Embodied Spatial Affordance: Spatial-Aware Affordance Learning for Embodied Navigation and Manipulation

Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang

***Abstract*—Embodied agents must effectively navigate and manipulate objects within their environments to accomplish tasks. A key challenge in this process is understanding the spatial context and the affordances of the environment, which involves recognizing how object can be interacted with (object affordance) and identifying suitable location for movement and object placement (free space affordance). Despite the recent adoption of Vision-Language Models (VLMs) to control robot behavior, these models often struggle to translate reasoning outcomes into precise executable actions, focusing instead on high-level spatial question answering and task planning. To address this gap, we introduce the Embodied Spatial Affordance dataset, a comprehensive resource designed to enable embodied agents to reason about both object and free space affordances. This dual focus enhances agents' ability to navigate their environment, interact with specific objects, and identify appropriate locations for object placement. By integrating reasoning about object properties with spatial context, we aim to improve the robustness and versatility of embodied intelligence. Using the proposed ESA dataset, we develop a novel model, EspA, which predicts both object and free space affordances based on observed images and language instructions. The outputs of EspA are affordance keypoints providing actionable insights that facilitate real-time decision-making for embodied agents. Extensive experimental results demonstrate that EspA outperforms existing state-of-the-art Vision-Language Models (VLMs), both open-source and closed-source, in object and free space affordance prediction. Furthermore, it exhibits superior performance in real-world embodied navigation and manipulation experiments. We believe this work paves the way for more robust and versatile embodied agents capable of effectively interacting with complex environments. The dataset, benchmark, and evaluation code will be publicly available to facilitate future research. Project website: *https://embodied-spatial-affordance.github.io/*.**

***Index Terms*—Embodied Spatial Affordance, Spatial Reasoning, Embodied Navigation, Embodied Manipulation**

I. Introduction

Enabling embodied agents to navigate and manipulate objects within complex 3D environments poses a significant challenge in the field of embodied AI. Achieving this goal requires not only sophisticated perception and control mechanisms [1], [2] but also a profound understanding of spatial relationships and affordances—defined as the action possibilities presented by objects and spaces. Recent advancements in Vision-Language Models (VLMs) [3]–[8] have markedly enhanced robots' perceptual and reasoning capabilities, leading to impressive strides in various embodied tasks. These include vision-language navigation [9]–[12], embodied manipulation [13]–[15], and embodied reasoning [6], [7], [16]. However, a notable gap remains between high-level reasoning and precise action execution in current VLMs. While these models excel at abstract spatial question answering, they frequently overlook the actionable affordances essential for effective real-world interactions. This gap underscores the need for a more comprehensive understanding of spatial affordances to bridge the divide between reasoning and action in embodied agents.

Existing work seeks to address this gap by utilizing spatial reasoning through pointing, which translates language instructions into actionable targets, as shown by RoboPoint [17]. While this approach mitigates ambiguities inherent in language (*e.g.*, "next to the plate"), it still exhibits significant limitations. First, RoboPoint primarily focuses on free space references and lacks explicit modeling of object affordances that define *where* and *how* to interact with objects (*e.g.*, grasping a teapot by its handle). Second, the need for robots to comprehend complex spatial relationships and object references during navigation is largely overlooked in RoboPoint. These limitations arise from a deficiency in training data that explicitly captures the interplay between object affordance, free space affordance, and spatial reasoning. Consequently, robots struggle to translate high-level goals (*e.g.*, "bring me the book") into low-level actions (*e.g.*, navigating to the bookshelf, grasping the book, and returning it).

To address these limitations, this work introduces a comprehensive ***Embodied Spatial Affordance (ESA)*** dataset that explicitly integrates object affordances and free space affordances with spatial reasoning for embodied agents. As shown in Fig. 1, the ESA dataset empowers embodied agents with four essential capabilities: (1) *Object Affordances for Navigation*, enabling agents to navigate towards and interact with specific objects based on their affordances (*e.g.*, navigating to a trash can to dispose of waste); (2) *Free Space Affordances for Navigation*, allowing agents to identify navigable areas and avoid obstacles; (3) *Object Affordances for Manipulation*, guiding agents to recognize functional parts of objects for interaction (*e.g.*, grasping a teapot by its handle); and (4) *Free Space Affordances for Manipulation*, facilitating agents in finding suitable locations for placing objects (*e.g.*, positioning

Xiaoshuai Hao is with the Beijing Academy of Artificial Intelligence, Beijing, China (e-mail: xshao@baai.ac.cn).

Yingbo Tang is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China (e-mail: tangyingbo2020@ia.ac.cn).

Lingfeng Zhang is with the Shenzhen International Graduate School, Tsinghua University, Shenzhen, China (e-mail: lfzhang715@gmail.com).

Corresponding author: Yingbo Tang.

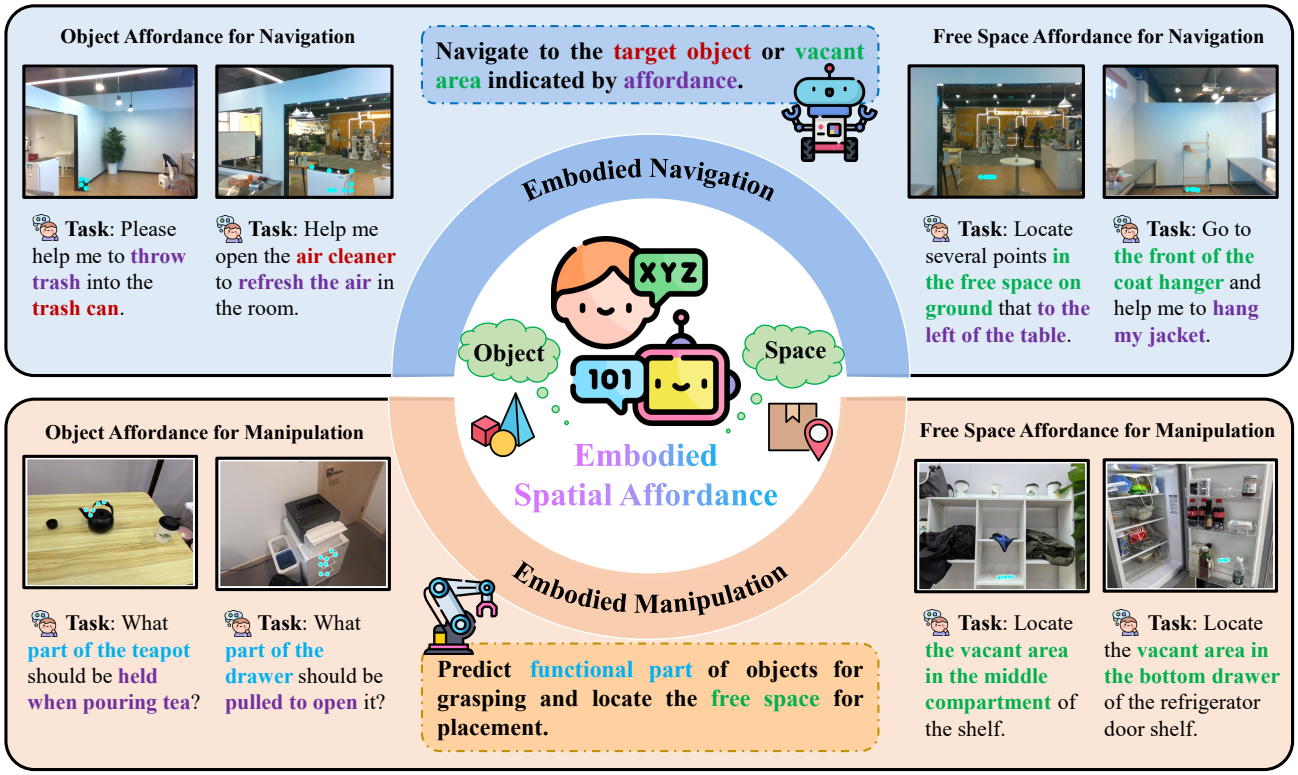

Fig. 1: Overview of the Embodied Spatial Affordance (ESA) dataset. The dataset is categorized into object affordances and free space affordances, supporting both navigation and manipulation tasks. In navigation tasks, the agent must reach a target object or a designated empty area; in manipulation tasks, the agent learns to identify functional parts for grasping and suitable spaces for placement. The ESA dataset integrates spatial reasoning with affordance understanding in an embodied context.

an item in the vacant space of a refrigerator). Building upon the ESA dataset, we propose **EspA**, a model that learns these affordances as a unified point representation, effectively bridging high-level planning (*e.g.*, "fetch the book") with low-level execution (*e.g.*, navigating to the bookshelf, grasping the book, and returning). By jointly reasoning about *object-centric* and *space-centric* affordances, EspA facilitates robust adaptation to unseen environments and diverse object configurations, effectively overcoming the limitations of conventional pointing paradigms. This capability is validated through rigorous benchmarks in navigation and manipulation tasks.

Our contributions are summarized as follows:

- We introduce the ***Embodied Spatial Affordance (ESA)*** dataset, a novel resource designed to enable embodied agents to reason about both object and free space affordances. ESA provides fine-grained 2D point annotations, allowing agents to understand these affordances within a unified spatial context.
- We propose ***EspA***, an innovative model that utilizes the ESA dataset to predict both object and free space affordances from observed images and language instructions. EspA integrates affordance prediction with spatial reasoning, generating keypoints that offer actionable insights for downstream navigation and manipulation tasks.
- Extensive experiments demonstrate that ***EspA*** significantly outperforms existing state-of-the-art Vision-Language Models (VLMs) in object and free space affordance prediction on our constructed benchmark. Additionally, EspA shows superior performance in embodied navigation and manipulation, highlighting its effectiveness in real-world applications.

The rest of this paper is organized as follows. We briefly review some related works in Section II. In Section III, we introduce our proposed Embodied Spatial Affordance dataset. We then give a detailed explanation of our method in section IV. A variety of experimental results are presented in Section V. Finally, Section VI concludes this paper.

II. Related Work

Datasets for Affordance Learning Affordance learning has become a pivotal area in computer vision and robotics, focusing on understanding potential interactions enabled by objects and environments. For instance, a chair affords "sitting", while a cup affords "grasping" or "pouring". This functional understanding is crucial for intelligent agents to interact naturally with the physical world. To support this goal, various datasets have been developed for training and evaluating affordance models. Early works concentrated on image-based functional part recognition, utilizing semantic segmentation for object region annotations. The UMD dataset [18] provided one of the first pixel-level annotations for common household tools. The IIT-AFF dataset [19] expanded the diversity of object categories and affordance types. Subsequent datasets like

TABLE I: Comparison of existing affordance datasets.

Dataset	Domain			Tasks		Affordance Type		#Spatial	#Images	#QAs
	Generic	Indoor	Tabletop	Navigation	Manipulation	Object Affordance	Free Space Affordance	Relations		
UMD [18]	✗	✗	✓	✗	✓	✓	✓	✗	30K	-
IIT-AFF [19]	✓	✗	✗	✗	✓	✓	✓	✗	8.8K	-
PAD [20]	✓	✗	✗	✗	✓	✓	✓	✗	4K	-
AGD-20K [21]	✓	✗	✗	✗	✓	✓	✓	✗	26.1K	-
3DOI [22]	✓	✓	✗	✗	✓	✓	✓	✗	10K	-
RoboPoint [17]	✓	✗	✓	✓	✓	✓	✗	✓	-	1.4M
RoboSpatial [23]	✗	✓	✓	✗	✓	✗	✗	✓	1.0M	3.0M
ESA (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	**870K**	**2.0M**

PAD [20], [24] and AGD-20K [21] enhanced annotation quality and coverage, introducing richer interaction types such as "hold", "open", and "cut". The 3DOI dataset [22] introduced multi-scale functional annotations, capturing affordances at various granularities. Beyond static images, some approaches explored video-based affordance transfer, using videos of human-object interactions to generate heatmaps for target images [25], [26]. These techniques leverage temporal cues to model affordances in dynamic settings. Recent efforts have focused on 3D affordance modeling, with datasets like [27]–[29] providing 3D annotations that model affordances as contact regions or trajectories, enhancing learning for applications like grasp planning. Despite these advances, existing datasets are often limited in scope, either being object-centric with detailed functional part annotations or scene-centric, focusing on global interaction contexts. This disconnect poses challenges for real-world learning scenarios where both object details and spatial context are vital. To address this gap, we introduce ***Embodied Spatial Affordance (ESA)***, a dataset that connects object-level and scene-level affordance understanding. ***ESA*** employs an *interactive point prediction* framework, combining precise localization of functional parts of objects with language-guided spatial awareness in real-world contexts. By integrating vision, language, and interaction intent, ***ESA*** serves as a comprehensive benchmark for multimodal affordance learning, enhancing intelligent, context-aware robotic behavior.

Spatial Reasoning with VLMs Spatial reasoning is crucial for robots to interact effectively with the physical world. To enhance the capabilities of 2D Vision-Language Models (VLMs), several works have extracted 3D spatial information from 2D images to generate large-scale, spatially-related question answering pairs for model fine-tuning. For example, SpatialVLM [30] converts 2D images into object-centric 3D point clouds using metric depth estimation, synthesizing VQA data with 3D spatial reasoning supervision. SpatialRGPT [31] enhances region-level spatial reasoning through region proposals and 3D scene graph construction. RoboPoint [17] introduces a synthetic dataset for free space references, predicting points for precise action spaces. Other studies, such as SpatialBot [32] and RoboSpatial [23], leverage both RGB and depth images to provide VLMs with comprehensive spatial information, leading to improved spatial understanding and reasoning. As large models advance, recent works have focused on optimizing the reasoning process to enhance spatial understanding. For instance, SpatialCoT [33] utilizes bidirectional alignment of spatial coordinates and language, significantly bolstering the spatial reasoning capabilities of VLMs. VILASR [34] introduces a drawing-to-reason paradigm where VLMs perform spatial reasoning through visual annotations, its progressive training framework systematically cultivate visual reasoning capabilities. Additionally, MetaSpatial [35] incorporates a multi-turn reinforcement learning-based optimization framework with physics-aware constraints to enhance 3D spatial reasoning. Despite advancements, existing VLMs face limitations in spatial reasoning for real-world interactions. Their inadequate coverage of object functionality and spatial relations hampers accurate predictions of contact positions for affordances. When faced with unseen settings or objects, VLMs often struggle to transfer spatial knowledge from training, resulting in suboptimal performance. This paper explores methods to enhance VLMs' spatial reasoning capabilities, aiming to improve robotic interactions in real-world scenarios.

III. Embodied Spatial Affordance Dataset Construction

A. Data collection and Filtering

Our Embodied Spatial Affordance dataset integrates multimodal data from both real-world and synthetic sources, systematically addressing *Object Affordance* for interaction possibilities and *Free Space Affordance* for navigable and actionable regions. This comprehensive approach ensures robust coverage of various interaction scenarios and spatial contexts.

Object Affordance The data for object affordance learning is sourced from five datasets: LVIS [37], Pixmo-Points [38], Object Reference [17], PACO-LVIS [39], and our own NaviAfford. LVIS [37] provides 152K images with 2.2M high-quality instance segmentation masks across over 1,000 categories, which we convert into an object detection format with bounding box coordinates (x_1, y_1, x_2, y_2) to establish general object detection capabilities. For precise object pointing, we utilize the Pixmo-Points dataset along with 288K synthetic images from RoboPoint [17]. Given the densely repeated instances in Pixmo-Points, we implement a two-step filtering process: first, we discard annotations with more than 10 point labels for training simplicity; second, we apply GPT-4o [36] to retain only relevant indoor objects (*e.g.*, furniture, kitchenware), resulting in 63,907 images suitable for object pointing.

For manipulation scenarios requiring fine-grained part-level affordance prediction, we utilize the PACO-LVIS dataset [39], which contains bounding boxes and part segmentation masks for 45,790 images across 75 object categories and 200 part categories. We convert these ground truth labels into point annotations by sampling points within the masks to indicate object affordance. For navigation scenarios, we collected the

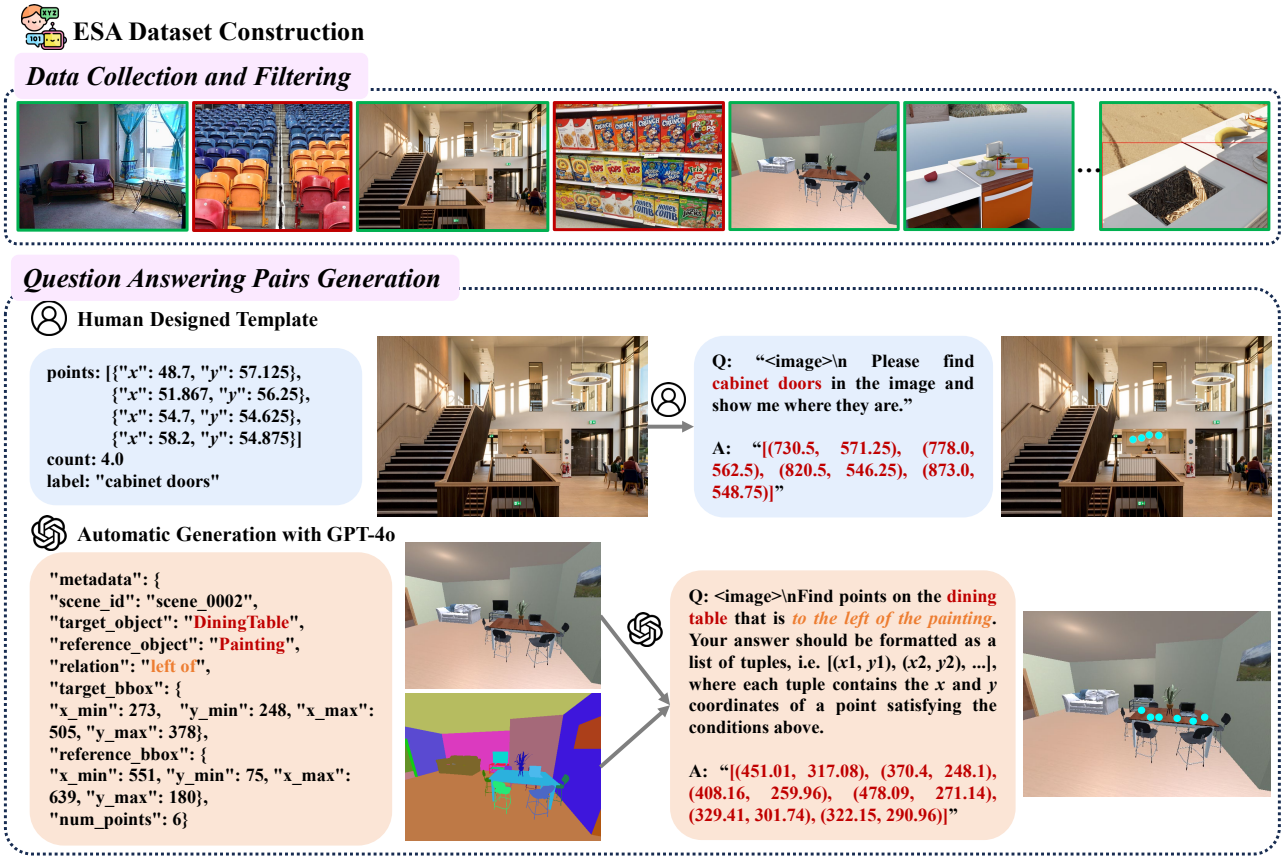

Fig. 2: Pipeline for constructing the ESA dataset. We begin by excluding images with densely repeated objects to ensure quality. Next, we generate question-answering pairs using either human-designed templates or the GPT-4o model [36], facilitating diverse and contextually relevant interactions.

NaviAfford dataset, comprising 50K ego-perspective images from 200 indoor scenes in the AI2Thor simulator [40]. Our collection pipeline randomly samples reachable locations, filters out areas with insufficient clearance (¿1.5m), and captures RGB images with instance segmentation masks from multiple viewpoints (0-360° random rotation, -15° to 15° horizontal rotation). Each image includes metadata such as visible object bounding boxes, 3D distances, and 2D coordinates. From these images, we extract spatial relationship annotations by identifying object pairs that meet specific spatial constraints (*e.g.*, left/right spacing ¿20 pixels). For each relation, we generate 4-8 pointing coordinates within the target object's bounding box, resulting in annotations formatted as "locate several points on target relation reference," yielding 50K object affordance samples for navigation training.

Free Space Affordance For free space affordance learning, we utilize the Region Reference dataset from [17], which encompasses 270K images across 8K instances and 262 categories. Each image features one or two colored bounding boxes to indicate relevant objects, with ground truths represented as series of points $[(x_1, y_1), (x_2, y_2), ...]$ for free space referencing. To enhance optimization efficiency, we sample these ground truth coordinates, limiting them to a maximum of ten points per annotation. This approach maintains consistent annotation density while ensuring spatial precision, facilitating effective learning of free space affordances. By incorporating diverse scenes and object types, the dataset enables robust training for models designed to navigate and interact within complex environments.

B. Question Answering Pairs Generation

We develop a question-answer generation pipeline for the ESA dataset, as shown in Fig. 2. By transforming the data into affordance-aware QAs, we enhance the engagement of vision-language models (VLMs) with the dataset, enabling them to learn and infer spatial relationships between objects and affordances.

Object Affordance QA Generation For object affordance task, we first use GPT-4o [36] to analyze scenes and filter out irrelevant outdoor images. We generate questions and answers using human-designed templates for Pixmo-Points [38], such as "Point to all occurrences of `<label>` in the image" and "Can you see any `<label>` in the image? Point to them", where "`<label>`" refers to ground truths from Pixmo-Points. We design 28 templates and randomly select one for each object pointing question. For the PACO-LVIS dataset [39] and our NaviAfford dataset, we generate QAs by prompting GPT-4o with images, metadata, and ground truth masks. Questions for the PACO-LVIS dataset focus on functional use without naming the object (*e.g.*, "What appliance can heat food

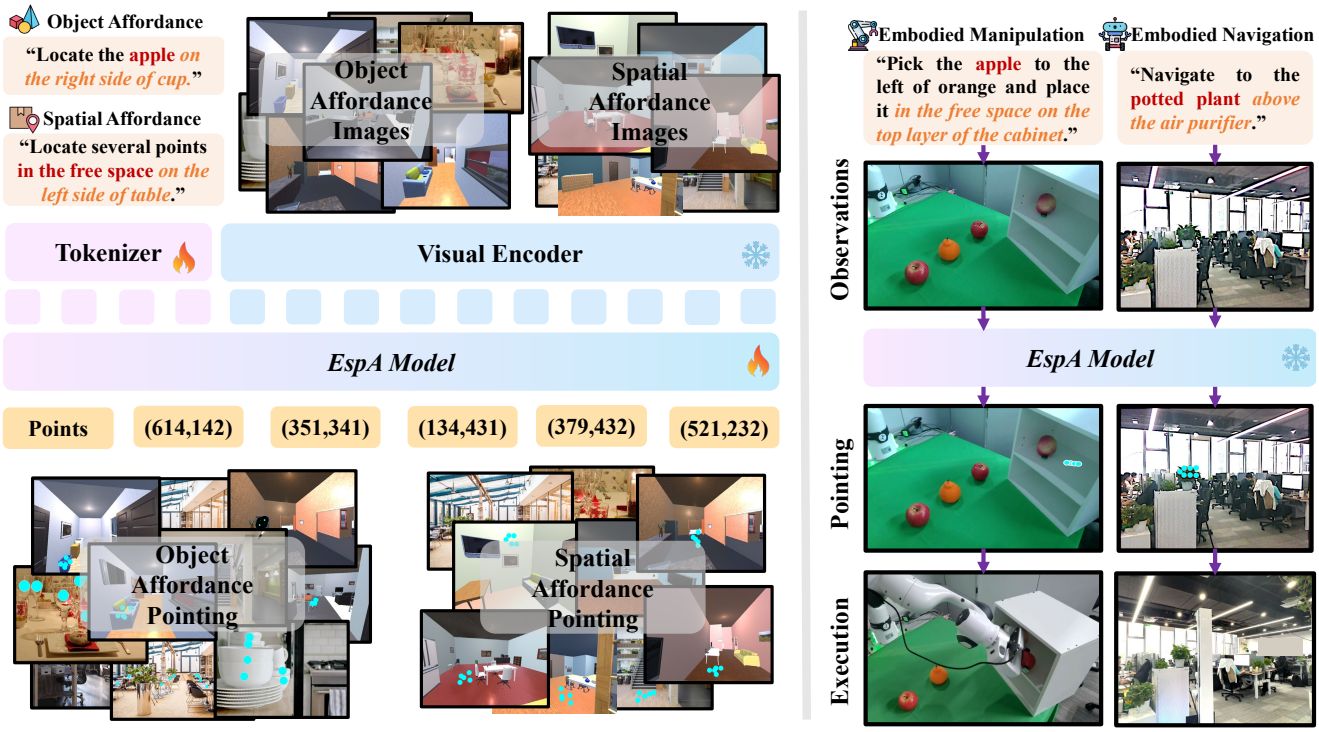

Fig. 3: Framework of EspA. We fine-tune a multimodal language model on the ESA dataset to enhance object and free space affordance capabilities. For downstream embodied navigation and manipulation tasks, we integrate depth images to convert 2D points representing affordances into 3D coordinates, which are then used as target positions for navigation and manipulation.

quickly?" for a microwave) and for object parts, we ask for part identification (*e.g.*, "Which part of a knife should be held to cut safely?" for the handle). For NaviAfford, we generate questions for referred objects based on spatial relations from the metadata. Ground truth answers include two formats: (1) bounding boxes for the target object or part, and (2) points sampled from the ground truth segmentation mask. This dual representation ensures accurate part grounding and enhances point-level object affordance prediction.

Free Space Affordance QA Generation For free space affordance learning, we develop a systematic QA generation process that converts raw spatial annotations into a unified format. We generate QAs by modifying annotations from RoboPoint [17], transforming each spatial relationship into contextual question-answer pairs. The process includes two key steps: first, we convert normalized coordinates to absolute positions to maintain accurate real-world scales and spatial relationships. Second, we create diverse question templates to capture various spatial reasoning scenarios, ensuring comprehensive coverage of free space affordance understanding. Ground-truth points are resampled to a maximum of ten per question for consistency across different spatial complexities, with instructions adjusted for answer format uniformity. This strategy balances computational efficiency with spatial precision. The conversion preserves the spatial relationships defined in RoboPoint while integrating them into our unified affordance framework, effectively capturing and enabling spatial reasoning for vision-language models.

C. ESA-Eval Benchmark

To evaluate object affordance, we manually annotated 114 questions for spatial-awareness, focusing on spatial relations among target and reference objects, and 124 questions for functional-awareness, emphasizing object functionality using images from the Where2Place dataset [17]. For free space affordances, we retained the original 100 questions, adjusting the prompts to shift predictions from normalized to absolute coordinates. Ground truth for each question consists of one or more human-annotated polygon masks corresponding to the parts or instances in the answers.

For each predicted point, we check if it falls within the ground truth masks. The accuracy for a question is the ratio of correctly located points to total predicted points, with overall accuracy being the average across all questions. To enforce stricter criteria, we penalize points outside the image boundaries, encouraging the model to learn absolute interactive positions more effectively.

IV. Methodology

A. Framework

Based on the ESA dataset, we introduce the EspA model, a vision-language model specifically designed for Embodied Spatial Affordance, featuring a multimodal architecture that includes a Vision Transformer (ViT) [41] as the vision encoder, an MLP projector, a language tokenizer, and the Qwen2.5 LLM [42], as shown in Fig. 3. Given an input image

$I \in \mathbb{R}^{H\times W\times 3}$ and a textual instruction T, our objective is to predict spatial coordinates via the mapping:

$$EspA : (I, T) \rightarrow \mathbf{P} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)\}, \tag{1}$$

where $\mathbf{P}$ represents the predicted 2D points corresponding to object and free space affordances.

The vision encoder $\mathcal{E}_v$ transforms the input image into a sequence of visual tokens:

$$\mathbf{V} = \mathcal{E}_v(I) = \{v_1, v_2, \ldots, v_m\} \in \mathbb{R}^{m\times d_v}, \tag{2}$$

where m is the number of visual patches and d_v is the visual feature dimension. Simultaneously, the textual instruction is tokenized and embedded using the language tokenizer:

$$\mathbf{E}_t = \mathcal{E}_l(T) = \{e_1, e_2, \ldots, e_n\} \in \mathbb{R}^{n\times d_l}, \tag{3}$$

with d_l as the language embedding dimension. To align visual and textual features in a shared embedding space, we employ an MLP projector $\mathcal{P}$ defined as:

$$\mathcal{P}(v_i) = \mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot v_i + \mathbf{b}_1) + \mathbf{b}_2, \tag{4}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_h\times d_v}$, $\mathbf{W}_2 \in \mathbb{R}^{d_l\times d_h}$, and biases $\mathbf{b}_1 \in \mathbb{R}^{d_h}$, $\mathbf{b}_2 \in \mathbb{R}^{d_l}$, with d_h as the hidden dimension, yielding aligned visual features:

$$\mathbf{V}' = \mathcal{P}(\mathbf{V}) \in \mathbb{R}^{m\times d_l}. \tag{5}$$

The aligned visual features are concatenated with the textual embeddings to form a unified multimodal representation:

$$\mathbf{X} = [\mathbf{V}'; \mathbf{E}_t] \in \mathbb{R}^{(m+n)\times d_l}, \tag{6}$$

which is then fed into the Qwen2.5 LLM [42] for joint spatial affordance reasoning:

$$\mathbf{H} = \text{Qwen2.5-LLM}(\mathbf{X}) \in \mathbb{R}^{(m+n)\times d_l}. \tag{7}$$

The final hidden states $\mathbf{H}$ are decoded to generate textual representations of spatial coordinates:

$$\hat{\mathbf{P}} = \text{Qwen2.5-LLM}_{\text{decode}}(\mathbf{H}), \tag{8}$$

resulting in natural language descriptions of point coordinates in the form $\mathbf{P} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)\}$. This architecture underwent supervised fine-tuning to enhance the model's spatial understanding, followed by evaluation on various downstream tasks.

B. Instruction Fine-tuning

We employ a multi-stage training strategy based on the LLaVA-1.5 instruction tuning framework [43], consisting of two phases: General Localization Learning and Embodied Spatial Affordance Enhancement. The first phase uses the LVIS [37] and Pixmo-Points [38] datasets, totaling 216K images and 703K QAs, to provide the model with fundamental object detection and pointing capabilities. The second phase includes Object Reference [17], PACO-LVIS [39], NaviAfford, and Region Reference [17], totaling 654K images and 1.33M QA pairs, which enhance object affordance prediction and free space localization. During training, we use a standard autoregressive language modeling loss for the coordinate tokens, given by:

$$\mathcal{L} = -\sum_{i=1}^{L} \log P(y_i | y_{<i}, \mathbf{X}), \tag{9}$$

where L is the length of the target sequence of spatial coordinates, y_i is the i-th token in the sequence, and $P(y_i|y_{<i}, \mathbf{X})$ represents the probability of generating token y_i given the preceding tokens and the multimodal input $\mathbf{X}$.

This multi-stage fine-tuning approach, leveraging diverse datasets, enables the model to develop hierarchical affordance reasoning, progressing from basic object grounding to advanced object and free space affordance tasks.

C. EspA for Embodied Navigation and Manipulation

Fig. 3 illustrates the effective application of the fine-tuned EspA model to downstream embodied navigation and manipulation tasks.

Embodied Navigation Given an instruction like "Navigate to the potted plant above the air purifier," the EspA model accurately localizes the target object through spatial reasoning. The predicted 2D coordinates (x_{2d}, y_{2d}) are then transformed into 3D world coordinates using the depth image D and camera intrinsics:

$$\begin{bmatrix} x_{3d} \\ y_{3d} \\ z_{3d} \end{bmatrix} = D(x_{2d}, y_{2d}) \begin{bmatrix} \frac{x_{2d}-c_x}{f_x} \\ \frac{y_{2d}-c_y}{f_y} \\ 1 \end{bmatrix}, \tag{10}$$

where f_x and f_y are the focal lengths, and (c_x, c_y) is the principal point of the image. The 3D coordinates are then converted to the robot's local frame using a transformation matrix $\mathbf{T}_{camera}^{robot}$, allowing the robot to navigate to the designated target position.

Embodied Manipulation For instructions like "Pick the apple to the left of the orange and place it in the free space on the top layer of the cabinet," EspA predicts the affordances for the specified apple and the cabinet's free space. The predicted 2D points are translated into 3D coordinates using depth-based projection, while object affordance is processed via SAM2 [44] and AnyGrasp [45] for grasping pose estimation:

$$\begin{cases} (x_{2d}, y_{2d}) \rightarrow (x_{3d}, y_{3d}, z_{3d}), & \text{if free space,} \\ (M, D) \rightarrow (x_{3d}, y_{3d}, z_{3d}, r, p, y, w), & \text{otherwise,} \end{cases} \tag{11}$$

where M denotes the 2D mask of object r, p, y represent the roll, pitch, and yaw angles for the gripper orientation, and w indicates the opening width. This pipeline transforms visual affordance predictions into actionable 6D pose commands, enabling precise manipulation tasks.

V. Experiments

A. Experimental Setup

Dataset and Benchmark We train our EspA model on the Embodied Spatial Affordance (ESA) dataset, a comprehensive multi-modal collection of real and synthetic data for object and free space affordance. We evaluate the model against

TABLE II: Comparison results of various VLMs on ESA-Eval benchmark.

Type	Models	Parameters	Object Affordance		Free Space Affordance↑	Average↑
			Spatial-Awareness↑	Functional-Awareness↑		
Closed-source Models	GPT-4o [36]	-	21.2	15.9	25.4	20.5
	Claude-3.5-Sonnet [48]	-	20.4	13.1	22.6	18.4
	Gemini-2.5-Flash [49]	-	20.4	21.7	29.4	23.5
	Gemini-2.5-Pro [50]	-	17.0	14.5	41.8	23.4
Open-source Models	Qwen2.5-VL [42]	3B	7.1	2.0	12.5	6.8
	Molmo [38]	7B	5.7	4.7	4.7	5.0
	Qwen2-VL [51]	7B	15.6	10.5	14.3	13.4
	LLaVA-Next [52]	8B	2.9	0.8	0.6	1.4
	SpaceMantis [30]	8B	3.6	4.8	12.0	6.5
	RoboPoint [17]	13B	55.7	35.0	44.2	44.7
	Qwen2.5-VL [42] (Baseline)	7B	19.5	8.3	21.9	16.1
	EspA (Ours)	**7B**	**70.5 (+51.0↑)**	**63.1 (+54.8↑)**	**55.8 (+33.9↑)**	**63.4 (+47.3↑)**

several state-of-the-art VLMs using the ESA-Eval benchmark, focusing on three abilities:

- **Object affordance with spatial awareness:** Assesses reasoning about spatial relationships between objects.
- **Object Affordance with Functional Awareness:** Assesses the model's understanding of object functionalities and interactions at the part level.
- **Free space affordance:** Measures the prediction of valid vacant areas for movement and placement.

Evaluation Metrics For the ESA-Eval benchmark, we use accuracy (Acc), defined as the ratio of correctly predicted points within ground truth masks to the total predicted points. For real-world navigation and manipulation, we adopt the success rate (SR), defined as the ratio of successful executions to the total attempts.

Implementation Details Our EspA model is initialized with pre-trained Qwen2.5-VL-7B-Instruct [46] weights and undergoes full-parameter supervised fine-tuning as outlined in [47]. Experiments are conducted on four H100 GPUs, utilizing AdamW as the optimizer with a learning rate of 10^{-5} over one epoch. Each GPU processes a batch size of 4, with gradient accumulation set to 2 steps.

B. Comparison Results of State-of-the-Art Models

Baseline Models We evaluate several state-of-the-art vision-language models (VLMs) using the ESA-Eval benchmark, encompassing both closed-source and open-source models. The closed-source models include GPT-4o [36], Claude-3.5-Sonnet [48], Gemini-2.5-Flash [49], and Gemini-2.5-Pro [50]. On the other hand, the open-source models feature general-purpose VLMs such as LLaVA-Next [52], Molmo [38], Qwen2-VL [51], and Qwen2.5-VL [42]. Additionally, we assess two models specifically designed for spatial awareness: SpaceMantis, a community implementation of SpatialVLM [30], and RoboPoint [17]. This diverse selection allows for a comprehensive evaluation of model performance across various capabilities.

Tab. II presents a quantitative comparison of baseline models on the ESA-Eval benchmark. Generic VLMs exhibit limited zero-shot generalization in object and free space affordance tasks, with top performers like Gemini-2.5-Flash and Gemini-2.5-Pro achieving average accuracies of only 23.5 and 23.4, respectively. In contrast, specialized models with spatial reasoning, such as RoboPoint, attain an average accuracy of 44.7, underscoring the significance of spatial reasoning in these tasks. Our EspA model, fine-tuned on the ESA dataset, outperforms all baseline models, achieving accuracies of 70.5 and 63.1 for object affordance with spatial and functional awareness, respectively, and 55.8 for free space affordance, resulting in an average accuracy of 63.4. Compared to the baseline Qwen2.5-VL-7B [42], our model improves object affordance accuracies by 51.0 and 54.8 for spatial and functional awareness, respectively, and achieves a 33.9-point gain in free space affordance. Furthermore, it surpasses RoboPoint [17] by 18.7 points in average accuracy, demonstrating the effectiveness of the ESA dataset in enhancing embodied spatial affordance understanding.

C. Ablation Study

We conduct an ablation study on the ESA dataset, as shown in Tab. III, to evaluate the contribution of each data component to affordance understanding. The full configuration "All" demonstrates optimal performance by integrating LVIS [37] for object detection, Pixmo-Points [38] for object pointing, Object Reference [17], PACO-LVIS [39], and our NaviAfford dataset for object affordance, alongside Region Reference [17] for free space affordance learning. Notably, the removal of any single component leads to significant performance degradation, with reductions in spatial awareness and functional awareness accuracy, underscoring the critical role of each dataset. For instance, excluding NaviAfford results in an average accuracy drop to 57.9, highlighting its importance. Overall, the study confirms that a diverse dataset composition is essential for enhancing the model's capability in embodied spatial affordance understanding.

D. EspA for Downstream Embodied Tasks

Embodied Navigation We evaluate the performance of the EspA model in downstream embodied navigation tasks through four targeted scenarios, as shown in Fig. 4. These tasks assess EspA's ability to integrate affordance learning with path planning for goal-oriented navigation. In Task 1, the model successfully localizes a trash can and plans an obstacle-free path using object affordance predictions and real-time depth

TABLE III: Ablation on the data composition in ESA dataset.

Data Configuration	Object Affordance		Free Space Affordance↑	Average↑
	Spatial-Awareness↑	Functional-Awareness↑		
EspA (Ours)	70.5	**63.1**	55.8	**63.4**
w/o LVIS [37]	68.5	62.1	52.3	61.4
w/o Pixmo-Points [38]	69.0	53.6	57.0	59.8
w/o Object Reference [17]	65.3	62.4	51.7	60.2
w/o Region Reference [17]	69.0	60.3	20.1	51.4
w/o PACO-LVIS [39]	**71.0**	42.3	50.8	54.5
w/o NaviAfford	66.1	54.3	**57.9**	59.3

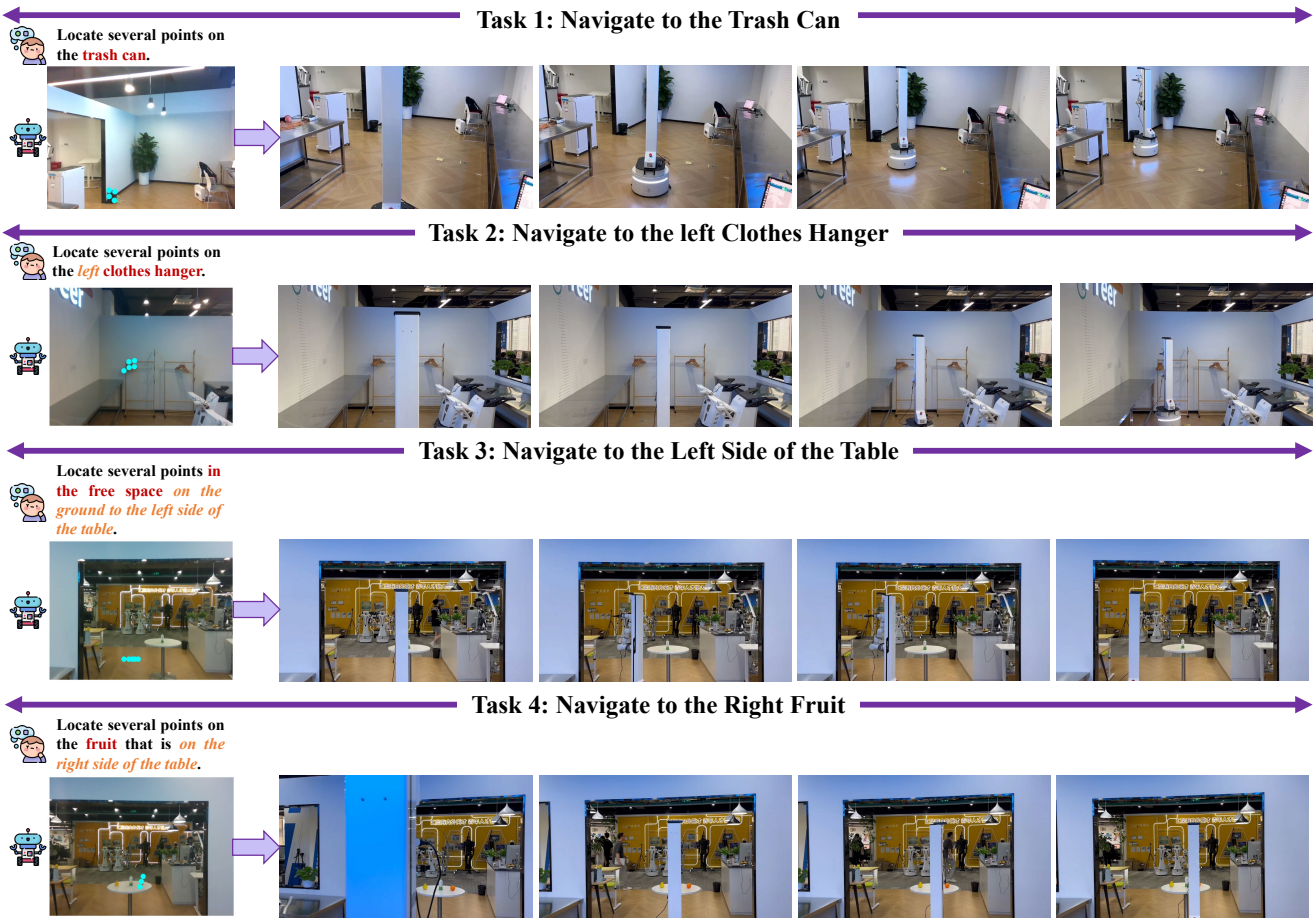

Fig. 4: Results of deploying EspA model to downstream robotic navigation tasks.

TABLE IV: Real-world navigation comparison of different methods.

Models	Trash Can	Potted Plant	Air Purifier	Fruit	Coffee Machine	Refrigerator	Free Space	Avg. SR↑
GPT-4o [36]	3/10	2/10	1/10	4/10	3/10	2/10	1/10	22.9%
Qwen2.5-VL-7B [42]	1/10	0/10	0/10	2/10	1/10	0/10	0/10	5.7%
RoboPoint [17]	4/10	3/10	2/10	6/10	4/10	3/10	2/10	34.3%
EspA (Ours)	**7/10**	**6/10**	**6/10**	**9/10**	**8/10**	**7/10**	**6/10**	**70.0% (+35.7%↑)**

perception. Task 2 showcases EspA's capability to resolve semantic ambiguities by navigating to the leftmost clothes hanger among multiple candidates. Tasks 3 and 4 highlight robust spatial reasoning, with the model identifying navigable regions next to a table and locating fruit on the right side.

Quantitative results in Tab. IV demonstrate EspA's significant advantages over state-of-the-art models, achieving a 70.0% average success rate (SR), markedly higher than GPT-4o [36] (22.9%) and RoboPoint [17] (34.3%). EspA excels in fruit localization (9/10 success rate) and maintains strong performance on complex targets like refrigerators (7/10) and free space navigation (6/10). The model's 35.7% improvement

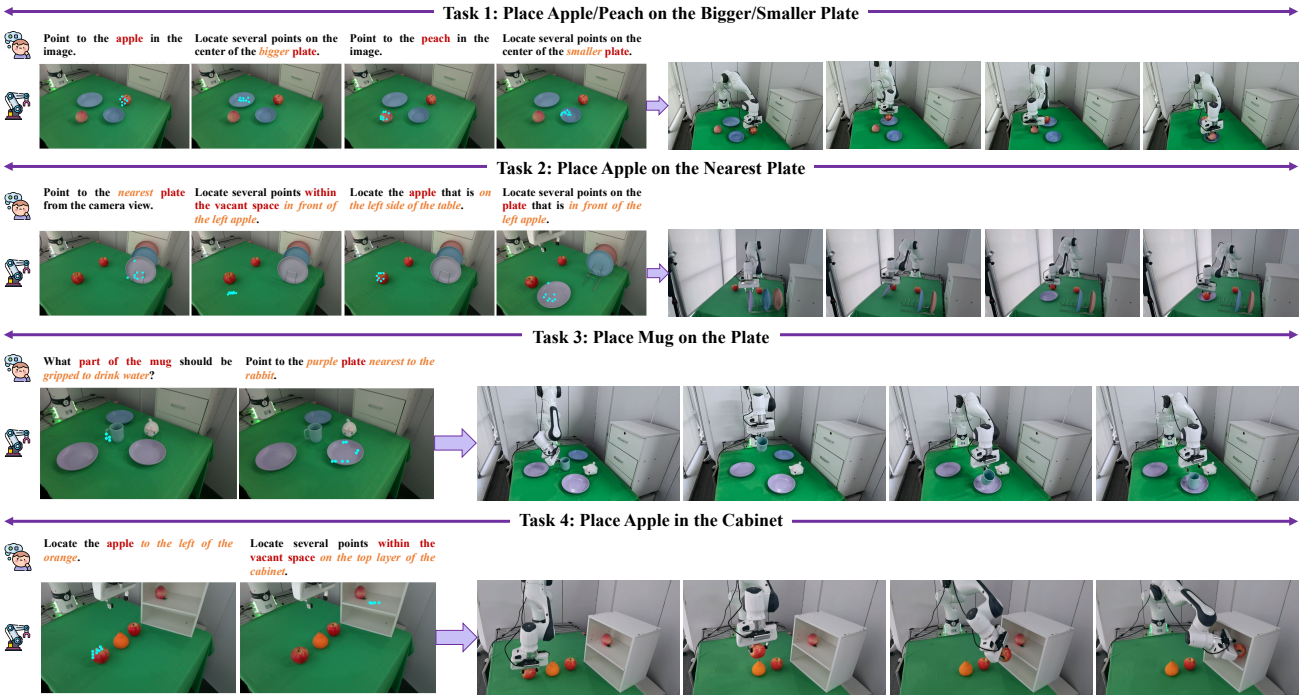

Fig. 5: Results of deploying EspA model to downstream robotic manipulation tasks.

TABLE V: Real-world manipulation comparison of different methods.

Model	Apple	Mug	Drawer	in Bowl	on Plate	Top-Cabinet	Bottom-Cabinet	Avg. SR↑
GPT-4o [36]	2/10	1/10	0/10	2/10	2/10	1/10	0/10	11.4%
Qwen2.5-VL-7B [42]	1/10	0/10	0/10	2/10	1/10	0/10	0/10	5.7%
RoboPoint [17]	5/10	2/10	1/10	6/10	6/10	2/10	3/10	35.7%
EspA (Ours)	**7/10**	**5/10**	**4/10**	**8/10**	**8/10**	**6/10**	**5/10**	**61.4% (+25.7%↑)**

over the best baseline underscores its strengths in object affordance understanding and free space reasoning, validating the effectiveness of the ESA dataset for training models in real-world navigation challenges.

Embodied Manipulation We rigorously evaluate the EspA model's manipulation capabilities through four hierarchical tasks in both tabletop and cabinet environments, as shown in Fig. 5. We mainly investigate the effectiveness of EspA in facilitating feasible pick-and-place manipulations. Tasks 1-2 assess basic object affordance understanding by distinguishing plate sizes and spatial distances for retrieving the nearest plate relative to the camera view. Task 3 showcases object affordance with function-awareness through mug grasping, and spatial awareness to find the nearest plate to the rabbit. Task 4 involves placing the specified apple within the cabinet, requiring understanding of the complex 3D structure with multiple layers. By identifying the optimal location on the top layer for apple placement, this case demonstrates the model's adaptability in complex spatial arrangements. These results collectively highlight EspA's advanced ability to combine object affordance prediction with spatial reasoning in real-world manipulation.

The quantitative comparison in Tab. V shows that EspA outperforms other models in real-world manipulation tasks, achieving an average success rate (SR) of 61.4%, significantly surpassing GPT-4o [36] (11.4%), Qwen2.5-VL-7B [42] (5.1%), and RoboPoint [17] (35.7%). EspA excels in both object affordance tasks (grasping an apple, mug, and opening a drawer) and free space tasks (placement in a bowl, on a plate, and in the cabinet). Specifically, it achieves high success rates in grasping an apple (7/10), a mug (8/10), and opening a drawer (4/10), demonstrating a robust understanding of object interactions. In free space tasks, EspA consistently performs well, placing objects in a bowl (8/10), on a plate (8/10), and in cabinet areas (6/10 for the top layer and 5/10 for the bottom), highlighting its advanced spatial reasoning capabilities. The 25.7% improvement over the best baseline, RoboPoint [17], emphasizes EspA's strengths in precise object manipulation and accurate spatial reasoning, validating its effectiveness for real-world robotic applications.

E. Visualization Results

Fig. 6 illustrates the qualitative results of the EspA model on the ESA-Eval benchmark, demonstrating its strong performance across various real-world manipulation tasks. The model effectively identifies object affordances, such as gripping a mug's handle, turning a faucet knob, and unscrewing a bottle cap, while also localizing free space affordances

Fig. 6: Qualitative results of EspA model, where cyan points indicate the object and free space affordances.

through contextual spatial reasoning. Notably, EspA adeptly manages complex spatial relationships, such as "bowl between the two mugs", "bread out of the plate", and "plate in the right sink", as shown in the first row of Fig. 6. This highlights its ability to generalize in complex scenarios with multiple objects and occlusions. The affordance points in the figure visually validate EspA's strengths in understanding object affordance with spatial and functional awareness, as well as free space prediction with spatial reasoning. These results underscore EspA's practical capabilities in real-world applications, merging accurate affordance understanding with spatial awareness.

VI. Conclusion

In this paper, we introduce ***ESA***, a large-scale dataset for object and free space affordance learning, comprising 2 million question-answer pairs with fine-grained 2D point annotations. This dataset enhances vision-language models (VLMs) by improving their reasoning about object affordance with spatial and functional awareness, while facilitating free space localization in context. Building on this, we propose **EspA**, a unified vision-language model for joint object and free space affordance reasoning, generating actionable affordance keypoints that bridge high-level instructions and executable actions. Experimental results confirm EspA's superior performance over state-of-the-art VLMs in the ESA-Eval benchmark and real-world tasks, demonstrating significant improvements in affordance prediction accuracy and task success rates. This work advances affordance-aware learning for robotic navigation and manipulation, effectively connecting high-level reasoning with low-level interactions. Future efforts will focus on extending EspA to dynamic environments and multi-agent collaboration, paving the way for next-generation embodied intelligence that adapts to ever-changing real-world conditions.

References

[1] D. Li, Y. Jin, Y. Sun, H. Yu, J. Shi, X. Hao, P. Hao, H. Liu, F. Sun, J. Zhang *et al.*, "What foundation models can bring for robot learning in manipulation: A survey," *arXiv preprint arXiv:2404.18201*, 2024.

[2] Y. Wu, H. Lyu, Y. Tang, L. Zhang, Z. Zhang, W. Zhou, and S. Hao, "Evaluating gpt-4o's embodied intelligence: A comprehensive empirical study," *TechRxiv preprint techrxiv.174495686.69962588/v1*, 2025.

[3] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang *et al.*, "Palm-e: An embodied multimodal language model," 2023.

[4] Y. Ji, H. Tan, J. Shi, X. Hao, Y. Zhang, H. Zhang, P. Wang, M. Zhao, Y. Mu, P. An *et al.*, "Robobrain: A unified brain model for robotic manipulation from abstract to concrete," *arXiv preprint arXiv:2502.21257*, 2025.

[5] B. R. Team, M. Cao, H. Tan, Y. Ji, M. Lin, Z. Li, Z. Cao, P. Wang, E. Zhou, Y. Han *et al.*, "Robobrain 2.0 technical report," *arXiv preprint arXiv:2507.02029*, 2025.

[6] G. R. Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl *et al.*, "Gemini robotics: Bringing ai into the physical world," *arXiv preprint arXiv:2503.20020*, 2025.

[7] H. Tan, Y. Ji, X. Hao, M. Lin, P. Wang, Z. Wang, and S. Zhang, "Reason-rft: Reinforcement fine-tuning for visual reasoning," *arXiv preprint arXiv:2503.20752*, 2025.

[8] S. Zhang, X. Hao, Y. Tang, L. Zhang, P. Wang, Z. Wang, H. Ma, and S. Zhang, "Video-cot: A comprehensive dataset for spatiotemporal understanding of videos based on chain-of-thought," *arXiv preprint arXiv:2506.08817*, 2025.

[9] G. Zhou, Y. Hong, and Q. Wu, "Navgpt: Explicit reasoning in vision-and-language navigation with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7641–7649.

[10] L. Zhang, X. Hao, Q. Xu, Q. Zhang, X. Zhang, P. Wang, J. Zhang, Z. Wang, S. Zhang, and R. Xu, "Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation," *arXiv preprint arXiv:2502.13451*, 2025.

[11] L. Zhang, Q. Zhang, H. Wang, E. Xiao, Z. Jiang, H. Chen, and R. Xu, "Trihelper: Zero-shot object navigation with dynamic assistance," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024, pp. 10 035–10 042.

[12] L. Zhang, H. Wang, E. Xiao, X. Zhang, Q. Zhang, Z. Jiang, and R. Xu, "Multi-floor zero-shot object navigation policy," *arXiv preprint arXiv:2409.10906*, 2024.

[13] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, "Robotic control via embodied chain-of-thought reasoning," *arXiv preprint arXiv:2407.08693*, 2024.

[14] R. Xu, J. Zhang, M. Guo, Y. Wen, H. Yang, M. Lin, J. Huang, Z. Li, K. Zhang, L. Wang *et al.*, "A0: An affordance-aware hierarchical model for general robotic manipulation," *arXiv preprint arXiv:2504.12636*, 2025.

[15] Y. Tang, S. Zhang, X. Hao, P. Wang, J. Wu, Z. Wang, and S. Zhang, "Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter," *arXiv preprint arXiv:2503.00778*, 2025.

[16] W. Zhang, M. Wang, G. Liu, X. Huixin, Y. Jiang, Y. Shen, G. Hou, Z. Zheng, H. Zhang, X. Li *et al.*, "Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks," *arXiv preprint arXiv:2503.21696*, 2025.

[17] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox, "Robopoint: A vision-language model for spatial affordance prediction for robotics," *arXiv preprint arXiv:2406.10721*, 2024.

[18] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos, "Affordance detection of tool parts from geometric features," in *IEEE International Conference on Robotics and Automation*, 2015, pp. 1374–1381.

[19] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Object-based affordances detection with convolutional neural networks and dense conditional random fields," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017, pp. 5908–5915.

[20] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "One-shot affordance detection," *arXiv preprint arXiv:2106.14747*, 2021.

[21] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. C. Tao, "Learning affordance grounding from exocentric images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2252–2261.

[22] S. Qian and D. F. Fouhey, "Understanding 3d object interaction from a single image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 753–21 763.

[23] C. H. Song, V. Blukis, J. Tremblay, S. Tyree, Y. Su, and S. Birchfield, "Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics," *arXiv preprint arXiv:2411.16537*, 2024.

[24] W. Zhai, H. Luo, J. Zhang, Y. Cao, and D. Tao, "One-shot object affordance detection in the wild," *International Journal of Computer Vision*, vol. 130, no. 10, pp. 2472–2500, 2022.

[25] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 778–13 790.

[26] J. Chen, D. Gao, K. Q. Lin, and M. Z. Shou, "Affordance grounding from demonstration video to target image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6799–6808.

[27] S. Deng, X. Xu, C. Wu, K. Chen, and K. Jia, "3d affordancenet: A benchmark for visual object affordance understanding," in *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1778–1787.

[28] Y. Li, N. Zhao, J. Xiao, C. Feng, X. Wang, and T.-s. Chua, "Laso: Language-guided affordance segmentation on 3d object," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 251–14 260.

[29] A. Delitzas, A. Takmaz, F. Tombari, R. Sumner, M. Pollefeys, and F. Engelmann, "Scenefun3d: fine-grained functionality and affordance understanding in 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 531–14 542.

[30] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia, "Spatialvlm: Endowing vision-language models with spatial reasoning capabilities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 455–14 465.

[31] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu, "Spatialrgpt: Grounded spatial reasoning in vision language models," *arXiv preprint arXiv:2406.01584*, 2024.

[32] W. Cai, I. Ponomarenko, J. Yuan, X. Li, W. Yang, H. Dong, and B. Zhao, "Spatialbot: Precise spatial understanding with vision language models," *arXiv preprint arXiv:2406.13642*, 2024.

[33] Y. Liu, D. Chi, S. Wu, Z. Zhang, Y. Hu, L. Zhang, Y. Zhang, S. Wu, T. Cao, G. Huang *et al.*, "Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning," *arXiv preprint arXiv:2501.10074*, 2025.

[34] J. Wu, J. Guan, K. Feng, Q. Liu, S. Wu, L. Wang, W. Wu, and T. Tan, "Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing," *arXiv preprint arXiv:2506.09965*, 2025.

[35] Z. Pan and H. Liu, "Metaspatial: Reinforcing 3d spatial reasoning in vlms for the metaverse," *arXiv preprint arXiv:2503.18470*, 2025.

[36] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[37] A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5356–5364.

[38] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini *et al.*, "Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models," *arXiv preprint arXiv:2409.17146*, 2024.

[39] V. Ramanathan, A. Kalia, V. Petrovic, Y. Wen, B. Zheng, B. Guo, R. Wang, A. Marquez, R. Kovvuri, A. Kadian *et al.*, "Paco: Parts and attributes of common objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7141–7151.

[40] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, M. Deitke, K. Ehsani, D. Gordon, Y. Zhu *et al.*, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[42] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[43] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 34 892–34 916, 2023.

[44] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[45] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, 2023.

[46] Q. Team, "Qwen2.5-vl," January 2025. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5-vl/

[47] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, and Y. Ma, "Llamafactory: Unified efficient fine-tuning of 100+ language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[48] Anthropic, "The claude 3 model family: Opus, sonnet, haiku," 2024.

[49] Google, "Gemini 2.5 flash," 2025, accessed: 2025-05-20. [Online]. Available: https://deepmind.google/models/gemini/flash/

[50] G. Gemini Team, "Gemini 2.5 pro," 2025, accessed: 2025-05-06. [Online]. Available: https://deepmind.google/models/gemini/pro/

[51] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[52] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 296–26 306.

Xiaoshuai Hao received his Ph.D. from the Institute of Information Engineering, Chinese Academy of Sciences, in 2023. He is currently a researcher at the Beijing Academy of Artificial Intelligence, specializing in embodied multimodal large models. His research interests encompass embodied intelligence, multimodal learning, and autonomous driving. Dr. Hao has published over 30 papers in top-tier journals and conferences, including TIP, Information Fusion, NeurIPS, ICLR, ICML, CVPR, ICCV, ECCV, ACL, AAAI, and ICRA. He has achieved outstanding results in international competitions, securing top-three placements at prestigious conferences like CVPR and ICCV. Additionally, he serves on the editorial board of Data Intelligence and is an organizer for the RoDGE Workshop at ICCV 2025 and The RoboSense Challenge at IROS 2025.

Yingbo Tang received the B.E. degree from North China Electric Power University, Beijing, China, in 2020. She is currently pursuing the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her research interests include computer vision and robotic manipulation.

Lingfeng Zhang received the B.S. degree in electronic information engineering from the Beijing Institute of Technology and MPhil degree in microelectronics from the Hong Kong University of Science and Technology (Guangzhou). He is currently working toward the Ph.D. degree with the Shenzhen International Graduate School, Tsinghua University. His research interests are embodied AI and vision-based navigation.