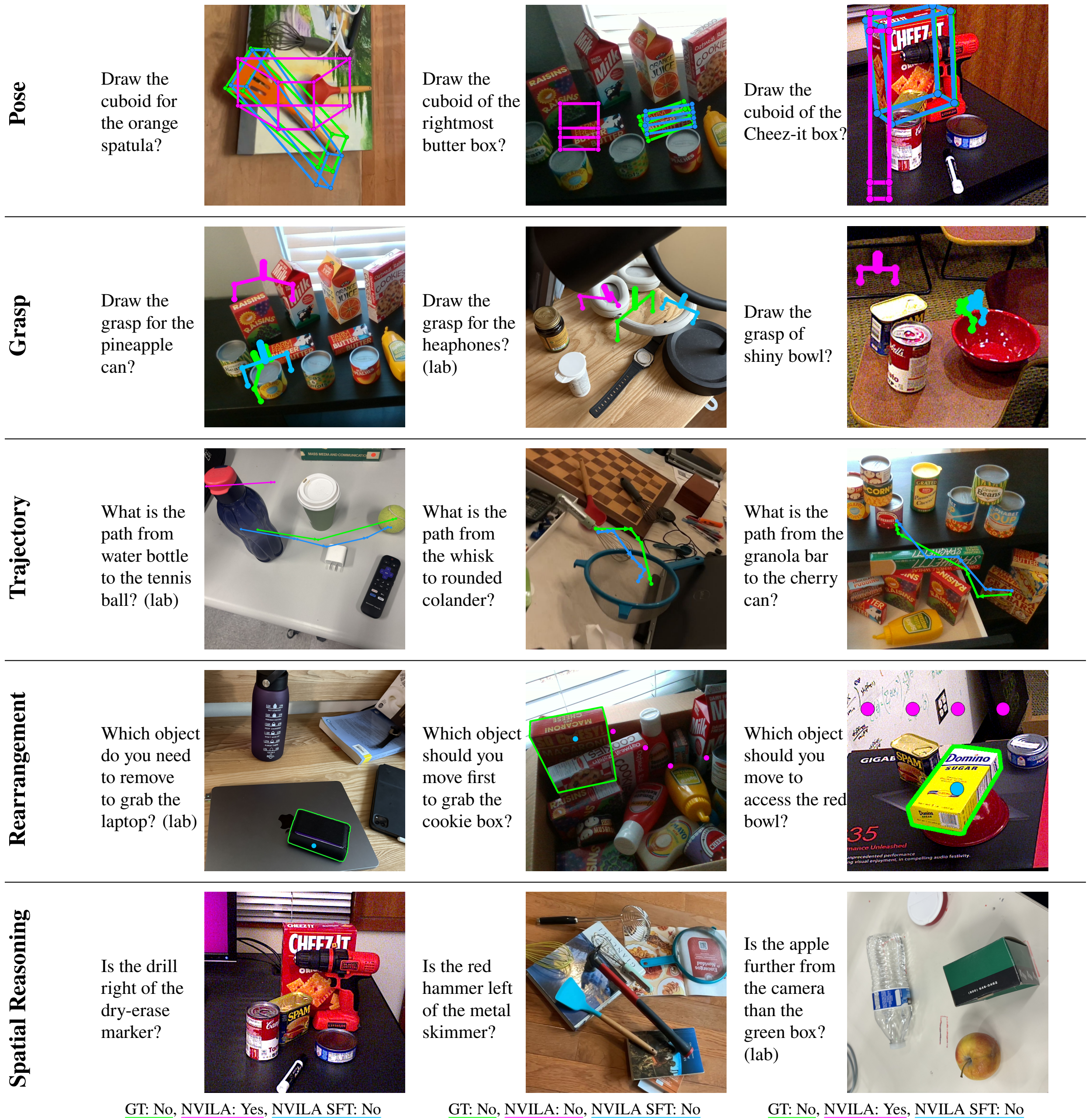

Figure 3. Predictions from samples in BOP-Ask-core and BOP-Ask-lab (identified by (lab)), showing improvements gained from fine-tuning on BOP-Ask. Predictions from NVILA (shown in magenta) and NVILA SFT (shown in blue) are shown alongside the Ground Truth (in green). For the 'Rearrangement' task, the Ground Truth shape delineates the area of valid predictions. Absence of a colored prediction indicates none was made or it was out of frame. Images are from HOPE [55], HANDAL [16], and YCB-V [59].

obtain a single human baseline per question. We computed mean performance across tasks to derive the human benchmark reported in Table 3.

5.3. Results

BOP-Ask-core. Training on BOP-Ask leads to consistent performance gains across all six evaluated tasks, as summarized in Table 3. For *object pose estimation*, open-source VLMs typically default to predicting 2D bounding boxes rather than full 3D cuboid poses. Fine-tuning on our dataset enables NVILA and Qwen-VL to adapt to this richer prediction space while preserving their inherent visual grounding capabilities, resulting in more precise and physically consistent pose predictions. *Trajectory generation* and *grasp prediction* emerge as challenging skills, primarily because existing pretraining corpora lack supervision for continuous spatial motion or 6-DoF grasp understanding. Training on BOP-Ask exposes these models to explicit geometric cues, improving their ability to reason about motion continuity and feasible grasp locations. While Gemini and Molmo exhibit strong performance on pointing-based tasks, they occasionally misidentify objects in scenes with high clutter or visual distractors. *Spatial reasoning* and *relative depth perception* tasks show higher accuracies, as these typically reduce to binary or comparative judgments (e.g., "left of," "closer"), which are well-aligned with the relational reasoning priors acquired during pretraining. Finally, *object rearrangement* remains a difficult task: the best-performing models only reach about 60% accuracy. This task demands fine-grained understanding of object–object relationships, 3D coordinate alignment, and clutter dynamics—highlighting a key opportunity for future progress in physically grounded reasoning. Figure 3 presents qualitative examples across all six tasks, comparing predictions from two representative models against the ground truth. Examples across different rows illustrate that while the baseline NVILA model (magenta) struggles with complex reasoning tasks, fine-tuning on BOP-Ask enables accurate geometric priors (blue) essential for robotic manipulation.

Table 4. Comparing models trained on BOP-Ask with baselines on OOD test sets. RS-H [50], CV-B [54], SB [4]. RoboSpatial results are from original paper. "–" denotes unavailable scores. Metrics: Pose (IoU), Grasping (Gr; NCE ↓), Trajectories (Traj; success rate), Spatial–Depth (S-D; accuracy).

Models	RS-H	CV-B	SB	BOP-Ask-lab			
				Pose	Gr	Traj	S-D
RoboSpatial [50]	78.0	-	<u>64.7</u>	-	-	-	-
NVILA [33]	63.4	78.2	47.5	6.1	4.2	0	70.0
+ BOP-Ask	69.1 ↑	<u>89.3</u> ↑	50.0 ↑	<u>16.2</u> ↑	**1.1** ↑	<u>28.2</u> ↑	<u>81.2</u> ↑
Qwen-VL 2.5 [2]	<u>78.1</u>	88.8	60.0	12.6	3.6	0	74.4
+ BOP-Ask	**81.3** ↑	**92.4** ↑	**65.0** ↑	**25.3** ↑	<u>1.3</u> ↑	**37.1** ↑	**85.8** ↑

BOP-Ask-lab & out of distribution Table 4 reports the improvement observed on models fine-tuned on BOP-Ask and tested on BOP-Ask-lab. We observe improvement on all metrics for both Qwen-VL 2.5 and NVILA, demonstrating that training on BOP-Ask enhances 3D grounding and manipulation capabilities, while underscoring remaining challenges of open-world grounded interaction, see Figure 3 for examples taken from BOP-Ask-lab. Although BOP-Ask focuses on indoor scenes, its spatial and depth reasoning tasks enable strong domain transfer to out-of-domain benchmarks. We evaluate fine-tuned models on three popular spatial reasoning datasets—*Configuration* from RoboSpatial-HOME [50], *Relation* task from CV-Bench [54], and *Reach* task from SpatialBench [4] — chosen for their close alignment with the spatial reasoning tasks in BOP-Ask such as *"Is the picture above the desk?"*. Table 4 (left-side) reports the results for the RoboSpatial [50] baseline, and Qwen-VL 2.5 [2] and NVILA [33] before and after fine-tuning on BOP-Ask. Fine-tuned models achieve consistent accuracy gains, highlighting effective zero-shot transfer of spatial reasoning capabilities learnt from BOP-Ask.

Robot Experiments To evaluate real-world performance, we deployed models on 15 pick-and-place tasks executed by a Franka robot (e.g., "Move the tomato sauce can to the location of the green box"). Across all tasks, the base NVILA failed to complete any, whereas NVILA fine-tuned on BOP-Ask succeeded on 10 out of 15. Additional details are provided in the supplementary material.

Table 5. Ablations on incremental training of NVILA on images within BOP-Ask. **BA-YCBV:** YCB-V only; **BA-YCBV+H:** YCB-V + HANDAL; **BA-YCBV+H+L:** YCB-V + HANDAL + LINEMOD; **BA-NoSpatDep:** BOP-Ask excluding spatial and depth (yes/no) questions.

Method	Pose IOU ↑	Traj. SR ↑	Grasps NCE ↓	Spat. SR ↑	Depth SR ↑	Rearr. Rec. (%) ↑
NVILA (Base)	6.5	0	8.15	65.5	52.5	8.3
+ BA-YCBV	31.7	24.2	6.37	65	75.0	16.4
+ BA-YCBV+H	54.4	30.8	2.85	87.5	90.2	21.8
+ BA-YCBV+H+L	67.2	<u>51.8</u>	2.02	<u>93.5</u>	<u>92.6</u>	39.2
+ BOP-Ask	<u>77.4</u>	**64.2**	**1.40**	**94.2**	**94.6**	**57.7**
+ BA-NoSpatDep	**78.2**	50.0	<u>1.69</u>	62.5	50.6	<u>50.3</u>

Ablations on data recipe. We evaluate the effect of combining data from multiple BOP families—HOPE [55], HANDAL [16], YCB-V [59], and LINEMOD [18]—and analyze performance across question types (Table 5). Performance consistently improves as additional datasets are introduced, suggesting that increased visual and semantic diversity—spanning object geometries, textures, and spatial layouts—enhances generalization in fine-grained spatial reasoning. The base model initially fails in trajectory questions but quickly adapts during fine-tuning, reaching 24% accuracy using only VQAs from YCB-V. This underscores spatial pre-training, which provides transferable visual grounding for new tasks. Subsequent inclusion of additional datasets further boosts performance across all skills. Notably, removing binary spatial and depth reasoning tasks (*BA-NoSpatDep*) leads to a drop in overall performance, confirming that such auxiliary tasks strengthen object-interaction reasoning through multi-skill co-training.

6. Conclusion

We presented BOP-Ask, a large-scale training dataset that introduces object-interaction reasoning. It includes tasks such as 3D object poses, grasp affordances, motion trajectories, and object rearrangements. In addition, we introduced two complementary benchmarks, BOP-Ask-core and BOP-Ask-lab, to evaluate models in seen and unseen environments. Experiments on in-domain and out-of-domain test sets show the rich spatio-geometric priors encoded in BOP-Ask enhance trained models ability to reason and act upon objects in complex scenes.

7. Acknowledgments

We are grateful to Huy Le for supporting an earlier stage of this dataset. We also thank Naren Devarakonda, Amrita Mazumdar, Alex Zook, Siyi Chen, Chan Hee (Luke) Song, and Faisal Ladhak for their helpful input on this manuscript. Also thank you to all the anonymous participant to our human study. This paper is supported in part by the New York University Abu Dhabi (NYUAD) Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award CG010, and by NSF under grant number 2208189.

References

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 8

[3] Vineet Bhat, Naman Patel, Prashanth Krishnamurthy, Ramesh Karri, and Farshad Khorrami. Maplegrasp: Mask-guided feature pooling for language-driven efficient robotic grasping. *arXiv preprint arXiv:2506.06535*, 2025. 5

[4] Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2025. 8

[5] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465, 2024. 1, 2, 3

[6] Kaiyuan Chen, Shuangyu Xie, Zehan Ma, Pannag R Sanketi, and Ken Goldberg. Robo2vlm: Improving visual question answering using large-scale robot manipulation data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 3

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 6

[8] 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 models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2, 3

[9] Matt Deitke, Christopher Clark, Sangho Lee, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 4, 6

[10] Abhay Deshpande, Yuquan Deng, Arijit Ray, Jordi Salvador, Winson Han, Jiafei Duan, Kuo-Hao Zeng, Yuke Zhu, Ranjay Krishna, and Rose Hendrix. Graspmolmo: Generalizable task-oriented grasping via large-scale synthetic data generation, 2025. 5

[11] David H. Douglas and Thomas K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 10(2):112–122, 1973. 4

[12] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2024. 2, 3

[13] Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. AHA: A vision-language-model for detecting and reasoning over failures in robotic manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[14] Kuan Fang, Fangchen Liu, Pieter Abbeel, and Sergey Levine. Moka: Open-world robotic manipulation through mark-based visual prompting. *Robotics: Science and Systems (RSS)*, 2024. 1

[15] Xingyu Fu, Yushi Hu, Bangzheng Li, et al. Blink: Multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 148–166. Springer, 2024. 2, 3

[16] Andrew Guo, Bowen Wen, Jianhe Yuan, Jonathan Tremblay, Stephen Tyree, Jeffrey Smith, and Stan Birchfield. Handal: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11428–11435, 2023. 5, 7, 8

[17] Yi Han, Cheng Chi, Enshen Zhou, Shanyu Rong, Jingkun An, Pengwei Wang, Zhongyuan Wang, Lu Sheng, and Shanghang Zhang. Tiger: Tool-integrated geometric reasoning in vision-language models for robotics. *arXiv preprint arXiv:2510.07181*, 2025. 3

[18] Stefan Hinterstoißer, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Kurt Konolige, Gary R. Bradski, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian Conference on Computer Vision*, 2012. 5, 8

[19] Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9488–9495, 2024. 3

[20] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of

relational keypoint constraints for robotic manipulation. In *8th Annual Conference on Robot Learning*, 2024. 3

[21] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709, 2019. 3

[22] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1724–1734, 2025. 2

[23] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 3

[24] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[25] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 2, 3

[26] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, et al. Openvla: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning (CoRL)*, pages 2679–2713. PMLR, 2025. 3

[27] 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. 3

[28] Steven M. LaValle and James J. Kuffner. Rapidly-exploring random trees: A new tool for path planning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 1999. 4

[29] Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025. 2, 3

[30] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500, 2023. 1

[31] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. In *Advances in Neural Information Processing Systems*, 2024. NeurIPS. 3

[32] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 2, 3

[33] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Haotian Tang, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Jinyi Hu, Sifei Liu, Ranjay Krishna, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4122–4134, 2025. 6, 8

[34] Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. *Advances in neural information processing systems*, 37:68803–68832, 2024. 3

[35] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. 3

[36] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3d vision language reasoning. In *CVPR*, 2024. 3

[37] Soroush Nasiriany, Fei Xia, Wenhao Yu, et al. Pivot: iterative visual prompting elicits actionable knowledge for vlms. In *Proceedings of the International Conference on Machine Learning (ICML)*. JMLR.org, 2024. 3

[38] Van Nguyen Nguyen, Stephen Tyree, Andrew Guo, Mederic Fourmy, Anas Gouda, Taeyeop Lee, Sungphill Moon, Hyeontae Son, Lukas Ranftl, Jonathan Tremblay, et al. Bop challenge 2024 on model-based and model-free 6d object pose estimation. *arXiv preprint arXiv:2504.02812*, 2025. 2, 5

[39] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, et al. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024. 3

[40] OpenAI. Introducing gpt-5. `https://openai.com/index/introducing-gpt-5/`, 2025. OpenAI Blog. 6

[41] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, et al. Gpt-4 technical report, 2024. 3

[42] Navid Rajabi and Jana Kosecka. Towards grounded visual spatial reasoning in multi-modal vision language models, 2024. 3

[43] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12977–12987, 2024. 3

[44] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Dynamic spatial aptitude training for multimodal language models, 2025. 3

[45] Olinde Rodrigues. Des lois géométriques qui régissent les déplacements d'un système solide dans l'espace, et de la variation des coordonnées provenant de ces déplacements considérés indépendamment des causes qui peuvent les produire. *Journal de mathématiques pures et appliquées*, 5:380–440, 1840. 4

[46] Leonard Salewski, A. Sophia Koepke, Hendrik P. A. Lensch, and Zeynep Akata. Clevr-x: A visual reasoning dataset for natural language explanations. In *xxAI - Beyond explainable Artificial Intelligence*, pages 85–104. Springer, 2022. 3

[47] Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, et al. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024. 2, 3

[48] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530, 2023. 1, 3

[49] Chan Hee Song, Jihyung Kil, Tai-Yu Pan, Brian M. Sadler, Wei-Lun Chao, and Yu Su. One step at a time: Long-horizon vision-and-language navigation with milestones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15482–15491, 2022. 3

[50] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15768–15780, 2025. 2, 3, 8

[51] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, 2019. Association for Computational Linguistics. 3

[52] Emilia Szymanska, Mihai Dusmanu, Jan-Willem Buurlage, Mahdi Rad, and Marc Pollefeys. Space3D-Bench: Spatial 3D Question Answering Benchmark. In *European Conference on Computer Vision (ECCV) Workshops*, 2024. 3

[53] Gemini Robotics Team. Building the next generation of physical agents with gemini robotics-er 1.5. https://developers.googleblog.com/en/building-the-next-generation-of-physical-agents-with-gemini-robotics-er-15/, 2025. Google Developers Blog. 2, 6

[54] 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, 3, 8

[55] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022. 5, 7, 8

[56] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. Gpt-4v(ision) for robotics: Multimodal task planning from human demonstration. *IEEE Robotics and Automation Letters*, 9(11): 10567–10574, 2024. 3

[57] Tai Wang, Xiaohan Mao, Chenming Zhu, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[58] Youngsun Wi, Mark Van der Merwe, Pete Florence, Andy Zeng, and Nima Fazeli. Calamari: Contact-aware and language conditioned spatial action mapping for contact-rich manipulation. In *7th Annual Conference on Robot Learning*, 2023. 3

[59] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Proceedings of Robotics: Science and Systems*, 2017. 5, 7, 8

[60] Houjian Yu, Mingen Li, Alireza Rezazadeh, Yang Yang, and Changhyun Choi. A parameter-efficient tuning framework for language-guided object grounding and robot grasping. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14353–14360, 2025. 5

[61] Wentao Yuan, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. M2t2: Multi-task masked transformer for object-centric pick and place. *arXiv preprint arXiv:2311.00926*, 2023. 4

[62] 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 in robotics. In *8th Annual Conference on Robot Learning*, 2024. 1, 2

[63] Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. SPARTUN3d: Situated spatial understanding of 3d world in large language model. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[64] Enyu Zhao, Vedant Raval, Hejia Zhang, Jiageng Mao, Zeyu Shangguan, Stefanos Nikolaidis, Yue Wang, and Daniel Seita. Manipbench: Benchmarking vision-language models for low-level robot manipulation. *arXiv preprint arXiv:2505.09698*, 2025. 3

[65] Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and Shanghang Zhang. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics, 2025. 2, 3, 6

[66] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, et al. Rt-2: Vision-language-action models transfer web knowl-

edge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, pages 2165–2183. PMLR, 2023. 3