MANIPULATION

A retrieval-augmented framework enabling VLM spatial awareness for object-centric robot manipulation

Kai Chen[1], Chengkun Li[1], Chang Tu[1], Jiahui Pan[1], Yiyao Ma[1], Wei Chen[2], Zhongxiang Zhou[3], Xuecheng Xu[3], Stephen James[4], Chi-Wing Fu[1], Rong Xiong[3,5], Pieter Abbeel[6], Yun-Hui Liu[2], Qi Dou[1]*

Connecting the semantic reasoning of vision-language models (VLMs) to the precise geometric demands of robotic manipulation remains a fundamental challenge. Although VLMs can interpret high-level commands, they lack the intrinsic spatial intelligence required for tasks demanding precise object placement, orientation, and physical reasoning. Here, we introduce Retrieval-Augmented Manipulation (RAM), an object-centric framework that endows general-purpose vision foundation models with the spatial reasoning necessary for robust manipulation. RAM bridges the semantic-to-geometric gap by grounding abstract concepts into an explicit, object-centric three-dimensional (3D) representation. This grounded information is then provided as augmented context to the VLM, empowering it to decompose complex instructions into a sequence of spatially precise and physically plausible subgoals. We demonstrate that RAM, in a zero-shot setting on a real-world robot, can execute these subgoals to fulfill complex spatial language instructions, complete spatially aware manipulation under the guidance of a single 2D image, and adaptively replan tasks by reasoning about physical constraints like object size and collisions. Quantitative evaluations on the Common Object in 3D (CO3D) dataset also validated that RAM's core vision module generalizes to previously unseen object categories and is robust to variations in shape and occlusions. By providing a structured bridge between semantic intent and geometric execution, RAM represents a critical step toward developing more physically intelligent and general-purpose robotic systems.

INTRODUCTION

The long-standing pursuit of creating general-purpose robots capable of seamlessly assisting humans in diverse and unstructured settings has recently been invigorated by transformative advances (*1*–*4*). At the forefront of this revolution are vision-language models (VLMs), which have demonstrated a remarkable capacity for parsing high-level, abstract human commands and decomposing them into logical sequences of subtasks (*5*, *6*). By leveraging the vast knowledge embedded in internet-scale data (*7*, *8*), these models can function as a central "brain" for robots, enabling them to reason about complex, long-horizon goals, from brewing a cup of coffee to tidying up the living room (*9*–*11*). This breakthrough in semantic understanding represents a pivotal step toward building more versatile and intelligent robotic systems.

However, a fundamental chasm remains between the semantic plans generated by VLMs and the physical realities of robotic manipulation. At its core, robotic manipulation is an inherently spatial endeavor. Its success is not merely determined by what to do but critically by how and where, which requires precise reasoning about object poses, contact points, and relational configurations in a three-dimensional (3D) world (*12*–*21*). Current VLMs, although proficient in high-level reasoning, often lack the fine-grained spatial awareness needed to inform these crucial geometric details. This creates a critical gap between abstract intent and concrete physical execution, representing the primary bottleneck that hinders the deployment of autonomous robots in spatially aware manipulation scenarios, in which tasks would have a higher demand on precision, reliability, and safety.

In response to this challenge, many works have attempted to bridge this gap by augmenting VLM-based systems with specialized perception modules. One common strategy is to use language-driven expert models to predict task-oriented grasps (*22*, *23*) or to generate affordance maps that highlight functional regions relevant to the command (*24*–*27*). However, these methods often decouple high-level planning from low-level geometric validation. This separation forces the VLM to formulate its plan in a physical vacuum, unaware of the feasibility of subsequent steps, which can lead to suboptimal or impossible-to-execute actions. Another direction seeks to improve the VLM's planning context by providing it with enhanced visual inputs. Examples include abstracting the scene into object key points (*28*) or overlaying orientation vectors onto the image to guide the VLM spatial reasoning (*29*, *30*). These approaches typically provide local cues. Such fragmented information often does not allow the VLM to have a holistic understanding of the physical scene, making it difficult to reason about the complex, multiobject spatial relationships required in long-horizon tasks.

The limitations of these approaches point to a more fundamental problem, rooted in the nature of how VLMs acquire and represent knowledge. Primarily trained on vast corpora of 2D images and text from the internet, these models lack direct, grounded experience with the 3D physics and geometry that govern the real world. Consequently, when prompted for spatially aware details, such as stable placements, functional parts, or precise orientations, they are often forced to "hallucinate," generating plausible-sounding but physically inaccurate or unreliable information (*31*). Although recent efforts to fine-tune these models on 3D data aim to embed this knowledge directly (*17*, *32*, *33*), this approach immediately confronts a steep practicality barrier rooted in data. The combinatorial complexity of

[1]Department of Computer Science and Engineering, Chinese University of Hong Kong, HKSAR, China. [2]Department of Mechanical and Automation Engineering, Chinese University of Hong Kong, HKSAR, China. [3]Zhejiang Humanoid Robot Innovation Center Co. Ltd., Ningbo, Zhejiang, China. [4]Imperial College London, London, UK. [5]College of Control Science and Engineering, Zhejiang University, Hangzhou, Zhejiang, China. [6]University of California, Berkeley, CA, USA.
*Corresponding author. Email: qidou@cuhk.edu.hk

3D objects—their functional parts, orientations, and physical interactions—demand training data on a scale that grows exponentially beyond that of 2D images and text. Moreover, unlike the readily available corpora of the web, high-fidelity 3D scenarios and contextualized interaction data remain scarce and prohibitively expensive to acquire. As a result, the current implicit and opaque spatial understanding obtained via fine-tuning on 3D data remains quite brittle. This implicit knowledge is not only difficult to verify or interpret but also lacks reliability when encountering previously unseen objects or configurations.

To circumvent this problem, we propose a shift in methodology. Instead of attempting to embed all physical knowledge in the VLM's parameters, we advocate for a framework that augments the model's spatial awareness with an explicit, verifiable, and controllable source of external knowledge. It is realized through a retrieval-augmented approach (*34*–*38*). At its core, this strategy empowers the VLM to actively query a structured knowledge base at inference time, grounding its abstract task plans with precise, factual information. The crux of our proposed framework lies in the nature of this knowledge base, which we structure as category-level, object-centric priors. This object-centric formulation ensures that the retrieved information is directly relevant to the manipulation task, and the category-level abstraction provides the critical ability to generalize to unseen object instances in a known class, such as any bowl rather than just a specific one. These priors encapsulate essential geometric and functional properties, such as canonical coordinates, stable grasping, and functional regions, that are vital for any spatially aware manipulation task.

To this end, we instantiate a Retrieval-Augmented Manipulation (RAM) framework that bridges the gap between semantic reasoning and physical execution from a unified, object-centric perspective. As shown in Fig. 1, RAM is built on an external, extensible object-centric engine. This engine maintains a library of canonical shape templates, one for each object category, annotated with a rich vocabulary of geometry-relevant information, such as canonical coordinate map and symmetry, as well as manipulation-relevant knowledge like stable grasp poses and functional planes. We then leverage a boosted vision foundation model to construct a 3D visual grounding model that transfers the rich geometry- and manipulation-relevant knowledge to various objects of different shapes and previously unseen categories in a generalizable manner. The grounded, object-centric knowledge then is provided as augmented context to the VLM for task planning. Given this augmented context, the VLM planning and motion trajectory optimization are integrated via shared object-centric spatial constraints. This integration transforms the VLM from a naive planner into a spatially aware reasoner, which decomposes complex task instructions into a sequence of spatially precise and physically plausible subgoals expressed with object-centric spatial constraints. These planned constraints are then realized via a trajectory optimization scheme, in which we expand the supported spectrum of geometric primitives and spatial relationship constraints. In this way, our method effectively translates task instructions into executable robot trajectories, thereby accommodating diverse manipulation tasks characterized by complex spatial relationships. We evaluated RAM across 14 distinct, spatially aware manipulation scenarios. These tasks were designed to span a diverse range of capabilities, including spatial instruction following, image-guided spatially aware manipulation, and tasks that require complex spatial reasoning. The results demonstrate that RAM can enhance the spatial awareness of general VLMs for spatially aware robotic manipulation.

RESULTS

System overview

RAM was designed to endow a general VLM with robust spatial intelligence at inference time. Our implementation was built on Gemini-2.5-Pro (*39*) as the core VLM. The central tenet of RAM was a process whereby retrieved object-centric geometric and manipulation knowledge informed and refined the VLM's task planning, bridging the gap between task planning and concrete physical execution. As shown in Fig. 1, this was achieved through three interconnected modules.

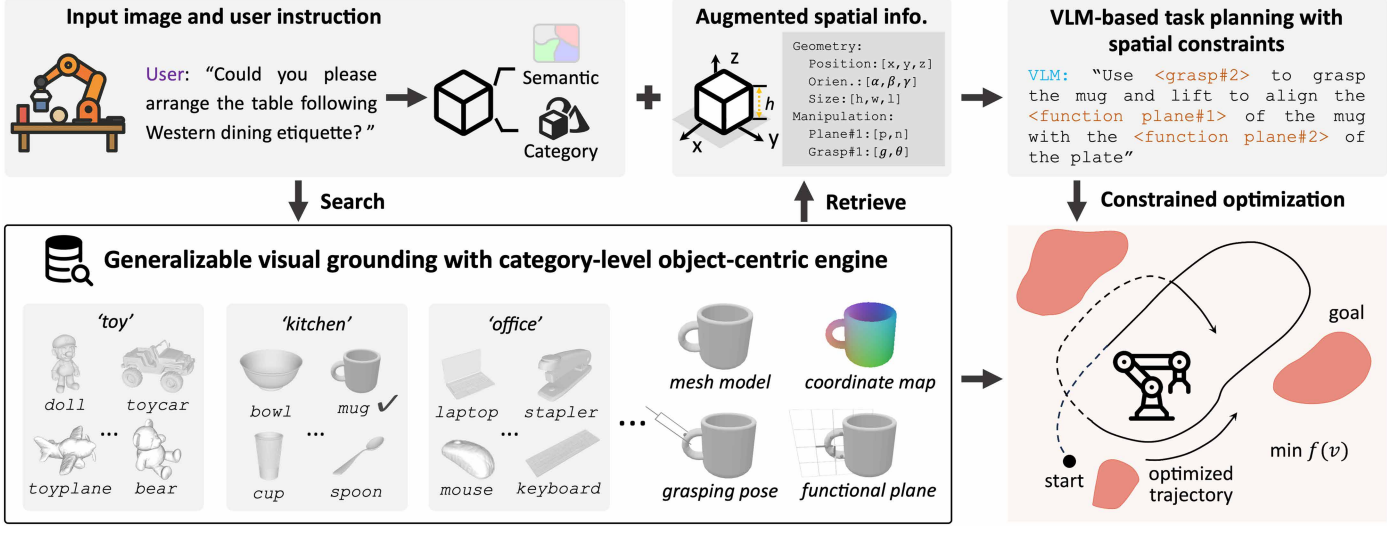

Fig. 1. Overview of the RAM framework. Given the image and user instruction, RAM first parses the command and identifies the task and involved object. To acquire the necessary spatial knowledge, RAM then queries the object-centric knowledge base with the object category. The retrieved category-level priors, including canonical pose, size, stable grasping configurations, and functional regions, are then grounded to the specific object instance in the scene. On the basis of the augmented context, the VLM will further decompose the task into substeps with explicit spatial constraints, and the robot trajectory will be optimized with constraints for task execution.

Visual grounding with category-level object-centric priors

Given an RGB-D (red, green, blue, depth) image and a user instruction, this module first used a 2D grounding model (*40*) to segment the objects relevant to the task. On the basis of these segmented regions, it then used a 3D category-level grounding model built on DINO-v2 (*41*). The pretrained DINO-v2 features were further augmented via a lifting module to construct discriminative 3D object representations, enabling the transfer of rich geometry and manipulation priors from canonical templates to the observed object instances. Such representations allowed the model to be robust to intraclass object shape variations and generalizable to unseen object categories. As depicted in Fig. 1, these transferred priors included object pose and size, stable grasping configurations, and object functional planes, providing a detailed geometric and manipulation-oriented understanding of the objects for downstream tasks.

Task decomposition with spatially enhanced context

This module leveraged the grounded spatial information to produce a spatially aware manipulation plan for a given task. To achieve this, it first retrieved a textual description for each object-centric prior, such as "a horizontal plane located in the bowl opening area and parallel to the bowl rim." These text descriptions, along with the transferred information (such as pose and size parameters), were then structured to enhance the context of the task. This spatially augmented context, along with the original image and user instruction, was fed back into the VLM, which then decomposed the user's high-level task into a sequence of physically plausible and spatially precise substeps, each expressed by a set of object-centric spatial constraints. Crucially, this allowed the VLM to adaptively select and parameterize the relevant spatial priors needed to define the robot's motion. For instance, it could then generate a subgoal like "align <plane #1> of the bowl with <plane #2> of the plate," directly grounding abstract actions in concrete spatial constraints.

Trajectory optimization with spatial constraints

This module was responsible for translating the VLM's spatially aware plan into the robot trajectory. It achieved this by leveraging both the spatial constraints associated with each substep and the detailed visual grounding information from the grounding module. Similar to a prior work (*27*), we formulated the robot's workspace as a series of voxelized maps. The spatial constraints planned by the VLM were encoded into these maps and then composed into a unified cost field to guide a trajectory optimizer. More specifically, we defined four key maps to construct this field: an affordance map to direct the robot toward the target object, an avoidance map for collision-free motion, a rotation map to control the end effector's orientation, and a gripper map to command the gripper's state. By optimizing a trajectory through this combined cost field, the robot could execute the planned motions to complete the manipulation task.

Zero-shot performance of RAM in real-world robot tasks

To evaluate the zero-shot capabilities of the RAM framework, we designed a suite of 14 challenging spatially aware manipulation tasks that demanded a nuanced understanding of spatial relationships. These experiments were conducted on a real-world robotic platform involving 31 distinct object instances from 11 categories. The tasks were structured to probe three critical dimensions of spatially aware manipulation. In the following sections, we demonstrate how RAM enabled the precise execution of spatial-relevant language commands, facilitated image-guided spatial-aware manipulation, and supported adaptive task planning with spatial reasoning of the scene. These results collectively showcase RAM's ability to bridge the gap between task planning and physical execution in diverse real-world spatially aware manipulation scenarios.

RAM enabled precise execution of spatial instructions

Although recent advancements have enabled robots to follow high-level language commands, a critical gap persists in their ability to comprehend and act upon precise spatial-relevant instructions. Existing methods often falter when instructions involve nuanced positional and orientational relationships, leading to ambiguous or incorrect physical executions. To evaluate RAM, we designed a suite of experiments to probe its capacity for translating complex, spatially rich language instructions into precise physical actions. To this end, we curated 12 robot manipulation tasks (four positional tasks and eight combined position-and-orientation tasks), each defined by a unique spatial instruction. These tasks were grouped into three types on the basis of the complexity. The first category comprised four single-object, single-step tasks, which required precise control over a single object based on short-horizon commands, such as "rotate the spoon on the table 90° counterclockwise." The second category consisted of four multiobject, single-step tasks, which tested the understanding of spatial relationships between multiple objects, such as "position the camera to face the teddy bear directly." The third category included four multiobject, multistep tasks, which demanded long-horizon planning and a consistent understanding of evolving spatial relationships, such as "align all the scattered spoons to the right of the blue spoon, matching their orientations." Figure 2 (A to C) presents an example for each of these three types of tasks. The complete task list can be found in the Supplementary Materials.

To ensure a robust evaluation, we tested each of the 12 tasks 10 times with varied initial object states. As summarized in Fig. 2D, across the 120 total trials, RAM achieved an average success rate of 89.17%. For the most challenging multiobject, multistep tasks, which required a combination of long-horizon planning and precise spatial control, RAM maintained an average success rate of 80.00%. These results validated that RAM could consistently follow complex spatial-relevant instructions to control the position and orientation of single or multiple objects in both short- and long-horizon contexts, demonstrating a robust capability for spatial instruction following. In the meantime, we evaluated the RAM framework with three additional VLMs, including two leading proprietary models, GPT-4o (*42*) and Claude Sonnet 4 (*43*), and one open-source model, Qwen-VL (*44*). As can be observed from Fig. 2D, RAM with different VLMs exhibited similar capabilities of spatial instruction following, especially for single-object, single-step tasks. Furthermore, we observed that RAM's performance in spatial instruction following exhibited a correlation with the inherent planning capabilities of its underlying VLM. This was evidenced by the performance trend distinguishing the proprietary models from the open-source counterpart. Although RAM's effectiveness was anchored on the capability of the adopted VLM, this result also demonstrates that RAM was effective at leveraging and channeling the general reasoning power of a given VLM for the domain of spatially aware manipulation.

RAM enabled image-guided spatial-aware manipulation

Natural language instruction was a highly flexible interface for commanding robots. Yet, its expressive bandwidth proved insufficient when specifying tasks with intricate, multiobject spatial arrangements. In contrast, a single image could convey complex goal configurations more directly and unambiguously. However, translating the rich information in a goal image into robotic action was a formidable

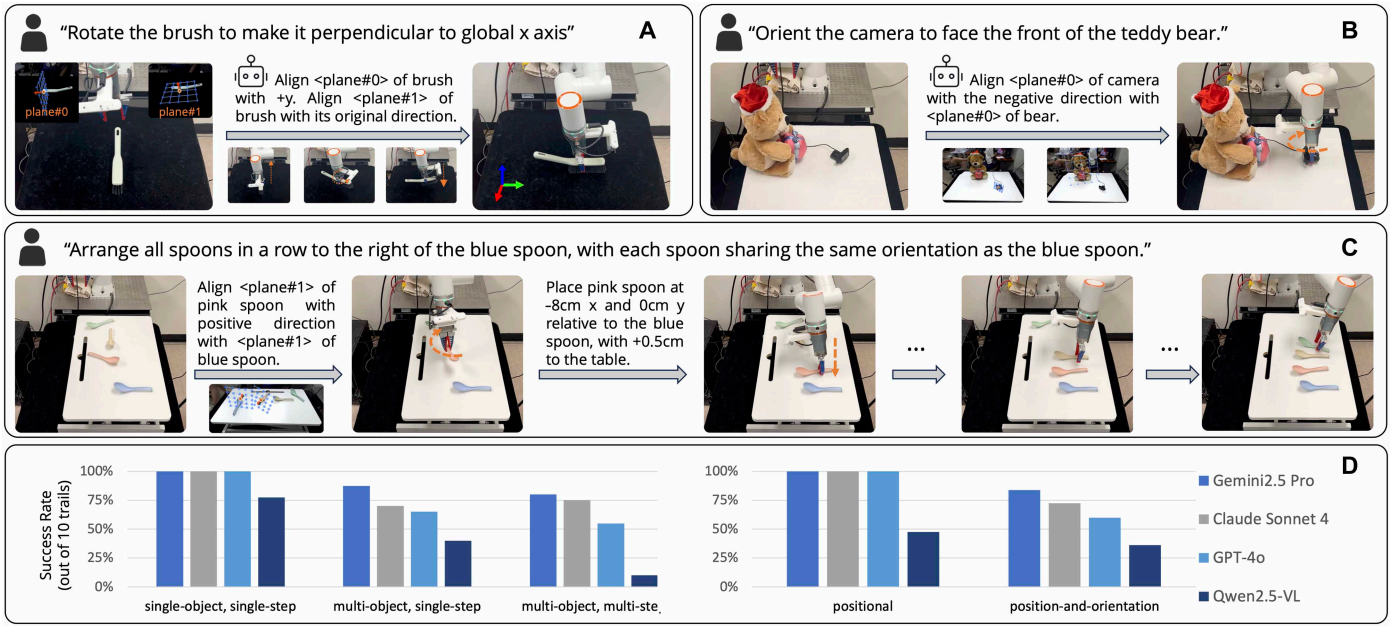

Fig. 2. Illustration and results of RAM in different spatial instruction-following tasks. These tasks include single-object, single-step tasks (**A**); multiobject, single-step tasks (**B**); and multiobject, multistep tasks (**C**). (**D**) RAM results with different VLMs.

challenge. This demanded not only the ability to perceive and abstract latent spatial relationships from a 2D view but also the critical capacity to ground these abstract concepts as precise, actionable constraints in the robot's 3D physical workspace.

To evaluate RAM's performance in this setting, we took the task of tableware rearrangement as a representative example. Conventional methods (*45*, *46*) for this task were often constrained by specific goal-image perspectives, such as a fixed top-down view, to reduce this challenging task to a 2D arrangement problem on a single table plane. We evaluated RAM with a single goal image captured from an arbitrary viewpoint. This unconstrained setup increased the difficulty, demanding a higher level of spatial understanding and reasoning. As the subtask decomposition and execution results in Fig. 3 demonstrate, RAM could bridge the observed and goal images, allowing the system to correctly infer the intended 3D spatial relationships. RAM then leveraged this spatial understanding to decompose the long-horizon rearrangement task into a coherent series of subtasks, each endowed with explicit spatial constraints derived from the goal image for guided rearrangement.

Specifically, we collected 10 different goal images with varying objects and arrangements. The evaluation was structured into two levels of difficulty. In the regular setting, all objects were initialized with random positions and orientations on a single horizontal plane. The hard setting introduced a greater challenge by placing objects on different vertical planes, requiring a higher level of 3D spatial understanding. We tested the regular setting on five goal images and the hard setting on the other five goal images. For each goal image, the rearrangement experiment was repeated five times with randomized initial object positions and orientations. RAM achieved a 92.00% average success rate in the regular setting and maintained a 72.00% task success rate in the hard setting. These results validated that RAM could effectively translate complex visual goals into precise, multistep manipulation plans, showcasing a sophisticated level of image-guided spatial awareness for spatially aware robot manipulation.

RAM enables adaptive task planning with spatial reasoning

Many manipulation tasks, particularly those requiring precise tool use or interaction with cluttered environments, cannot be resolved with semantic understanding alone. They are constrained by the spatial realities of the scene. This work suggested RAM to elevate the spatial intelligence of general VLMs, enabling the robot to reason about these physical constraints and make plans adaptively for robot manipulation.

To evaluate RAM's performance, we designed a desk-cleaning task where semantic guidance alone is insufficient, creating ambiguity that must be resolved by spatial reasoning. Figure 4A shows the tools used in this task, including a table with a continuously adjustable height (15 to 60 cm), four bins of varying sizes, a dustpan, and a brush. By varying the table height and the available tools, we simulated diverse desk-cleaning scenarios. In these situations, a general VLM relying solely on task descriptions and scene semantics might generate plans that are plausible in semantics but physically infeasible, for example, selecting an incorrectly sized bin or attempting to sweep debris into a bin whose opening was higher than the tabletop. Beyond this need for spatial reasoning in task planning, the task itself demanded high precision in manipulation. Specifically, in the 1.2 m–by–1.2 m workspace, the dustpan's edge had to be flush with the table's edge; the brush had to be precisely oriented for sweeping; and the dustpan had to be accurately aligned over the bin's opening to pour without spilling. Under this complex task setting, we conducted 20 replicate experiments, with different tools and varying table heights in each. As shown in Fig. 4 (B to D), we observed that RAM could autonomously identify implicit spatial constraints and adaptively generate a physically feasible plan, achieving an average success rate of 65%.

Analysis of RAM results with benchmark performance

We conducted additional experiments on public benchmark datasets to evaluate RAM's performance and compare it against existing

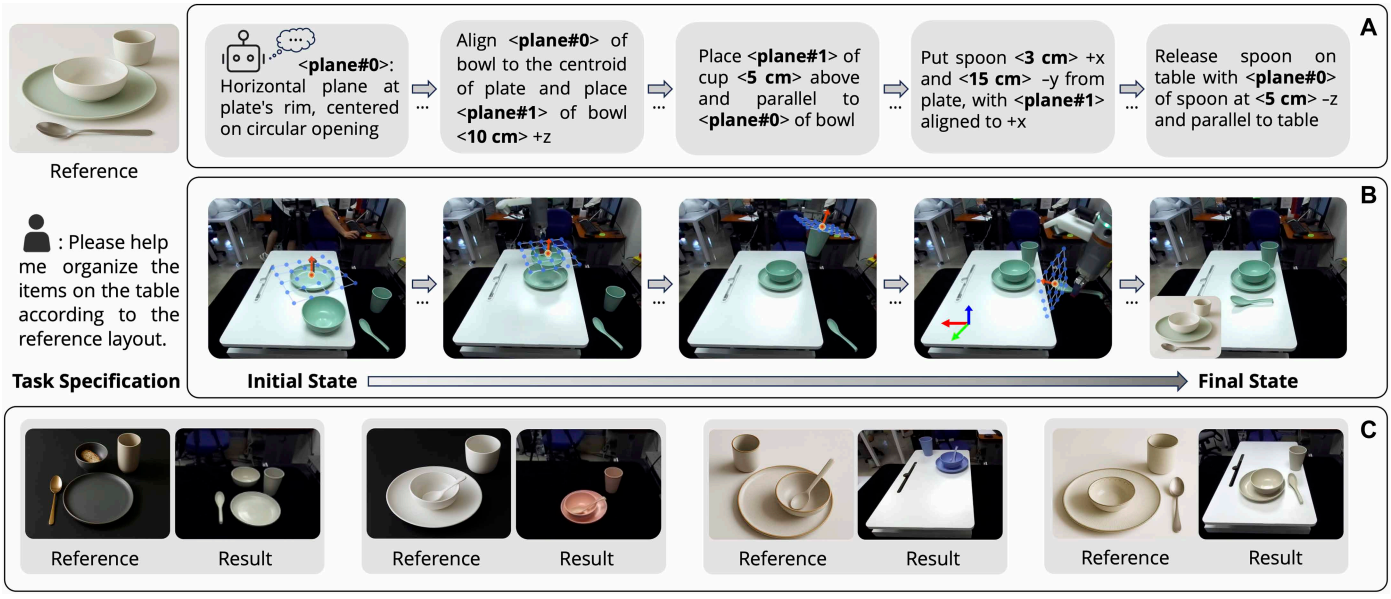

Fig. 3. Illustration and results of RAM in image-guided spatial-aware manipulation tasks. (**A**) RAM can decompose a long-horizon task into a series of substeps with coherent spatial constraints with the given goal image. (**B**) Visualization of the robot execution after the task planning. (**C**) RAM is robust to different image specifications for tableware rearrangement.

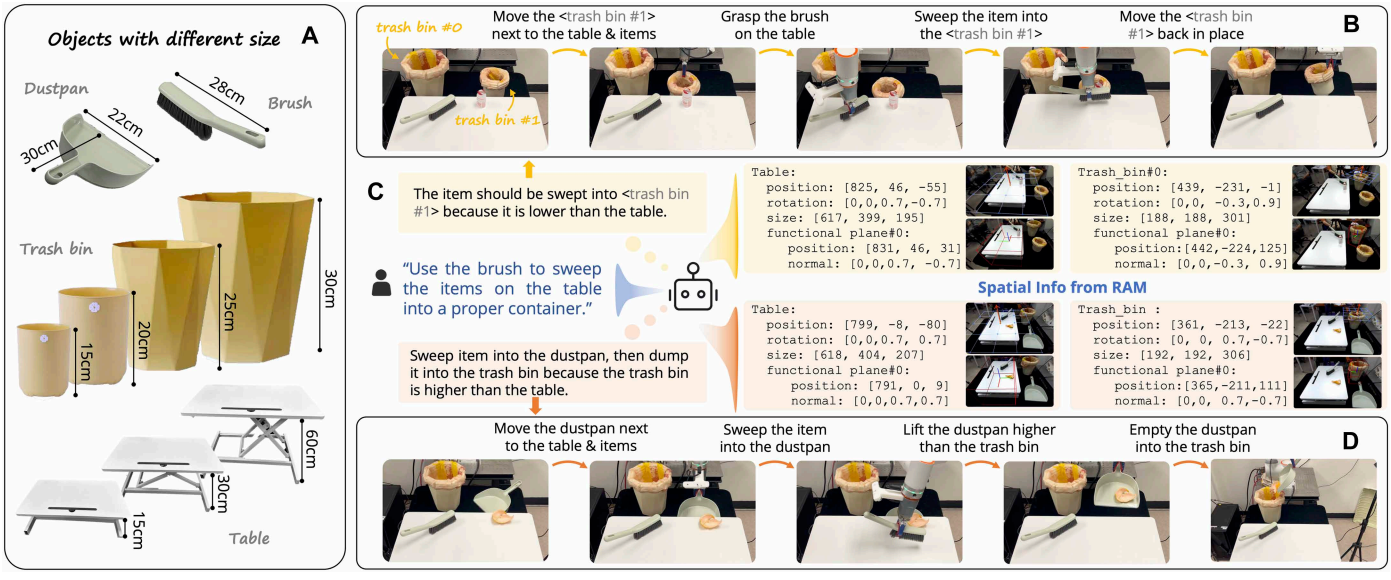

Fig. 4. Illustration and results of RAM in adaptive task-planning tasks. (**A**) Visualization of tools used in the desk-cleaning task and their physical sizes. (**B**) Execution of a direct sweeping strategy. (**C**) Robot's decision-making process. (**D**) Execution of an indirect sweeping strategy.

approaches. These experiments aimed to further systematically assess RAM's generalization to unseen object categories, robustness to environments, and spatial understanding and reasoning capabilities in comparison with existing VLMs.

Evaluation of generalization capability to unseen categories

The category-level visual grounding module of RAM was first trained on a synthetic dataset comprising 20 base categories. The objective of this phase was twofold: to enhance the pretrained 2D image representations with 3D point cloud coordinates and to adapt the 2D vision foundation model for our target 3D grounding scenarios. To quantitatively evaluate the generalization performance of this trained module on various unseen object categories, we conducted experiments on the CO3D dataset (*47*). We used precise ground-truth pose labels from (*48*) to evaluate RAM's category-level visual grounding module on 10 distinct object categories, varying from common household items (such as laptop and backpack) to larger objects (such as motorcycle and chair). Each category contained around 1020 images captured from diverse viewpoints, across 10 unique object instances with varying textures, shapes, and sizes. Following the methodology of (*49*), we used the average accuracy in

terms of geodetic rotation error of 15° to evaluate the 3D grounding results. Figure 5A presents the experimental results. In comparison with two baseline models, zero-shot pose (ZSP) (*48*) and FoundPose (*50*), that are based on pretrained DINO-v2 (*41*), RAM's visual grounding model exhibited a consistent performance improvement across all tested categories. This superiority was primarily attributed to our feature lifting strategy, which integrated 3D position embeddings to enhance viewpoint discrimination, thereby reducing orientation ambiguity compared with the 2D feature matching used in baselines. Furthermore, our dense coordinate prediction established robust dense correspondences, offering greater resilience to shape variations than sparse keypoint matching. The result indicated that despite being fine-tuned exclusively on synthetic data from base categories, RAM could effectively scale its generalization capabilities to a wide array of real-world objects from unseen categories.

Evaluation of robustness to shape variations and occlusions

A critical consideration for the practical deployment of RAM is the robustness of its object-centric grounding to shape variations between the shape template and a target object instance, as well as to the choice of the template itself. To this end, we quantitatively evaluated RAM's robustness to shape variation on the CO3D dataset using the same metric based on geodetic rotation error. Given that precise mesh models for each instance in CO3D were not available, it is infeasible to directly measure the shape difference between an object instance and its corresponding shape template. Therefore, we devised a proxy evaluation: We used multiple, distinct 3D mesh models as shape templates for the same category and then assessed robustness by comparing the performance differences when using these different templates. As shown in Fig. 5B, five distinct shape templates were selected for a category on the basis of their mutual shape difference, measured by point cloud Chamfer distance. We observed that RAM was robust to both the choice of shape template and the inherent shape variations among objects, evidenced by the similar average accuracies and accuracy distributions achieved across different object instances regardless of the template used.

Concurrently, we evaluated RAM's performance under varying degrees of environmental occlusion. Following a methodology similar to that in (*51*), we simulated partial occlusion by randomly masking a specified percentage of image pixels and evaluated the average accuracy at different occlusion ratios. The results show that, although RAM maintained robustness under moderate occlusion levels (below 50%), its accuracy degraded as the occlusion ratio increased further. This degradation under heavy occlusion represents a known challenge for purely vision-based systems, highlighting a promising direction of fusing tactile sensing with active perception strategies (*52*) to mitigate such failures. Furthermore, to ground this analysis in practical use cases, we conducted tests on our robotic platform in two common manipulation scenarios: object-object and object-robot occlusion. As shown by the representative qualitative results in Fig. 5C, RAM demonstrated robustness in both of these practical occlusion scenarios.

Evaluation of spatial understanding capability with VQA

To evaluate RAM's spatial understanding capabilities in robot manipulation tasks, we constructed a visual question answering (VQA) dataset using our robotic platform, designed to assess a broader spectrum of spatial reasoning skills. Our VQA dataset comprised 100 manually annotated samples, with each sample consisting of an RGB image paired with a corresponding multiple-choice question. Effective robot manipulation necessitates advanced spatial reasoning, including assessing interobject positional relationships, discerning object orientations and rotations, evaluating the outcomes of operational behaviors, performing task planning, and conducting quantitative scene measurements. Consequently, our VQA questions were categorized into five distinct types: positional relationship, judgment, rotation, planning, and measurement, with 20 questions allocated to each category. All questions and their corresponding ground-truth answers were designed and human annotated. Representative examples from our VQA dataset are presented in Fig. 6.

RAM's performance was benchmarked against its baseline, VLM Gemini-2.5-Pro (*39*), and two VLMs that had been fine-tuned for improving the spatial awareness, SpaceOM and SpaceThinker (*31*). As

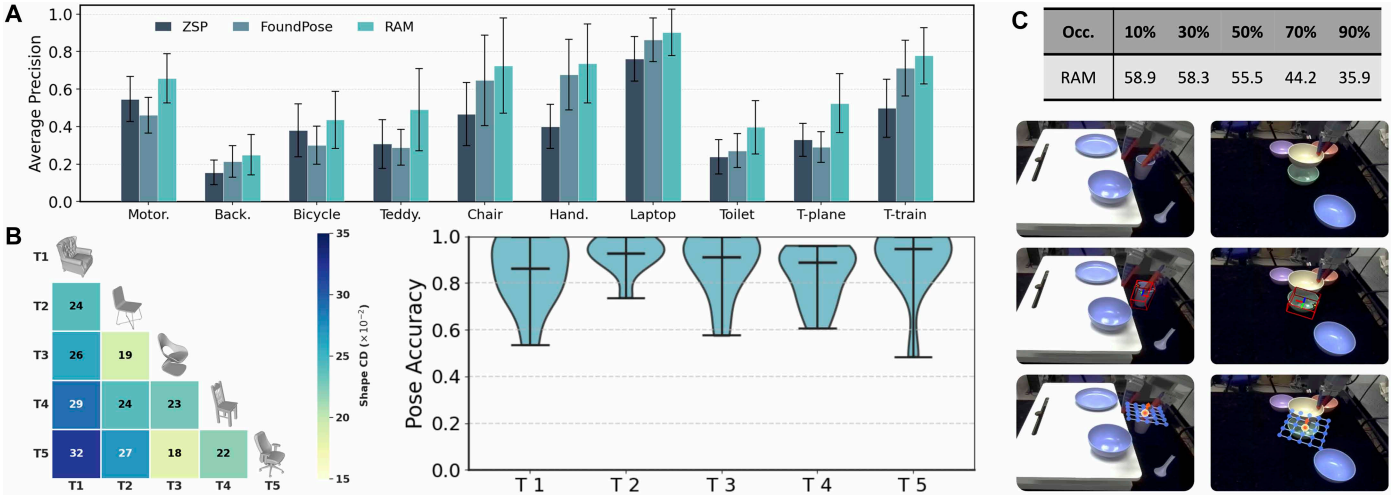

Fig. 5. Benchmark performance of RAM. (**A**) Generalization performance of RAM on unseen object categories with average accuracies in terms of geodetic rotation error. Bar plots with error bars show the mean values ± SD, with $n = 10$ independent experiments. (**B**) Robustness analysis of RAM to object shape variations. Left: Five distinct chair templates (T1 to T5) and their shape differences, quantified by Chamfer distance. Right: A violin plot showing the corresponding pose accuracy when using each template. (**C**) Robustness analysis of RAM to environmental occlusions (Occ.) on both the benchmark dataset and real-world robot manipulation scenarios.

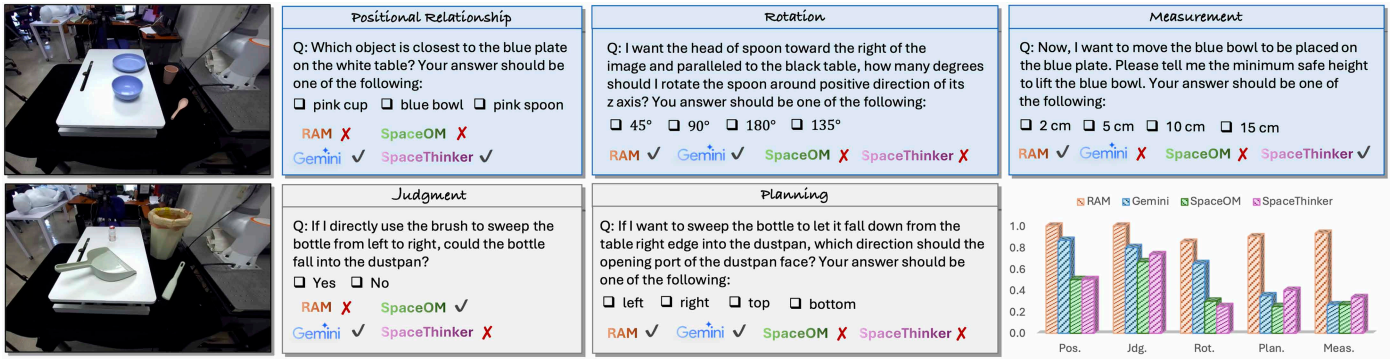

Fig. 6. Spatial comprehension evaluation on a custom VQA dataset. Representative examples from the VQA dataset, which was created to assess a broad spectrum of spatial reasoning skills. The questions were divided into five categories: positional relationship (Pos.), judgment (Jdg.), rotation (Rot.), planning (Plan.), and measurement (Meas.). The bar chart compares the performance of RAM against other state-of-the-art models—Gemini-2.5-Pro, SpaceOM, and SpaceThinker—demonstrating RAM's superior accuracy across all five categories.

depicted in the accompanying bar chart (Fig. 6), the results unequivocally demonstrated RAM's consistent superior performance across all five categories compared with the other VLMs. Benefiting from its larger model capacity, Gemini-2.5-Pro demonstrated superior performance over the 3B models [SpaceOM and SpaceThinker, which are based on Qwen2.5VL-3B (*44*)] in tasks involving spatial positional relationships and object orientation. However, in more intricate spatial perception tasks, including judgment, measurement, and planning, Gemini-2.5-Pro exhibited comparable performance to SpaceOM and SpaceThinker, whereas RAM exhibited consistent performance improvement when compared with the three competing VLMs.

DISCUSSION

Our results demonstrate that the RAM framework provides a robust and effective solution to one of the most pressing challenges in modern robotics: bridging the semantic-geometric gap for foundation models. We have shown that, by explicitly grounding the abstract reasoning of a VLM with a retrievable, object-centric knowledge base, a robot can achieve a level of spatial awareness and precision in manipulation that is unattainable with ungrounded, end-to-end approaches. The success of RAM in complex, contact-rich tasks validates our core hypothesis: The key to unlocking physically intelligent behavior is not a choice between classical structured methods and modern learned models but their synergistic integration. Our framework offers a concrete mechanism to resolve the tension between the powerful, zero-shot semantic capabilities of VLMs and the nonnegotiable physical constraints of the real world, moving beyond spatially naive planning to enable meaningful physical interaction.

The implications of this work extend beyond the immediate tasks demonstrated. The concept of an object-centric engine, particularly its use of "functional planes," can be seen as a computational instantiation of the theory of affordances. It suggests that a robot's spatial understanding of an object should not be limited to its identity or geometry but should also include its potential for interaction. To translate this high-level understanding into actionable robot skills, our current implementation of RAM focuses on spatial-aware placement and rearrangement tasks. In these scenarios, the core difficulty involves reasoning about and achieving precise spatial configurations among rigid objects. However, the proposed framework is inherently extensible beyond these boundaries. We have demonstrated that our object-centric formulation is robust to handle nonrigid geometry. By using a multitemplate strategy, RAM can precisely ground articulation mechanisms, such as revolute and prismatic axes, for articulated objects and manage local shape variations in deformable tasks like clothes folding, despite being trained primarily on rigid data. Furthermore, as detailed in the Supplementary Materials, the framework could be agnostic to the end effector. The knowledge base can readily accommodate multifingered dexterous hands by extending grasp representations to include local contact maps and semantic descriptors. This allows the VLM to select valid dexterous grasps on the basis of functional suitability without retraining the visual grounding model. Moreover, this structured knowledge base also serves as an anchor for multimodal integration. Our experiments show that, by fusing tactile feedback with geometric priors, the VLM can reason about unobservable physical properties, such as friction or weight distribution, and correct improper grasps via an adaptive replanning scheme. Furthermore, this structured spatial representation could have a profound effect on downstream applications. For instance, it can improve sim-to-real transfer by providing simulators with richer object-level constraints and enable more sophisticated task and motion planning by reducing the search space for valid physical interactions.

It is equally important to acknowledge the current limitations of the RAM framework, which, in turn, define avenues for future research. First, our failure analysis (refer to the "Failure analysis" section of the Supplementary Materials) reveals that the system's reliability is constrained by specific bottlenecks across the pipeline, including upstream segmentation errors from the 2D perception module, occasional suboptimal logic by the VLM in long-horizon planning, and downstream kinematic constraints that are not yet fully integrated into the high-level reasoning loop. Second, a practical hurdle for scalability is the manual creation and annotation of shape templates in the object-centric engine. Although this "one-time investment" pays substantial dividends in generalization, scaling this approach to thousands of object categories requires a more automated process. A promising future direction involves developing methods to learn these object-centric knowledge templates automatically from diverse, multimodal data sources, such as videos of human interaction, product manuals, or large-scale 3D object datasets. Furthermore, our current work focuses on rigid objects. Regarding the handling of articulated and deformable objects, although we have demonstrated that RAM

can encompass these categories through a multitemplate strategy, we frankly acknowledge that using a finite set of discrete templates to approximate the continuous state space of nonrigid objects is a strategic trade-off. Despite offering a data-efficient and practical solution for many manipulation tasks, this approach may not be the optimal representation for scenarios involving highly unstructured deformations or infinite degrees of freedom. Thus, we position our method as a robust solution for handling structured variations in nonrigid objects, leaving the development of a universal solver for all dynamic complexities as a challenging but essential next step.

In conclusion, the RAM framework represents a step toward building robots that are both linguistically competent and physically intelligent. By demonstrating how to effectively ground the powerful reasoning of foundation models in the geometric and functional reality of the physical world, our work provides a blueprint for the robotic system with a hybrid architecture, whereby structured knowledge and large-scale learning empower one another, paving the way for robots that can more safely, reliably, and capably collaborate with humans in the complex, unstructured environments of our daily lives.

MATERIALS AND METHODS

The category-level object-centric engine

The category-level object-centric engine served as an extensible knowledge base designed to bridge high-level semantic reasoning with physical execution. It stored a standard template for every object category, which was capable of accommodating object-centric priors tailored to different robot embodiments (such as parallel-jaw grippers or dexterous hands). These priors were then transferred to different object instances via a unified visual grounding process. The template included extensive annotations that offered a thorough understanding of the object category. Specifically, it contained an object category label $\mathcal{E}$, which was a short phrase describing the object category, alongside a 3D mesh $\mathcal{X}$ that represented the typical shape of the category with complete 3D geometry. Furthermore, the template incorporated a canonical pose $o \in \mathrm{SO}(3)$, indicating the standard orientation of the object and serving as a reference for aligning multiple instances from the same category. It also included a set of valid grasp configurations $\mathcal{G}$ defined on $\mathcal{X}$. To ensure that the grasp representation can fit different end effectors, we characterized each grasp $\mathbf{G}_i \in \mathcal{G}$ by a tuple consisting of a normalized approaching direction and an object-centric contact relationship with the object. The exact form of this relationship varied according to the end effector. For instance, in the case of a parallel-jaw gripper, $\mathbf{G}_i = \langle \mathbf{u}_i, \mathbf{t}_i \rangle$, such that $\mathbf{u}_i, \mathbf{t}_i \in \mathbb{R}^3$, $\| \mathbf{u}_i \| = 1, \forall i = 1, 2, \ldots, |\mathcal{G}|$. These configurations were demonstrated to enable effective and stable grasps of $\mathcal{X}$ in a physical simulator. In addition, the template provided semantic grasp descriptions, offering a standardized textual description for each valid grasp configuration in the format of `<#semantics, #orientation, #functional description>`. These descriptions identified the semantic region, approach direction, and functional purpose, thereby decoupling high-level grasp planning from the policy used to sample the template grasp and facilitating our method to accommodate grasps sampled from diverse policies in a unified manner. Last, the template defined functional planes $\mathcal{F}$, which were essential for object manipulation on $\mathcal{X}$. Each plane $\mathbf{F}_j \in \mathcal{F}$ was represented by a tuple of normal and center point, defined as $\mathbf{F}_j = \langle \mathbf{n}_j, \mathbf{c}_j \rangle$, such that $\mathbf{n}_j, \mathbf{c}_j \in \mathbb{R}^3$, $\|\mathbf{n}_j\| = 1, \forall j = 1, 2, \ldots, |\mathcal{F}|$.

By defining the template for each object category, we could transfer the knowledge to various observed instances in the same category. This approach enhanced the capability of robotic systems, enabling them to perform complex manipulation tasks more efficiently.

Category-level visual grounding

We developed a category-level visual grounding model to enhance spatial understanding of each observed object in the scene. It was achieved by estimating a dense correspondence between the object and its categorical template in the category-level object-centric engine. Considering the inherent intraclass variations of objects in terms of textures and shapes, we built our visual grounding model on top of the robust image representations from DINO-v2 (*41*) to transfer the category-level knowledge to the observed object. A foundation feature lifting module that we presented in our previous work (*53*) was leveraged to integrate the pretrained image features with positional embeddings derived from point cloud coordinates for robust visual grounding.

For a given observed object, we first retrieved its categorical template from our object-centric engine on the basis of the object category label. Subsequently, following the method of (*53*), we uniformly sampled a set of viewpoints of the 3D mesh $\mathcal{X}$ in the template from a hemisphere centered on its geometric center and rendered a synthetic RGB-D image for each view using Blender. Then, we predicted the object-template correspondence in two stages. First, an image-level matching was performed to identify the rendered viewpoint that most closely aligned with the observed view of the object. The procedure was written as

$$\nu^* = \arg \max_{\nu = 1, \ldots N} \left(\Phi_{\text{view}}\left(I_q, D_q\right) \cdot \Phi_{\text{view}}\left(I_{\text{r}}^{\nu}, D_{\text{r}}^{\nu}\right) \right) \tag{1}$$

where $\left(I_q, D_q\right)$ denoted the RGB and depth patch for the observed object q, and $\{\left(I_{\text{r}}^{\nu}, D_{\text{r}}^{\nu}\right) | \nu = 1, \ldots, N\}$ denoted N pairs of rendered RGB and depth images of the 3D mesh $\mathcal{X}$. Φ_{view} denoted a viewpoint encoder that extracted viewpoint features from a pair of RGB and depth inputs. The viewpoint encoder was trained with contrastive learning to generate similar features for comparable viewpoints in an object category while producing distinct features for substantially different viewpoints. Thus, the rendered viewpoint of $\mathcal{X}$ closest to that of the observed object q, denoted as ν^*, was determined by finding the ν whose viewpoint feature had the highest similarity with that of the observed object q.

Second, on the basis of the image-level matching result, we further performed a pixel-level matching to estimate a dense correspondence between the observed object and the categorical template. It could be formulated as

$$\mathcal{M} = \text{Softmax}\left(\Phi_{\text{pixel}}\left(I_q, D_q\right) \times \Phi_{\text{pixel}}\left(I_{\text{r}}^{\nu^*}, D_{\text{r}}^{\nu^*}\right)^{\text{T}} \right) \tag{2}$$

where ν^* denoted the best-match view identified through image-level matching. Φ_{pixel} was a pixel-level feature encoder that outputs d-dimensional features for each pixel in an RGB and depth image pair. Let n and m denote the number of pixels from the observed object and the 3D mesh at viewpoint ν^*, respectively, we had $\Phi_{\text{pixel}}\left(\mathrm{I}_q, \mathrm{D}_q\right) \in \mathbb{R}^{n \times d}$ and $\Phi_{\text{pixel}}\left(I_{\text{r}}^{\nu^*}, D_{\text{r}}^{\nu^*}\right) \in \mathbb{R}^{m \times d}$. The result, $\mathcal{M} \in \mathbb{R}^{n \times m}$, was a normalized matching matrix that represented the correspondence between each pair of pixels from the observed object and the rendered 3D mesh at view ν^*, respectively, with higher values indicating stronger correspondence.

On the basis of the estimated matching matrix $\mathcal{M}$, we could ground different types of prior information from the categorical template to the observed object. For object pose and size, we first followed (*53*) to define a canonical object coordinate map $\mathcal{P}_{\mathrm{r}}^{\mathrm{ccm}}$, consisting of m points each corresponding to a pixel from the rendered 3D mesh image at view v^*. Then, we matched the canonical coordinates for the observed object point cloud by $\mathcal{P}_{\mathrm{match}} = \mathcal{M} \times \mathcal{P}_{\mathrm{r}}^{\mathrm{ccm}}$, consisting of n points. After that, the object pose and size parameters could be jointly estimated by minimizing the alignment error $\|\mathcal{P}_q - (s\mathbf{R} \times \mathcal{P}_{\mathrm{match}} + \mathbf{t})\|_2^2$, where $\mathcal{P}_q \in \mathbb{R}^{n\times 3}$ denoted the observed object point cloud in the camera view and (s, $\mathbf{R}$, $\mathbf{t}$) denoted the object size, rotation, and translation parameters, respectively. We adopted Umeyama (*54*) and RANSAC (*55*) for a robust estimation.

The grounding of the grasping pose and functional plane followed a similar procedure. For the grasping pose, we used the grasping pose parameters to transform the canonical object coordinate map into the grasping frame. For the functional plane, we used the plane center and normal to transform the canonical object coordinate map into the plane frame. Then, we could estimate the parameters of the grasping and functional plane by minimizing similar alignment errors.

Visual grounding model training with synthetic data

We trained our category-level visual grounding model using photorealistic synthetic data rendered with Blender. To enhance diversity in the training dataset, we collected 20 object categories from ShapeNet (*56*), each featuring six unique object instances. For each category, we randomly selected one instance and a rendered view, resulting in 10,000 RGB-D images along with their corresponding object masks and ground-truth object coordinate maps. In total, we rendered 200,000 images for training.

For each object category, we randomly chose one object instance to serve as the shape template, whereas the other five instances were used as query objects during training. To train the image-level object viewpoint estimation, we adopted the pose-aware contrastive loss in (*57*). It harnessed both the positive and the negative viewpoint pairs to train Φ_{view}. The training loss was written as

$$\mathcal{L}_{\mathrm{view}} = -\log\left(\frac{\exp\left(f_{\mathrm{view}}^{q} \cdot f_{\mathrm{view}}^{r^+}/\tau\right)}{\sum_{k=1}^{K} d\left(\mathbf{R}_q, \mathbf{R}_r^k\right)\exp\left(f_{\mathrm{view}}^{q} \cdot f_{\mathrm{view}}^{r^k}/\tau\right)}\right) \tag{3}$$

where K represented the number of samples used for comparison with the query; q denoted the query viewpoint, r^k denoted each compared viewpoint, and r^+ denoted the closest reference viewpoint, with their extracted viewpoint features denoted f_{view}^{q}, $f_{\mathrm{view}}^{r^k}$, and $f_{\mathrm{view}}^{r^+}$, respectively; τ was a temperature parameter; $d(\cdot,\cdot)$ was the normalized viewpoint difference between query and reference, which was defined as $d\left(\mathbf{R}_q, \mathbf{R}_r^k\right) = \arccos\left(\frac{\mathrm{tr}\left((\mathbf{R}_q)^{\mathrm{T}}\mathbf{R}_r^k\right)-1}{\pi}\right)$. We followed (*58*) to handle the symmetric objects.

To train the pixel-level coordinate map estimation, we first obtained the point cloud of the query object $\mathcal{P}_q'$ and that of the reference object $\mathcal{P}_r'$. Next, we calculated the matched point cloud using the matching matrix $\mathcal{M}$ obtained with Eq. 2, which was written as

$$\mathcal{P}'_{\mathrm{match}} = \mathcal{M}\mathcal{P}'_q \tag{4}$$

where $\mathcal{P}'_{\mathrm{match}}$ was the matched point cloud. Then, we regarded $\mathcal{P}_r'$ as the ground-truth, compared $\mathcal{P}'_{\mathrm{match}}$ with it, and obtained the training loss $\mathcal{L}_{\mathrm{coor}}$ for pixel-level coordinate map estimation. For symmetrical objects, the training loss was designed on the basis of the Chamfer distance, which was written as

$$\mathcal{L}_{\mathrm{coor}} = CD\left(\mathcal{P}'_{\mathrm{match}}, \mathcal{P}'_r\right) \tag{5}$$

where $CD(\cdot)$ denoted the function of Chamfer distance. The equation meant that, for symmetrical objects, the matching parameters were satisfactory as long as the predicted result $\mathcal{P}'_{\mathrm{match}}$ matched well with $\mathcal{P}'_r$. For asymmetrical objects, we used a training loss based on the smoothed L1 loss, which was written as

$$\mathcal{L}_{\mathrm{coor}} = \sum\nolimits_{(x,y,z)\in\mathcal{P}'_{\mathrm{match}}} \psi\left(x, x_r\right) + \psi\left(y, y_r\right) + \psi\left(z, z_r\right) \tag{6}$$

where $(x, y, z) \in \mathcal{P}'_{\mathrm{match}}$ indicate a point in the $\mathcal{P}'_{\mathrm{match}}$, whereas (x_r, y_r, z_r) indicate the corresponding point in $\mathcal{P}_r'$ that was closest to (x, y, z) measured by Euclidean distance. $\psi(\cdot)$ denotes the smoothed L1 loss.

By fusing global viewpoint loss and local coordinate map loss, our core training objective was to fuse the rich semantic representations from 2D foundation models with explicit 3D spatial information derived from object coordinates. Although our implementation was trained exclusively on rigid objects from ShapeNet, we empirically observed that the adopted foundation feature lifting strategy fostered a robust understanding of intrinsic semantic-geometric correspondences that went beyond rigid alignment. Consequently, the model maintained high grounding accuracy even in the presence of local deformations or kinematic changes. This inherent robustness served as a critical enabler for extending our framework to handle complex articulated and deformable objects without requiring additional training, as further validated in our supplementary experiments. Formally, the overall training loss for the visual grounding model on viewpoint knowledge transfer was $\mathcal{L} = \mathcal{L}_{\mathrm{view}} + \mathcal{L}_{\mathrm{coor}}$. The training for knowledge transfer of valid grasp configurations and functional planes followed a similar approach, with the preparation of training data tailored to each, allowing us to train the respective encoders.

Retrieval-augmented task planning

With the geometric knowledge for each object category provided in our category-level object-centric engine and the category-level visual grounding scheme to transfer the knowledge to an observed object, we further introduced the knowledge to a VLM, the Gemini-2.5-Pro, to more effectively plan the current task. At the start of the task planning, we captured the RGB-D observation (I, D) of the entire scene and provided a task description $\mathcal{T}$. The VLM-based task planning consisted of four steps.

First, we gathered all of the object category labels from the category-level object-centric engine. The list of object category labels $\mathcal{E}$, I, and $\mathcal{T}$ was supplied to the VLM to identify object categories relevant to the task. For each relevant object category, the VLM enriched the object description for each relevant object category, on the basis of its original object category label $\mathcal{E}$ and the visual input I, resulting in an enriched object description $\widehat{\mathcal{E}}$.

Second, both the whole-scene image I and all of the enriched descriptions $\widehat{\mathcal{E}}$ were input into Grounded SAM (*59*) to detect and segment each relevant object q in I, resulting in a precise 2D bounding box $\mathcal{B}_q$ and mask $\mathcal{A}_q$. Subsequently, the VLM reengaged to

establish a definitive correspondence between $\widehat{\mathcal{E}}$ and q. Using the RGB-D observation (I, D) of the entire scene, we cropped an RGB-D patch (I_q, D_q) for each object q using the bounding box $\mathcal{B}_q$ and then passed it as input to our visual grounding module, which extracted the geometric information from the categorical templates and transferred it to each q. The geometric information encompassed the object's estimated 6D pose ($\mathbf{R}_q$, $\mathbf{t}_q$) and size (w_q, l_q, h_q), one or more valid grasp configuration(s) $\mathcal{G}_q$, and one or more functional plane(s) $\mathcal{F}_q$.

Third, we performed task decomposition. We input $\mathcal{T}$, I, and $\{\widehat{\mathcal{E}}, \mathcal{B}_q, (w_q, l_q, h_q)\}$ to the VLM to break down the overall task into multiple subtasks. We generated a detailed text description $\mathcal{T}_s$ for the sth subtask, which offered clear instructions and context, ensuring that the subtasks were well defined and easily understood for the subsequent generation of manipulation actions.

Last, for the sth subtask, the VLM derived a standardized sequence of actions $\boldsymbol{a}_t^s, t = 1, 2, \ldots, T^s$, each in four primitive types: grasp, lift, move, and release. Each manipulation action $\boldsymbol{a}_t^s$ was defined by a set of specific constraints $\mathcal{C}_t^s$ that delineated the spatial relationships between the robotic gripper, the transferred valid grasp configurations $\mathcal{G}_q$, and functional planes $\mathcal{F}_q$ of the identified instance q in the scene. These constraints fell into three main categories: point-to-point, point-to-plane, and plane-to-plane, with each action type typically corresponding to a fixed set of constraints. For example, to execute the action of "place bowl on plate," the VLM generated a plane-to-plane constraint requiring the base of the bowl to be parallel to the top surface of the plate, along with a point-to-point constraint that aligned the center points of the two corresponding functional planes in a horizontal direction. Let $\mathbf{F}_{\text{bowl}} = \langle \mathbf{n}_{\text{bowl}}, \mathbf{c}_{\text{bowl}} \rangle$ and $\mathbf{F}_{\text{plate}} = \langle \in \mathbf{n}_{\text{plate}}, \mathbf{c}_{\text{plate}} \rangle$ denote the planes of the bowl's base and the plate's top surface, respectively. The constraints were

$$\begin{aligned} &\text{plane to plane: } \mathbf{n}_{\text{bowl}}{}^T \mathbf{n}_{\text{plate}} = 1 \\ &\text{point to point: } \begin{cases} \mathbf{c}_{\text{bowl}}.x = \mathbf{c}_{\text{plate}}.x \\ \mathbf{c}_{\text{bowl}}.y = \mathbf{c}_{\text{plate}}.y \end{cases} \end{aligned} \tag{7}$$

where $\mathbf{c}_*.x$ and $\mathbf{c}_*.y$ represented the x and y locations, respectively, of a center point $\mathbf{c}_*$.

Similarly, for the action of "pick up the tilted spoon," the VLM first established two point-to-plane constraints to (i) align the gripper's approach direction with the normal of the spoon's handle surface and (ii) align the gripper's grasp center with the valid grasp configuration on the spoon's handle along its normal direction. Let $\mathbf{G}_{\text{gripper}} = \langle \mathbf{u}_{\text{gripper}}, \mathbf{t}_{\text{gripper}} \rangle$ denote the pose of the gripper, and $\mathbf{F}_{\text{spoon}} = \langle \mathbf{n}_{\text{spoon}}, \mathbf{c}_{\text{spoon}} \rangle$ denote the plane of the spoon's handle, the constraints were written as

$$\text{point to point: } \begin{cases} \mathbf{u}_{\text{gripper}}{}^T \mathbf{n}_{\text{spoon}} = 1 \\ (\mathbf{c}_{\text{spoon}} - \mathbf{t}_{\text{gripper}})^T \mathbf{n}_{\text{spoon}} = 1 \end{cases} \tag{8}$$

After establishing the set of constraints for each action $\boldsymbol{a}_t^s$ (the tth action for the sth subtask), a specialized parsing module was used to determine the target 6–degree of freedom (DoF) pose of the gripper that satisfied these constraints. Details for this process were provided in the "Details on gripper pose determination" section of the Supplementary Materials. Then, a binary value $e \in \{0, 1\}$ was added to represent the closed/open state of the gripper, forming a 7-DoF gripper pose $\langle \mathbf{G}_{\text{gripper}}, e \rangle \in SO(3) \times \{0, 1\}$. Consequently, the task-planning process broke down the entire task into multiple subtasks s and further decomposed each subtask into multiple actions $\boldsymbol{a}_t^s$. This resulted in an ordered list of 7-DOF poses for the gripper $\left\{ \left\langle \mathbf{G}_{t;\text{gripper}}^s, e_t^s \right\rangle \mid \forall s, t \right\}$, with each pose corresponding to a specific action $\boldsymbol{a}_t^s$.

Motion execution and robot hardware

After planning the manipulation task into an ordered list of gripper poses $\left\langle \mathbf{G}_{t;\text{gripper}}^s, e_t^s \right\rangle$, we used a low-level motion planner to generate a feasible and smooth manipulation trajectory between consecutive gripper poses. For simplicity, we focused on the trajectory optimization between a single pair of consecutive gripper poses, and we denoted the starting pose as $\mathbf{G}_{\text{init}} = \langle \mathbf{u}_{\text{init}}, \mathbf{t}_{\text{init}} \rangle$ and the ending pose as $\mathbf{G}_{\text{target}} = \left\langle \mathbf{u}_{\text{target}}, \mathbf{t}_{\text{target}} \right\rangle$.

Inspired by VoxPoser (*27*), we implemented a voxel-based motion planning framework. Specifically, we discretized the robot's workspace into a dense voxel grid $\mathbf{V}$ with a size of 100^3. We denoted any voxel as $\mathbf{v} \in \mathbf{V}$, the specific voxel corresponding to $\mathbf{t}_{\text{init}}$ as $\mathbf{v}_{\text{init}} \in \mathbf{V}$ and the corresponding voxel of $\mathbf{t}_{\text{target}}$ as $\mathbf{v}_{\text{target}} \in \mathbf{V}$. To facilitate the trajectory optimization, four key maps are created on $\mathbf{V}$. First, the affordance map assigned a nonnegative real value to each voxel to guide the approach toward the target regions for each action, which was defined as $\Lambda_{\text{aff}}: \mathbf{v} \to \mathbb{R}^+$. Second, the avoidance map assigned a nonnegative real value to each voxel to prevent contact with the obstacles, which was defined as $\Lambda_{\text{avd}}: \mathbf{v} \to \mathbb{R}^+$. Furthermore, the rotation map assigned the gripper orientation to each voxel, which was defined as $\Lambda_{\text{rot}}: \mathbf{v} \to \mathbb{R}^3$. To achieve this, we first assigned $\mathbf{u}_{\text{init}}$ to $\mathbf{v}_{\text{init}}$. Then, we assigned $\mathbf{u}_{\text{target}}$ to $\mathbf{v}_{\text{target}}$. Last, we applied Gaussian filtering to smooth the values for all voxels between the two assigned voxels. Last, the state map assigned a binary value to each voxel to indicate the target closed/open state of the gripper, which was defined as $\Lambda_{\text{sta}}: \mathbf{v} \to \{0, 1\}$. Specifically, we first assigned the initial state of the gripper e_{init} to $\mathbf{v}_{\text{init}}$ and then assigned the target state of the gripper e_{target} to $\mathbf{v}_{\text{target}}$, similarly applying value smoothing using Gaussian filtering. Both Λ_{aff} and Λ_{avd} were constructed from the surface points on objects with the help of our visual grounding model, rather than simply using the geometric centers of the objects. This enabled our system to identify semantically meaningful locations, such as a mug's handle or a jar's lid, enhancing manipulation capabilities beyond simple grasping.

Last, we computed a trajectory for each action, starting from $\mathbf{G}_{\text{init}}$ and ending at $\mathbf{G}_{\text{target}}$. Inspired by (*27*), we used a cost-guided sampling-based method for optimization to find the trajectory of the gripper's location first and then assigned the gripper's orientation along the trajectory. We first established a cost function for gripper's location $f: \mathbf{v} \to \mathbb{R}$, which was built on the basis of the affordance and the avoidance key maps: $f = -2\Lambda_{\text{aff}} + \Lambda_{\text{avd}}$. Starting from $\mathbf{t}_{\text{init}}$, we sampled multiple next-step locations $\Delta\mathbf{t}_1$ around $\mathbf{t}_{\text{init}}$ and compared the cost values of their corresponding voxel obtained using $f(\mathbf{v})$. The optimized next-step location $\mathbf{t}_1 = \mathbf{t}_{\text{init}} + \Delta\mathbf{t}_1$ was the one having the lowest cost value. The optimization was performed iteratively, and we optimized a trajectory of the gripper's location $\mathbf{t}_{1,2,3\ldots}$ until the gripper's target location $\mathbf{t}_{\text{target}}$ was reached. After determining the trajectory for the gripper's position, we assigned the gripper's orientation and open/closed state by referencing the voxels $\mathbf{v}_{1,2,3\ldots}$ corresponding

to each intermediate location along the trajectory. This was done by referring to Λ_{rot} for the orientation and Λ_{sta} for the state. Thus, starting from the current gripper pose, we sampled multiple offsets and chose the one with the best cost function value.

The robotic manipulation was performed by a FAIR Innovation FR5 robot arm that was equipped with a soft parallel-jaw gripper based on a Z-EFG electric controller. For visual sensing, both a ZED-2 camera and a RealSense camera were used. The VGGT model (*60*) was used for scene reconstruction, using two RGB images from the left and right ZED-2 cameras, along with an additional image from the RealSense camera. Two GeForce RTX 3090 GPUs were used for local network computation.

Statistical analysis

Statistical analysis was conducted using Python (version 3.9), NumPy (version 1.26.4), and SciPy (version 1.11.2). Quantitative results for real-world robotic manipulation tasks were obtained by using the average success rate across replicate trials. For benchmark evaluations, model performance was quantified using VQA accuracy and average accuracy based on the geodetic rotation error. A Friedman test was used, with a *P* value smaller than 0.01 considered statistically significant. Detailed *P* values were provided in the supplementary data file.

Supplementary Materials

This PDF file includes:
Materials and Methods
Figs. S1 to S17
Tables S1 to S4
Legends for movies S1 and S2

Other Supplementary Material for this manuscript includes the following:
Movies S1 and S2
Data file S1

REFERENCES AND NOTES

1. A. Gupta, S. Savarese, S. Ganguli, F.-F. Li, Embodied intelligence via learning and evolution. *Nat. Commun.* **12**, 5721 (2021).
2. J. Cui, J. Trinkle, Toward next-generation learned robot manipulation. *Sci. Robot.* **6**, eabd9461 (2021).
3. L. Shao, T. Migimatsu, Q. Zhang, K. Yang, J. Bohg, Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *Int. J. Robot. Res.* **40**, 1419–1434 (2021).
4. R. Mon-Williams, G. Li, R. Long, W. Du, C. G. Lucas, Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nat. Mach. Intell.* **7**, 592–601 (2025).
5. J. Zhang, J. Zhang, K. Pertsch, Z. Liu, X. Ren, M. Chang, S.-H. Sun, J. J. Lim, "Bootstrap your own skills: Learning to solve new tasks with large language model guidance" in *Proceedings of the 7th Conference on Robot Learning*, J. Tan, M. Toussaint, K. Darvish, Eds., vol. 229 of *Proceedings of Machine Learning Research* (PMLR, 2023), pp. 302–325.
6. Y. Hu, F. Lin, T. Zhang, L. Yi, Y. Gao, "Look before you leap: Unveiling the power of GPT-4V in robotic vision-language planning" in *Proceedings of the First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024* (Open Review, 2024); https://openreview.net/forum?id=n82dpqpa7J.
7. A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, "Open X-embodiment: Robotic learning datasets and RT-X models: Open X-embodiment collaboration" in *Proceedings of the IEEE International Conference on Robotics and Automation* (IEEE, 2024), pp. 6892–6903.
8. A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Baijal, M. G. Castro, D. Chen, Q. Chen, T. Chung, J. Drake, E. P. Foster, J. Gao, D. A. Herrera, M. Heo, K. Hsu, J. Hu, D. Jackson, C. Le, Y. Li, R. Lin, Z. Ma, A. Maddukuri, S. Mirchandani, D. Morton, T. Nguyen, A. O'Neill, R. Scalise, D. Seale, V. Son, S. Tian, E. Tran, A. E. Wang, Y Wu, A. Xie, J. Yang, P. Yin, Y. Zhang, O. Bastani, G. Berseth, J. Bohg, K. Goldberg, A. Gupta, A. Gupta, D. Jayaraman, J. J. Lim, J. Malik, R. Martín-Martín, S. Ramamoorthy, D. Sadigh, S. Song, J. Wu, M. C. Yip, Y. Zhu, T. Kollar, S. Levine, C. Finn, "DROID: A large-scale in-the-wild robot manipulation dataset" in *Proceedings of Robotics: Science and Systems* (RSS Foundation, 2024).
9. P. Hao, S. Cui, J. Wei, T. Lu, Y. Cai, S. Wang, Learn-Gen-Plan: Bridging the gap between vision language models and real-world long-horizon dexterous manipulations. *IEEE Trans. Autom. Sci. Eng.* **22**, 15638–15649 (2025).
10. Y. Feng, J. Han, Z. Yang, X. Yue, S. Levine, J. Luo, "Reflective planning: Vision-language models for multi-stage long-horizon robotic manipulation" in *Proceedings of the 9th Conference on Robot Learning*, J. Lim, S. Song, H.-W. Park, Eds., vol. 305 of *Proceedings of Machine Learning Research* (PMLR, 2025), pp. 2038–2062.
11. K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, N. Suenderhauf, "SayPlan: Grounding large language models using 3D scene graphs for scalable robot task planning" in *Proceedings of the 7th Conference on Robot Learning*, J. Tan, M. Toussaint, K. Darvish, Eds., vol. 229 of *Proceedings of Machine Learning Research* (PMLR, 2023), pp. 23–72.
12. H. Huang, X. Chen, Y. Chen, H. Li, X. Han, Z. Wang, T. Wang, J. Pang, Z. Zhao, "RoboGround: Robotic manipulation with grounded vision-language priors" in *2025 Proceedings of the Computer Vision and Pattern Recognition Conference* (IEEE, 2025), pp. 22540–22550.
13. J. Gao, B. Sarkar, F. Xia, T. Xiao, J. Wu, B. Ichter, A. Majumdar, D. Sadigh, "Physically grounded vision-language models for robotic manipulation" in *Proceedings of the IEEE International Conference on Robotics and Automation* (IEEE, 2024), pp. 12462–12469.
14. R. Wang, J. Mao, J. Hsu, H. Zhao, J. Wu, Y. Gao, "Programmatically grounded, compositionally generalizable robotic manipulation" in *The Eleventh International Conference on Learning Representations* (Open Review, 2023); https://openreview.net/forum?id=rZ-wylY5VI.
15. V. Bhat, P. Krishnamurthy, R. Karri, F. Khorrami, HiFi-CS: Towards open vocabulary visual grounding for robotic grasping using vision-language models. arXiv:2409.10419 [cs.RO] (2024).
16. S. Huang, I. Ponomarenko, Z. Jiang, X. Li, X. Hu, P. Gao, H. Li, H. Dong, "ManipVQA: Injecting robotic affordance and physically grounded information into multi-modal large language models" in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2024), pp. 7580–7587.
17. W. Cai, I. Ponomarenko, J. Yuan, X. Li, W. Yang, H. Dong, B. Zhao, "SpatialBot: Precise spatial understanding with vision language models" in *Proceedings of the IEEE International Conference on Robotics and Automation* (IEEE, 2025), pp. 9490–9498.
18. Y. Ding, H. Geng, C. Xu, X. Fang, J. Zhang, S. Wei, Q. Dai, Z. Zhang, H. Wang, "Open6DOR: Benchmarking open-instruction 6-DoF object rearrangement and a VLM-based approach" in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2024), pp. 7359–7366.
19. M. Dorkenwald, N. Barazani, C. G. Snoek, Y. M. Asano, "Pin: Positional insert unlocks object localisation abilities in VLMs" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2024), pp. 13548–13558.
20. R. Li, S. Li, L. Kong, X. Yang, J. Liang, "Seeground: See and ground for zero-shot open-vocabulary 3D visual grounding" in *Proceedings of the Computer Vision and Pattern Recognition Conference* (IEEE, 2025), pp. 3707–3717.
21. M. Li, S. Zhao, Q. Wang, K. Wang, Y. Zhou, S. Srivastava, C. Gokmen, T. Lee, E. L. Li, R. Zhang, W. Liu, P. Liang, L. Fei-Fei, J. Mao, J. Wu, "Embodied agent interface: Benchmarking LLMs for embodied decision making" in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang, Eds. (Curran Associates, 2024), vol. 37, pp. 100428–100534.
22. K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, F.-F. Li, S. Savarese, Learning task-oriented grasping for tool manipulation from simulated self-supervision. *Int. J. Robot. Res.* **39**, 202–216 (2020).
23. C. Tang, D. Huang, W. Dong, R. Xu, H. Zhang, FoundationGrasp: Generalizable task-oriented grasping with foundation models. *IEEE Trans. Autom. Sci. Eng.* **22**, 12418–12435 (2025).
24. X. Li, M. Zhang, Y. Geng, H. Geng, Y. Long, Y. Shen, R. Zhang, J. Liu, H. Dong, "ManipLLM: Embodied multimodal large language model for object-centric robotic manipulation" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2024), pp. 18061–18070.
25. R. Wu, K. Cheng, Y. Zhao, C. Ning, G. Zhan, H. Dong, "Learning environment-aware affordance for 3D articulated object manipulation under occlusions" in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine, Eds. (Curran Associates, 2023), vol. 36, pp. 60966–60983.
26. R. Xu, Y. Shen, X. Li, R. Wu, H. Dong, NaturalVLM: Leveraging fine-grained natural language for affordance-guided visual manipulation. *IEEE Robot. Autom. Lett.* **9**, 10842–10849 (2024).
27. W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, F.-F. Li, "VoxPoser: Composable 3D value maps for robotic manipulation with language models" in *Proceedings of the 7th Conference on Robot Learning*, J. Tan, M. Toussaint, K. Darvish, Eds., vol. 229 of *Proceedings of Machine Learning Research* (PMLR, 2023), pp. 540–562.

28. W. Huang, C. Wang, Y. Li, R. Zhang, F.-F. Li, "ReKep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation" in *Proceedings of the 8th Conference on Robot Learning*, P. Agrawal, O. Kroemer, W. Burgard, Eds., vol. 270 of *Proceedings of Machine Learning Research* (PMLR, 2024), pp. 4573–4602.
29. H. Huang, F. Lin, Y. Hu, S. Wang, Y. Gao, "CoPa: General robotic manipulation through spatial constraints of parts with foundation models" in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2024), pp. 9488–9495.
30. M. Pan, J. Zhang, T. Wu, Y. Zhao, W. Gao, H. Dong, "OmniManip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2025), pp. 17359–17369.
31. B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, F. Xia, "SpatialVLM: Endowing vision-language models with spatial reasoning capabilities" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2024), pp. 14455–14465.
32. A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R.-H. Yang, J. Kautz, X. Wang, S. Liu, "SpatialRGPT: Grounded spatial reasoning in vision-language models" in *Proceedings of Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang, Eds. (Curran Associates, 2024), vol. 37, pp. 135062–135093.
33. W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, D. Fox, "RoboPoint: A vision-language model for spatial affordance prediction for robotics" in *Proceedings of the 8th Conference on Robot Learning*, P. Agrawal, O. Kroemer, W. Burgard, Eds., vol. 270 of *Proceedings of Machine Learning Research* (PMLR, 2024), pp. 4005–4020.
34. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks" in *Proceedings of Advances in Neural Information Processing System*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. (Curran Associates, 2020), vol. 33, pp. 9459–9474.
35. Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997 [cs.CL] (2023).
36. Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, G. Neubig, "Active retrieval augmented generation" in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, K. Bali, Eds. (Association for Computational Linguistics, 2023), pp. 7969–7992.
37. J. Chen, H. Lin, X. Han, L. Sun, "Benchmarking large language models in retrieval-augmented generation" in *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI, 2024), vol. 38, pp. 17754–17762.
38. W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, Q. Li, "A survey on rag meeting LLMs: Towards retrieval-augmented large language models" in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2024), pp. 6491–6501.
39. Gemini Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, D. Silver, M. Johnson, I. Antonoglou, J. Schrittwieser, A. Glaese, J. Chen, E. Pitler, T. Lillicrap, A. Lazaridou, O. Firat, J. Molloy, M. Isard, P. R. Barham, T. Hennigan, B. Lee, F. Viola, M. Reynolds, Y. Xu, R. Doherty, E. Collins, C. Meyer, E. Rutherford, E. Moreira, K. Ayoub, M. Goel, J. Krawczyk, C. Du, E. Chi, H.-T. Cheng, E. Ni, P. Shah, P. Kane, B. Chan, M. Faruqui, A. Severyn, H. Lin, Y. Li, Y. Cheng, A. Ittycheriah, M. Mahdieh, M. Chen, P. Sun, D. Tran, S. Bagri, B. Lakshminarayanan, J. Liu, A. Orban, F. Güra, H. Zhou, X. Song, A. Boffy, H. Ganapathy, S. Zheng, H. Choe, Á. Weisz, T. Zhu, Y. Lu, S. Gopal, J. Kahn, M. Kula, J. Pitman, R. Shah, E. Taropa, M. Al Merey, M. Baeuml, Z. Chen, L. El Shafey, Y. Zhang, O. Sercinoglu, G. Tucker, E. Piqueras, M. Krikun, I. Barr, N. Savinov, I. Danihelka, B. Roelofs, A. White, A. Andreassen, T. von Glehn, L. Yagati, M. Kazemi, L. Gonzalez, M. Khalman, J. Sygnowski, A. Frechette, C. Smith, L. Culp, L. Proleev, Y. Luan, X. Chen, J. Lottes, N. Schucher, F. Lebron, A. Rrustemi, N. Clay, P. Crone, T. Kocisky, J. Zhao, B. Perz, D. Yu, H. Howard, A. Bloniarz, J. W. Rae, H. Lu, L. Sifre, M. Maggioni, F. Alcober, D. Garrette, M. Barnes, S. Thakoor, J. Austin, G. Barth-Maron, W. Wong, R. Joshi, R. Chaabouni, D. Fatiha, A. Ahuja, G. S. Tomar, E. Senter, M. Chadwick, I. Kornakov, N. Attaluri, I. Iturrate, R. Liu, Y. Li, S. Cogan, J. Chen, C. Jia, C. Gu, Q. Zhang, J. Grimstad, A. J. Hartman, X. Garcia, T. S. Pillai, J. Devlin, M. Laskin, D. de Las Casas, D. Valter, C. Tao, L. Blanco, A. P. Badia, D. Reitter, M. Chen, J. Brennan, C. Rivera, S. Brin, S. Iqbal, G. Surita, J. Labanowski, A. Rao, S. Winkler, E. Parisotto, Y. Gu, K. Olszewska, R. Addanki, A. Miech, A. Louis, D. Teplyashin, G. Brown, E. Catt, J. Balaguer, J. Xiang, P. Wang, Z. Ashwood, A. Briukhov, A. Webson, S. Ganapathy, S. Sanghavi, A. Kannan, M.-W. Chang, A. Stjerngren, J. Djolonga, Y. Sun, A. Bapna, M. Aitchison, P. Pejman, H. Michalewski, T. Yu, C. Wang, J. Love, J. Ahn, D. Bloxwich, K. Han, P. Humphreys, T. Sellam, J. Bradbury, V. Godbole, S. Samangooei, B. Damoc, A. Kaskasoli, S. M. R. Arnold, V. Vasudevan, S. Agrawal, J. Riesa, D. Lepikhin, R. Tanburn, S. Srinivasan, H. Lim, S. Hodkinson, P. Shyam, J. Ferret, S. Hand, A. Garg, T. Le Paine, J. Li, Y. Li, M. Giang, A. Neitz, Z. Abbas, S. York, M. Reid, E. Cole, A. Chowdhery, D. Das, D. Rogozińska, V. Nikolaev, P. Sprechmann, Z. Nado, L. Zilka, F. Prost, L. He, M. Monteiro, G. Mishra, C. Welty, J. Newlan, D. Jia, M. Allamanis, C. H. Hu, R. de Liedekerke, J. Gilmer, C. Saroufim, S. Rijhwani, S. Hou, D. Shrivastava, A. Baddepudi, A. Goldin, A. Ozturel, A. Cassirer, Y. Xu, D. Sohn, D. Sachan, R. K. Amplayo, C. Swanson, D. Petrova, S. Narayan, A. Guez, S. Brahma, J. Landon, M. Patel, R. Zhao, K. Villela, L. Wang, W. Jia, M. Rahtz, M. Giménez, L. Yeung, J. Keeling, P. Georgiev, D. Mincu, B. Wu, S. Haykal, R. Saputro, K. Vodrahalli, J. Qin, Z. Cankara, A. Sharma, N. Fernando, W. Hawkins, B. Neyshabur, S. Kim, A. Hutter, P. Agrawal, A. Castro-Ros, G. van den Driessche, T. Wang, F. Yang, S.-y. Chang, P. Komarek, R. McIlroy, M. Lučić, G. Zhang, W. Farhan, M. Sharman, P. Natsev, P. Michel, Y. Bansal, S. Qiao, K. Cao, S. Shakeri, C. Butterfield, J. Chung, P. K. Rubenstein, S. Agrawal, A. Mensch, K. Soparkar, K. Lenc, T. Chung, A. Pope, L. Maggiore, J. Kay, P. Jhakra, S. Wang, J. Maynez, M. Phuong, T. Tobin, A. Tacchetti, M. Trebacz, K. Robinson, Y. Katariya, S. Riedel, P. Bailey, K. Xiao, N. Ghelani, L. Aroyo, A. Slone, N. Houlsby, X. Xiong, Z. Yang, E. Gribovskaya, J. Adler, M. Wirth, L. Lee, M. Li, T. Kagohara, J. Pavagadhi, S. Bridgers, A. Bortsova, S. Ghemawat, Z. Ahmed, T. Liu, R. Powell, V. Bolina, M. Iinuma, P. Zablotskaia, J. Besley, D.-W. Chung, T. Dozat, R. Comanescu, X. Si, J. Greer, G. Su, M. Polacek, R. L. Kaufman, S. Tokumine, H. Hu, E. Buchatskaya, Y. Miao, M. Elhawaty, A. Siddhant, N. Tomasev, J. Xing, C. Greer, H. Miller, S. Ashraf, A. Roy, Z. Zhang, A. Ma, A. Filos, M. Besta, R. Blevins, T. Klimenko, C.-K. Yeh, S. Changpinyo, J. Mu, O. Chang, M. Pajarskas, C. Muir, V. Cohen, C. Le Lan, K. Haridasan, A. Marathe, S. Hansen, S. Douglas, R. Samuel, M. Wang, S. Austin, C. Lan, J. Jiang, J. Chiu, J. A. Lorenzo, L. L. Sjösund, S. Cevey, Z. Gleicher, T. Avrahami, A. Boral, H. Srinivasan, V. Selo, R. May, K. Aisopos, L. Hussenot, L. B. Soares, K. Baumli, M. B. Chang, A. Recasens, B. Caine, A. Pritzel, F. Pavetic, F. Pardo, A. Gergely, J. Frye, V. Ramasesh, D. Horgan, K. Badola, N. Kassner, S. Roy, E. Dyer, V. C. Campos, A. Tomala, Y. Tang, D. El Badawy, E. White, B. Mustafa, O. Lang, A. Jindal, S. Vikram, Z. Gong, S. Caelles, R. Hemsley, G. Thornton, F. Feng, W. Stokowiec, C. Zheng, P. Thacker, Ç. Ünlü, Z. Zhang, M. Saleh, J. Svensson, M. Bileschi, P. Patil, A. Anand, R. Ring, K. Tsihlas, A. Vezer, M. Selvi, T. Shevlane, M. Rodriguez, T. Kwiatkowski, S. Daruki, K. Rong, A. Dafe, N. FitzGerald, K. Gu-Lemberg, M. Khan, L. A. Hendricks, M. Pellat, V. Feinberg, J. Cobon-Kerr, T. Sainath, M. Rauh, S. H. Hashemi, R. Ives, Y. Hasson, E. Noland, Y. Cao, N. Byrd, L. Hou, Q. Wang, T. Sottiaux, M. Paganini, J.-B. Lespiau, A. Moufarek, S. Hassan, K. Shivakumar, J. van Amersfoort, A. Mandhane, P. Joshi, A. Goyal, M. Tung, A. Brock, H. Sheahan, V. Misra, C. Li, N. Rakićević, M. Dehghani, F. Liu, S. Mittal, J. Oh, S. Noury, E. Sezener, F. Huot, M. Lamm, N. De Cao, C. Chen, S. Mudgal, R. Stella, K. Brooks, G. Vasudevan, C. Liu, M. Chain, N. Melinkeri, A. Cohen, V. Wang, K. Seymore, S. Zubkov, R. Goel, S. Yue, S. Krishnakumaran, B. Albert, N. Hurley, M. Sano, A. Mohananey, J. Joughin, E. Filonov, T. Kępa, Y. Eldawy, J. Lim, R. Rishi, S. Badiezadegan, T. Bos, J. Chang, S. Jain, S. G. S. Padmanabhan, S. Puttagunta, K. Krishna, L. Baker, N. Kalb, V. Bedapudi, A. Kurzrok, S. Lei, A. Yu, O. Litvin, X. Zhou, Z. Wu, S. Sobell, A. Siciliano, A. Papir, R. Neale, J. Bragagnolo, T. Toor, T. Chen, V. Anklin, F. Wang, R. Feng, M. Gholami, K. Ling, L. Liu, J. Walter, H. Moghaddam, A. Kishore, J. Adamek, T. Mercado, J. Mallinson, S. Wandekar, S. Cagle, E. Ofek, G. Garrido, C. Lombriser, M. Mukha, B. Sun, H. R. Mohammad, J. Matak, Y. Qian, V. Peswani, P. Janus, Q. Yuan, L. Schelin, O. David, A. Garg, Y. He, O. Duzhyi, A. Älgmyr, T. Lottaz, Q. Li, V. Yadav, L. Xu, A. Chinien, R. Shivanna, A. Chuklin, J. Li, C. Spadine, T. Wolfe, K. Mohamed, S. Das, Z. Dai, K. He, D. von Dincklage, S. Upadhyay, A. Maurya, L. Chi, S. Krause, K. Salama, P. G. Rabinovitch, M. P. K. Reddy, A. Selvan, M. Dektiarev, G. Ghiasi, E. Guven, H. Gupta, B. Liu, D. Sharma, I. H. Shtacher, S. Paul, O. Akerlund, F.-X. Aubet, T. Huang, C. Zhu, E. Zhu, E. Teixeira, M. Fritze, F. Bertolini, L.-E. Marinescu, M. Bölle, D. Paulus, K. Gupta, T. Latkar, M. Chang, J. Sanders, R. Wilson, X. Wu, Y.-X. Tan, L. N. Thiet, T. Doshi, S. Lall, S. Mishra, W. Chen, T. Luong, S. Benjamin, J. Lee, E. Andrejczuk, D. Rabiej, V. Ranjan, K. Styrc, P. Yin, J. Simon, M. R. Harriott, M. Bansal, A. Robsky, G. Bacon, D. Greene, D. Mirylenka, C. Zhou, O. Sarvana, A. Goyal, S. Andermatt, P. Siegler, B. Horn, A. Israel, F. Pongetti, C.-W. "L." Chen, M. Selvatici, P. Silva, K. Wang, J. Tolins, K. Guu, R. Yogev, X. Cai, A. Agostini, M. Shah, H. Nguyen, N. Ó Donnaile, S. Pereira, L. Friso, A. Stambler, A. Kurzrok, C. Kuang, Y. Romanikhin, M. Geller, Z. J. Yan, K. Jang, C.-C. Lee, W. Fica, E. Malmi, Q. Tan, D. Banica, D. Balle, R. Pham, Y. Huang, D. Avram, H. Shi, J. Singh, C. Hidey, N. Ahuja, P. Saxena, D. Dooley, S. P. Potharaju, E. O'Neill, A. Gokulchandran, R. Foley, K. Zhao, M. Dusenberry, Y. Liu, P. Mehta, R. Kotikalapudi, C. Safranek-Shrader, A. Goodman, J. Kessinger, E. Globen, P. Kolhar, C. Gorgolewski, A. Ibrahim, Y. Song, A. Eichenbaum, T. Brovelli, S. Potluri, P. Lahoti, C. Baetu, A. Ghorbani, C. Chen, A. Crawford, S. Pal, M. Sridhar, P. Gurita, A. Mujika, I. Petrovski, P.-L. Cedoz, C. Li, S. Chen, N. D. Santo, S. Goyal, J. Punjabi, K. Kappaganthu, C. Kwak, Pallavi L. V., S. Velury, H. Choudhury, J. Hall, P. Shah, R. Figueira, M. Thomas, M. Lu, T. Zhou, C. Kumar, T. Jurdi, S. Chikkerur, Y. Ma, A. Yu, S. Kwak, V. Ähdel, S. Rajayogam, T. Choma, F. Liu, A. Barua, C. Ji, J. H. Park, V. Hellendoorn, A. Bailey, T. Bilal, H. Zhou, M. Khatir, C. Sutton, W. Rzadkowski, F. Macintosh, R. Vij, K. Shagin, P. Medina, C. Liang, J. Zhou, P. Shah, Y. Bi, A. Dankovics, S. Banga, S. Lehmann, M. Bredesen, Z. Lin, J. E. Hoffmann, J. Lai, R. Chung, K. Yang, N. Balani, A. Bražinskas, A. Sozanschi, M. Hayes, H. F. Alcalde, P. Makarov, W. Chen, A. Stella, L. Snijders, M. Mandl, A. Kärrman, P. Nowak, X. Wu, A. Dyck, K. Vaidyanathan, R. Raghavender, J. Mallet, M. Rudominer, E. Johnston, S. Mittal, A. Udathu, J. Christensen, V. Verma, Z. Irving, A. Santucci, G. Elsayed, E. Davoodi, M. Georgiev, I. Tenney, N. Hua, G. Cideron, E. Leurent, M. Alnahlawi, I. Georgescu, N. Wei, I. Zheng, D. Scandinaro, H. Jiang, J. Snoek, M. Sundararajan, X. Wang, Z. Ontiveros, I. Karo, J. Cole, V. Rajashekhar, L. Tumeh, E. Ben-David, R. Jain, J. Uesato, R. Datta, O. Bunyan, S. Wu, J. Zhang, P. Stanczyk, Y. Zhang, D. Steiner, S. Naskar, M. Azzam, M. Johnson, A. Paszke, C.-C. Chiu, J. S. Elias, A. Mohiuddin, F. Muhammad, J. Miao, A. Lee, N. Vieillard, J. Park, J. Zhang, J. Stanway,

D. Garmon, A. Karmarkar, Z. Dong, J. Lee, A. Kumar, L. Zhou, J. Evens, W. Isaac, G. Irving, E. Loper, M. Fink, I. Arkatkar, N. Chen, I. Shafran, I. Petrychenko, Z. Chen, J. Jia, A. Levskaya, Z. Zhu, P. Grabowski, Y. Mao, A. Magni, K. Yao, J. Snaider, N. Casagrande, E. Palmer, P. Suganthan, A. Castaño, I. Giannoumis, W. Kim, M. Rybiński, A. Sreevatsa, J. Prendki, D. Soergel, A. Goedeckemeyer, W. Gierke, M. Jafari, M. Gaba, J. Wiesner, D. G. Wright, Y. Wei, H. Vashisht, Y. Kulizhskaya, J. Hoover, M. Le, L. Li, C. Iwuanyanwu, L Liu, K. Ramirez, A. Khorlin, A. Cui, T. Lin, M. Wu, R. Aguilar, K. Pallo, A. Chakladar, G. Perng, E. A. Abellan, M. Zhang, I. Dasgupta, N. Kushman, I. Penchev, A. Repina, X. Wu, T. van der Weide, P. Ponnapalli, C. Kaplan, J. Simsa, S. Li, O. Dousse, F. Yang, J. Piper, N. Ie, R. Pasumarthi, N. Lintz, A. Vijayakumar, D. Andor, P. Valenzuela, M. Lui, C. Paduraru, D. Peng, K. Lee, S. Zhang, S. Greene, D. D. Nguyen, P. Kurylowicz, C. Hardin, L. Dixon, L. Janzer, K. Choo, Z. Feng, B. Zhang, A. Singhal, D. Du, D. McKinnon, N. Antropova, T. Bolukbasi, O. Keller, D. Reid, D. Finchelstein, M. A. Raad, R. Crocker, P. Hawkins, R. Dadashi, C. Gaffney, K. Franko, A. Bulanova, R. Leblond, S. Chung, H. Askham, L. C. Cobo, K. Xu, F. Fischer, J. Xu, C. Sorokin, C. Alberti, C.-C. Lin, C. Evans, A. Dimitriev, H. Forbes, D. Banarse, Z. Tung, M. Omernick, C. Bishop, R. Sterneck, R. Jain, J. Xia, E. Amid, F. Piccinno, X. Wang, P. Banzal, D. J. Mankowitz, A. Polozov, V. Krakovna, S. Brown, M. H. Bateni, D. Duan, V. Firoiu, M. Thotakuri, T. Natan, M. Geist, S. tan Girgin, H. Li, J. Ye, O. Roval, R. Tojo, M. Kwong, J. Lee-Thorp, C. Yew, D. Sinopalnikov, S. Ramos, J. Mellor, A. Sharma, K. Wu, D. Miller, N. Sonnerat, D. Vnukov, R. Greig, J. Beattie, E. Caveness, L. Bai, J. Eisenschlos, A. Korchemniy, T. Tsai, M. Jasarevic, W. Kong, P. Dao, Z. Zheng, F. Liu, F. Yang, R. Zhu, T. Huey Teh, J. Sanmiya, E. Gladchenko, N. Trdin, D. Toyama, E. Rosen, S. Tavakkol, L. Xue, C. Elkind, O. Woodman, J. Carpenter, G. Papamakarios, R. Kemp, S. Kafle, T. Grunina, R. Sinha, A. Talbert, D. Wu, D. Owusu-Afriyie, C. Du, C. Thornton, J. Pont-Tuset, P. Narayana, J. Li, S. Fatehi, J. Wieting, O. Ajmeri, B. Uria, Y. Ko, L. Knight, A. Héliou, N. Niu, S. Gu, C. Pang, Y. Li, N. Levine, A. Stolovich, R. Santamaria-Fernandez, S. Goenka, W. Yustalim, R. Strudel, A. Elqursh, C. Deck, H. Lee, Z. Li, K. Levin, R. Hoffmann, D. Holtmann-Rice, O. Bachem, S. Arora, C. Koh, S. H. Yeganeh, S. Põder, M. Tariq, Y. Sun, L. Ionita, M. Seyedhosseini, P. Tafti, Z. Liu, A. Gulati, J. Liu, X. Ye, B. Chrzaszcz, L. Wang, N. Sethi, T. Li, B. Brown, S. Singh, W. Fan, A. Parisi, J. Stanton, V. Koverkathu, C. A. Choquette-Choo, Y. Li, T. J. Lu, A. Ittycheriah, P. Shroff, M. Varadarajan, S. Bahargam, R. Willoughby, D. Gaddy, G. Desjardins, M. Cornero, B. Robenek, B. Mittal, B. Albrecht, A. Shenoy, F. Moiseev, H. Jacobsson, A. Ghaffarkhah, M. Rivière, A. Walton, C. Crepy, A. Parrish, Z. Zhou, C. Farabet, C. Radebaugh, P. Srinivasan, C. van der Salm, A. Fidjeland, S. Scellato, E. Latorre-Chimoto, H. Klimczak-Plucińska, D. Bridson, D. de Cesare, T. Hudson, P. Mendolicchio, L. Walker, A. Morris, M. Mauger, A. Guseynov, A. Reid, S. Odoom, L. Loher, V. Cotruta, M. Yenugula, D. Grewe, A. Petrushkina, T. Duerig, A. Sanchez, S. Yadlowsky, A. Shen, A. Globerson, L. Webb, S. Dua, D. Li, S. Bhupatiraju, D. Hurt, H. Qureshi, A. Agarwal, T. Shani, M. Eyal, A. Khare, S. R. Belle, L. Wang, C. Tekur, M. S. Kale, J. Wei, R. Sang, B. Saeta, T. Liechty, Y. Sun, Y. Zhao, S. Lee, P. Nayak, D. Fritz, M. R. Vuyyuru, J. Aslanides, N. Vyas, M. Wicke, X. Ma, E. Eltyshev, N. Martin, H. Cate, J. Manyika, K. Amiri, Y. Kim, X. Xiong, K. Kang, F. Luisier, N. Tripuraneni, D. Madras, M. Guo, A. Waters, O. Wang, J. Ainslie, J. Baldridge, H Zhang, G. Pruthi, J. Bauer, F. Yang, R. Mansour, J. Gelman, Y. Xu, G. Polovets, J. Liu, H. Cai, W. Chen, X. Sheng, E. Xue, S. Ozair, C. Angermueller, X. Li, A. Sinha, W. Wang, J. Wiesinger, E. Koukoumidis, Y. Tian, A. Iyer, M. Gurumurthy, M. Goldenson, P. Shah, M. K. Blake, H. Yu, A. Urbanowicz, J. Palomaki, C. Fernando, K. Durden, H. Mehta, N. Momchev, E. Rahimtoroghi, M. Georgaki, A. Raul, S. Ruder, M. Redshaw, J. Lee, D. Zhou, K. Jalan, D. Li, B. Hechtman, P. Schuh, M. Nasr, K. Milan, V. Mikulik, J. Franco, T. Green, N. Nguyen, J. Kelley, A. Mahendru, A. Hu, J. Howland, B. Vargas, J. Hui, K. Bansal, V. Rao, R. Ghiya, E. Wang, K. Ye, J. M. Sarr, M. M. Preston, M. Elish, S. Li, A. Kaku, J. Gupta, I. Pasupat, D.-C. Juan, M. Someswar, Tejvi M., X. Chen, A. Amini, A. Fabrikant, E. Chu, X. Dong, A. Muthal, S. Buthpitiya, S. Jauhari, N. Hua, U. Khandelwal, A. Hitron, J. Ren, L. Rinaldi, S. Drath, A. Dabush, N.-J. Jiang, H. Godhia, U. Sachs, A. Chen, Y. Fan, H. Taitelbaum, H. Noga, Z. Dai, J. Wang, C. Liang, J. Hamer, C.-S. Ferng, C. Elkind, A. Atias, P. Lee, V. Listík, M. Carlen, J. van de Kerkhof, M. Pikus, K. Zaher, P. Müller, S. Zykova, R. Stefanec, V. Gatsko, C. Hirnschall, A. Sethi, X. F. Xu, C. Ahuja, B. Tsai, A. Stefanoiu, B. Feng, K. Dhandhania, M. Katyal, A. Gupta, A. Parulekar, D. Pitta, J. Zhao, V. Bhatia, Y. Bhavnani, O. Alhadlaq, X. Li, P. Danenberg, D. Tu, A. Pine, V. Filippova, A. Ghosh, B. Limonchik, B. Urala, C. K. Lanka, D. Clive, Y. Sun, E. Li, H. Wu, K. Hongtongsak, I. Li, K. Thakkar, K. Omarov, K. Majmundar, M. Alverson, M. Kucharski, M. Patel, M. Jain, M. Zabelin, P. Pelagatti, R. Kohli, S. Kumar, J. Kim, S. Sankar, V. Shah, L. Ramachandruni, X. Zeng, B. Bariach, L. Weidinger, T. Vu, A. Andreev, A. He, K. Hui, S. Kashem, A. Subramanya, S. Hsiao, D. Hassabis, K. Kavukcuoglu, A. Sadovsky, Q. Le, T. Strohman, Y. Wu, S. Petrov, J. Dean, O. Vinyals, Gemini: A family of highly capable multimodal models. arXiv:2312.11805 [cs.CL] (2023).

40. S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, L. Zhang, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection" in *Computer Vision—ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVII*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol, Eds., vol. 15105 of *Lecture Notes in Computer Science* (Springer, 2024), pp. 38–55.
41. M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, P. Bojanowski, DINOv2: Learning robust visual features without supervision. arXiv:2304.07193 [cs.CV] (2023).
42. OpenAI, A. Hurst, A. Lerer, A. P Goucher, A. Perelman, A. Ramesh, A. Clark, A. J. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Mądry, A. Baker-Whitcomb, A. Beutel, A. Borzunov, A. Carney, A. Chow, A. Kirillov, A. Nichol, A. Paino, A. Renzin, A. T. Passos, A. Kirillov, A. Christakis, A. Conneau, A. Kamali, A. Jabri, A. Moyer, A. Tam, A. Crookes, A. Tootoochian, A. Tootoonchian, A. Kumar, A. Vallone, A. Karpathy, A. Braunstein, A. Cann, A. Codispoti, A. Galu, A. Kondrich, A. Tulloch, A. Mishchenko, A. Baek, A. Jiang, A. Pelisse, A. Woodford, A. Gosalia, A. Dhar, A. Pantuliano, A. Nayak, A. Oliver, B. Zoph, B. Ghorbani, B. Leimberger, B. Rossen, B. Sokolowsky, B. Wang, B. Zweig, B. Hoover, B. Samic, B. McGrew, B. Spero, B. Giertler, B. Cheng, B. Lightcap, B. Walkin, B. Quinn, B. Guarraci, B. Hsu, B. Kellogg, B. Eastman, C. Lugaresi, C. Wainwright, C. Bassin, C. Hudson, C. Chu, C. Nelson, C. Li, C. J. Shern, C. Conger, C. Barette, C. Voss, C. Ding, C. Lu, C. Zhang, C. Beaumont, C. Hallacy, C. Koch, C. Gibson, C. Kim, C. Choi, C. McLeavey, C. Hesse, C. Fischer, C. Winter, C. Czarnecki, C. Jarvis, C. Wei, C. Koumouzelis, D. Sherburn, D. Kappler, D. Levin, D. Levy, D. Carr, D. Farhi, D. Mely, D. Robinson, D. Sasaki, D. Jin, D. Valladares, D. Tsipras, D. Li, D. P. Nguyen, D. Findlay, E. Oiwoh, E. Wong, E. Asdar, E. Proehl, E. Yang, E. Antonow, E. Kramer, E. Peterson, E. Sigler, E. Wallace, E. Brevdo, E. Mays, F. Khorasani, F. P. Such, F. Raso, F. Zhang, F. von Lohmann, F. Sulit, G. Goh, G. Oden, G. Salmon, G. Starace, G. Brockman, H. Salman, H. Bao, H. Hu, H. Wong, H. Wang, H. Schmidt, H. Whitney, H. Jun, H. Kirchner, H. Ponde de Oliveira Pinto, H. Ren, H. Chang, H. W. Chung, I. Kivlichan, I. O'Connell, I. O'Connell, I. Osband, I. Silber, I. Sohl, I. Okuyucu, I. Lan, I. Kostrikov, I. Sutskever, I. Kanitscheider, I. Gulrajani, J. Coxon, J. Menick, J. Pachocki, J. Aung, J. Betker, J. Crooks, J. Lennon, J. Kiros, J. Leike, J. Park, J. Kwon, J. Phang, J. Teplitz, J. Wei, J. Wolfe, J. Chen, J. Harris, J. Varavva, J. G. Lee, J. Shieh, J. Lin, J. Yu, J. Weng, J. Tang, J. Yu, J. Jang, J. Q. Candela, J. Beutler, J. Landers, J. Parish, J. Heidecke, J. Schulman, J. Lachman, J. McKay, J. Uesato, J. Ward, J. W. Kim, J. Huizinga, J. Sitkin, J. Kraaijeveld, J. Gross, J. Kaplan, J. Snyder, J. Achiam, J. Jiao, J. Lee, J. Zhuang, J. Harriman, K. Fricke, K. Hayashi, K. Singhal, K. Shi, K. Karthik, K. Wood, K. Rimbach, K. Hsu, K. Nguyen, K. Gu-Lemberg, K. Button, K. Liu, K. Howe, K. Muthukumar, K. Luther, L. Ahmad, L. Kai, L. Itow, L. Workman, L. Pathak, L. Chen, L. Jing, L. Guy, L. Fedus, L. Zhou, L. Mamitsuka, L. Weng, L. McCallum, L. Held, L. Ouyang, L. Feuvrier, L. Zhang, L. Kondraciuk, L. Kaiser, L. Hewitt, L. Metz, L. Doshi, M. Aflak, M. Simens, M. Boyd, M. Thompson, M. Dukhan, M. Chen, M. Gray, M. Hudnall, M. Zhang, M. Aljubeh, M. Litwin, M. Zeng, M. Johnson, M. Shetty, M. Gupta, M. Shah, M. Yatbaz, M. J. Yang, M. Zhong, M. Glaese, M. Chen, M. Janner, M. Lampe, M. Petrov, M. Wu, M. Wang, M. Fradin, M. Pokrass, M. Castro, M. O. Temudo de Castro, M. Pavlov, M. Brundage, M. Wang, M. Khan, M. Murati, M. Bavarian, M. Lin, M. Yesildal, N. Soto, N. Gimelshein, N. Cone, N. Staudacher, N. Summers, N. LaFontaine, N. Chowdhury, N. Ryder, N. Stathas, N. Turley, N. Tezak, N. Felix, N. Kudige, N. Keskar, N. Deutsch, N. Bundick, N. Puckett, O. Nachum, O. Okelola, O. Boiko, O. Murk, O. Jaffe, O. Watkins, O. Godement, O. Campbell-Moore, P. Chao, P. McMillan, P. Belov, P. Su, P. Bak, P. Bakkum, P. Deng, P. Dolan, P. Hoeschele, P. Welinder, P. Tillet, P. Pronin, P. Tillet, P. Dhariwal, Q. Yuan, R. Dias, R. Lim, R. Arora, R. Troll, R. Lin, R. G. Lopes, R. Puri, R. Miyara, R. Leike, R. Gaubert, R. Zamani, R. Wang, R. Donnelly, R. Honsby, R. Smith, R. Sahai, R. Ramchandani, R. Huet, R. Carmichael, R. Zellers, R. Chen, R. Chen, R. Nigmatullin, R. Cheu, S. Jain, S. Altman, S. Schoenholz, S. Toizer, S. Miserendino, S. Agarwal, S. Culver, S. Ethersmith, S. Gray, S. Grove, S. Metzger, S. Hermani, S. Jain, S. Zhao, S. Wu, S. Jomoto, S. Wu, S. (T.) Xia, S. Phene, S. Papay, S. Narayanan, S. Coffey, S. Lee, S. Hall, S. Balaji, T. Broda, T. Stramer, T. Xu, T. Gogineni, T. Christianson, T. Sanders, T. Patwardhan, T. Cunninghman, T. Degry, T. Dimson, T. Raoux, T. Shadwell, T. Zheng, T. Underwood, T. Markov, T. Sherbakov, T. Rubin, T. Stasi, T. Kaftan, T. Heywood, T. Peterson, T. Walters, T. Eloundou, V. Qi, V. Moeller, V. Monaco, V. Kuo, V. Fomenko, W. Chang, W. Zheng, W. Zhou, W. Manassra, W. Sheu, W. Zaremba, Y. Patil, Y. Qian, Y. Kim, Y. Cheng, Y. Zhang, Y. He, Y. Zhang, Y. Jin, Y. Dai, Y. Malkov, GPT-4o system card. arXiv:2410.21276 (2024).
43. Anthropic AI, Model card and evaluations for Claude models (2023); https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf.
44. S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, Qwen2.5-VL technical report. arXiv:2502.13923 [cs.CV] (2025).
45. Y. Zeng, M. Wu, L. Yang, J. Zhang, H. Ding, H. Cheng, H. Dong, LVDiffusor: Distilling functional rearrangement priors from large models into Diffusor. *IEEE Robot. Autom. Lett.* **9**, 8258–8265 (2024).
46. G. Zhai, X. Cai, D. Huang, Y. Di, F. Manhardt, F. Tombari, N. Navab, B. Busam, "SG-Bot: Object rearrangement via coarse-to-fine robotic imagination on scene graphs" in *Proceedings of IEEE International Conference on Robotics and Automation* (IEEE, 2024), pp. 4303–4310.
47. J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, D. Novotny, "Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction" in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2021), pp. 10901–10911.
48. W. Goodwin, S. Vaze, I. Havoutis, I. Posner, "Zero-shot category-level object pose estimation" in *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, S. Avidan, G. Brostow, M. Cissé,

G. M. Farinella, T. Hassner, Eds., vol. 13699 of *Lecture Notes in Computer Science* (Springer, 2022), pp. 516–532.
49. D. Q. Huynh, Metrics for 3D rotations: Comparison and analysis. *J. Math. Imaging Vis.* **35**, 155–164 (2009).
50. E. P. Örnek, Y. Labbé, B. Tekin, L. Ma, C. Keskin, C. Forster, T. Hodan, "FoundPose: Unseen object pose estimation with foundation features" in *Computer Vision—ECCV 2024: 18th European Conference*, *Milan*, *Italy*, *September 29–October 4*, *2024*, *Proceedings*, *Part XLVII*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol, Eds., vol. 15105 of *Lecture Notes in Computer Science* (Springer, 2024), pp. 163–182.
51. Y. Wu, X. Wang, X. Yang, M. Liu, D. Zeng, H. Ye, S. Li, "Learning occlusion-robust vision transformers for real-time UAV tracking" in *Proceedings of the IEEE/CVF Conference* on *Computer Vision and Pattern Recognition* (IEEE, 2025), pp. 17103–17113.
52. S. Suresh, H. Qi, T. Wu, T. Fan, L. Pineda, M. Lambeta, J. Malik, M. Kalakrishnan, R. Calandra, M. Kaess, J. Ortiz, M. Mukadam, NeuralFeels with neural fields: Visuotactile perception for in-hand manipulation. *Sci. Robot.* **9**, eadl0628 (2024).
53. K. Chen, Y. Ma, X. Lin, S. James, J. Zhou, Y.-H. Liu, P. Abbeel, Q. Dou, "Vision foundation model enables generalizable object pose estimation" in *Proceedings of Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang, Eds. (Curran Associates, 2024), vol. 37, pp. 19975–20002.
54. S. Umeyama, An eigen-decomposition approach to weighted graph matching problems. *IEEE Trans. Pattern Anal. Mach. Intell.* **10**, 695–703 (1988).
55. M. A. Fischler, R. C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **24**, 381–395 (1981).
56. A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, F. Yu, ShapeNet: An information-rich 3D model repository. arXiv:1512.03012 [cs.GR] (2015).
57. Y. Xiao, Y. Du, R. Marlet, "PoseContrast: Class-agnostic object viewpoint estimation in the wild with pose-aware contrastive learning" in *Proceedings of International Conference on 3D Vision* (IEEE, 2021), pp. 74–84.
58. G. Pitteri, M. Ramamonjisoa, S. Ilic, V. Lepetit, "On object symmetries and 6D pose estimation from images" in *Proceedings of International Conference on 3D Vision* (IEEE, 2019), pp. 614–622.
59. T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, L. Zhang, Grounded SAM: Assembling open-world models for diverse visual tasks. arXiv: 2401.14159 [cs.CV] (2024).
60. J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, D. Novotny, "VGGT: Visual geometry grounded transformer" in *Proceedings of the Computer Vision and Pattern Recognition Conference* (IEEE, 2025), pp. 5294–5306.

Acknowledgments

Funding: This work was supported in part by a grant from the National Natural Science Foundation of China (project no. 62322318), the InnoHK of the government of Hong Kong via the Hong Kong Centre for Logistics Robotics (HKCLR), a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (project no. 14208424), and the Joint Funds of the National Natural Science Foundation of China (project nos. U24A20128 and U25A6013). **Author contributions:** Q.D., Y.-H.L., P.A., and R.X. conceived the study and identified research objectives and goals. Q.D., Y.-H.L., P.A., S.J., and K.C. designed the work. Q.D. and K.C. managed the project timeline, coordination, and team communication. K.C., C.L., C.T., Y.M., J.P., and W.C. developed the methodology and conducted experiments with data analysis. Q.D., Y.-H.L., and R.X. provided equipment to support the evaluation of the proposed algorithm on multiple hardware platforms. W.C., Z.Z., and X.X. provided technical support for trajectory planning and robot control. K.C., Y.M., C.T., and C.L. designed the supplementary movies. K.C., J.P., Y.M., C.L., and Q.D. cowrote the initial manuscript, with all coauthors providing constructive comments and editing. Q.D., Y.-H.L., P.A., R.X., C.-W.F., and S.J. provided guidance and mentorship throughout this project. **Competing interests:** P.A. holds concurrent appointments as a professor at UC Berkeley and as an Amazon Scholar. This paper describes work performed at UC Berkeley and is not associated with Amazon. The other authors declare that they have no competing financial interests. **Data, code, and materials availability:** All data and code needed to evaluate the conclusions in the paper are available in the Zenodo repository at https://doi.org/10.5281/zenodo.19325674. No new materials were generated from this study.

Submitted 1 July 2025
Accepted 31 March 2026
Published 29 April 2026
10.1126/scirobotics.aea2092

Science Robotics

A retrieval-augmented framework enabling VLM spatial awareness for object-centric robot manipulation

Kai Chen, Chengkun Li, Chang Tu, Jiahui Pan, Yiyao Ma, Wei Chen, Zhongxiang Zhou, Xuecheng Xu, Stephen James, Chi-Wing Fu, Rong Xiong, Pieter Abbeel, Yun-Hui Liu, and Qi Dou

Sci. Robot. **11** (113), eaea2092. DOI: 10.1126/scirobotics.aea2092

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.