

Embodied Scene Rearrangement Planning

Canzhi Chen*, Zan Wang*, Siqu Zhu*, Qi Wu*, Yixuan Li*, and Wei Liang†

Abstract—This paper introduces Embodied Scene Rearrangement Planning (ESRP), a novel task requiring embodied agents to rearrange furniture in 3D scenes to match a target configuration using only egocentric observations and a top-down target layout. Unlike prior rearrangement tasks, ESRP precludes global state access and introduces mutual object occlusions, reflecting the practical constraints of real-world robotic deployment. These factors make aligning partial egocentric observations with the global target layout particularly challenging for long-horizon planning. To facilitate research, we present ESRP-Bench, a comprehensive benchmark built on OmniGibson featuring over 5,400 scene pairs and 8,200 objects. We define three multi-level metrics to evaluate rearrangement quality and provide four baselines: a hierarchical task-and-motion planning method, a vision-language-model-based method, and two learning-based approaches (IL and RL). Experimental results demonstrate that current methods struggle to complete the task efficiently, highlighting ESRP as a challenging frontier for embodied agents in scene understanding and long-horizon task planning. This work serves as a stepping stone toward deploying intelligent agents in real-world scenarios. Project page: <https://pie-lab.cn/ESRP/>.

Index Terms—Mobile Manipulation, Semantic Scene Understanding, Scene Rearrangement Planning, Embodied AI.

I. INTRODUCTION

EXISTING studies on rearrangement planning generally fall into two distinct categories. (i) The first category focuses on *tabletop rearrangement*, where robotic manipulators are used to reposition small objects within confined workspaces [1]–[4]. These methods are limited to tabletop-scale and overlook larger scene-level rearrangement, which introduces additional challenges such as long-horizon navigation and spatial reasoning within complex environments. (ii) Another line of research targets *scene-level object rearrangement*, but typically operates on a simplified 2D plane with access to global observations [5]–[8]. However, abstracting the environment to 2D significantly limits their applicability to embodied agents in realistic 3D settings, where spatial geometric constraints and partial observations critically affect the feasibility of rearrangement actions. These limitations highlight a substantial

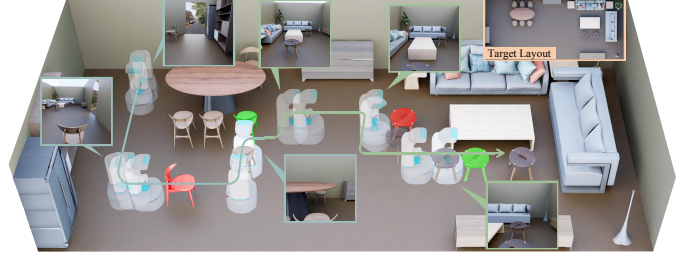


Fig. 1: **Overview of the ESRP task.** Given a top-down target layout, an embodied agent needs to rearrange furniture in a 3D scene from an initial configuration (red) to a target state (green). The agent operates under realistic constraints, relying only on egocentric observations without access to global state information or ground-truth localization.

gap between current research efforts and the practical demands of furniture rearrangement in real-world 3D scenes, such as domestic and hospitality environments.

To bridge these gaps, we introduce Embodied Scene Rearrangement Planning (ESRP), a novel task that requires embodied agents to rearrange a 3D scene from an initial layout to a target configuration, given only egocentric observations and a top-down target layout image, as shown in Fig. 1. Compared to prior works, ESRP introduces three distinct challenges:

- **Partial Observability:** Relying on egocentric observations significantly increases task difficulty. The policy must align partial, first-person views with the target layout and infer spatial information solely from local perception.
- **Complex Inter-Object Dependencies:** Manipulating large furniture items transforms the task into a long-horizon planning problem. The agent must generate plausible action sequences while reasoning about inter-object dependencies to avoid physical deadlocks and mutual occlusions.
- **Dynamic Scene Evolution:** The rearrangement process dynamically transforms the 3D scene, imposing heightened demands on the agent’s perception, localization, and navigation capabilities within a non-stationary environment.

Correspondingly, we introduce ESRP-Bench, a comprehensive benchmark built upon OmniGibson [9] and providing the systematic evaluation of ESRP. ESRP-Bench comprises over 5,400 fully furnished indoor scenes featuring more than 8,200 movable furniture items sourced from the 3D-FRONT dataset [10], covering diverse room types and furniture categories. Each scenario consists of three key components: (i) an initial scene layout configuration, (ii) a target layout represented as a top-down view image, and (iii) expert demonstration trajectories to support imitation learning. Furthermore, ESRP-Bench introduces three multi-level metrics for rigorous assessment and facilitates evaluation across varying difficulty levels, determined by the number of objects to be rearranged.

We develop four distinct baseline policies to tackle ESRP, categorized into three paradigms. (i) **Learning-based ap-**

Manuscript received: February 14, 2026; Revised: June 20, 2026; Accepted: July 24, 2026. This paper was recommended for publication by Editor Aniket Bera upon evaluation of the Associate Editor and Reviewers’ comments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62172043.

*C. Chen, Z. Wang, S. Zhu, Q. Wu, and Y. Li contributed equally.

† W. Liang is the corresponding author.

All authors are with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: chencanzhi@bit.edu.cn; wangzan@bit.edu.cn; zhuisiqi@gmail.com; wuqi@bit.edu.cn; liyixuan@gmail.com; liangwei@bit.edu.cn).

Digital Object Identifier (DOI): 10.1109/LRA.2026.3728329.

© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

proaches: We implement both an Imitation Learning (IL) and an Reinforcement Learning (RL) agent, which leverage egocentric observations and a top-down view of the goal layout. The RL baseline employs Proximal Policy Optimization (PPO) [11], while the IL baseline utilizes Behavior Cloning (BC) [12]. The PPO method adopts a vision encoder followed by a single-layer LSTM to process egocentric observations and maintain temporal memory, whereas BC adopts a Diffusion Policy [13] that predicts future action sequences through an iterative denoising process conditioned on visual observations and the gripper state. (ii) **Foundation-model-based approach:** We additionally introduce *ESRP-VLM*, a multimodal ReAct [14] agent built on top of a pretrained Vision-Language Model (VLM), which examines whether off-the-shelf generalist priors transfer to ESRP via in-context prompting alone. (iii) **Planning-based approach:** We additionally deploy a baseline that exploits privileged global information. The *ESRP-PLAN* baseline implements a Task and Motion Planning (TAMP) framework that hierarchically decomposes the problem: a task planner determines the optimal rearrangement order, while a motion planner computes collision-free paths for fetching and placing each object.

Despite employing diverse baseline models, experimental results indicate that performance remains suboptimal. For instance, even the strongest baseline achieves a success rate of only 30.20%, plummeting to nearly 0% in complex scenarios, underscoring the significant challenges inherent to ESRP.

We summarize our contributions as threefold: (i) We propose ESRP, a novel and practical task that requires embodied agents to rearrange furniture within 3D scenes using only egocentric observations. (ii) We introduce a comprehensive benchmark, ESRP-Bench, comprising over 5,400 initial-target scene layout pairs and three multi-level evaluation metrics that provide a rigorous assessment of rearrangement quality. (iii) We provide four baseline models as a first attempt to tackle ESRP, accompanied by comprehensive experimental results. These experiments reveal that different paradigms face distinct drawbacks in scene understanding and long-horizon task planning, highlighting key challenges and promising future research directions.

II. RELATED WORK

Rearrangement Planning Rearrangement entails reconfiguring objects from an initial state to a target layout [15]. A primary stream of research focuses on manipulating small objects, employing either planning-based methods [1], [2] or learning-based policies [4], [16]. However, these approaches are typically restricted to constrained workspaces and do not scale well to room-level tasks. Conversely, research addressing larger items, such as furniture [6], [8], [17], often relies on simplified 2D planar abstractions. This simplification overlooks intricate 3D spatial relationships and vertical geometric constraints, limiting the transferability of these methods to real-world deployment. To bridge this gap, we propose a task that requires rearranging furniture within fully 3D scenes using only egocentric perception, marking a significant step toward practical embodied AI applications.



Fig. 2: **Initial-target scene layout pairs.** We group scene layouts into three difficulty levels (as divided by green lines), with the first row showing the initial layout and the second row displaying the corresponding target layout. The top left shows **easy-level** scenes, where only 1 furniture object needs to be rearranged; the top right presents **medium-level** scenes with 2 ~ 3 objects to be rearranged; and the bottom shows **hard-level** scenes with 4 ~ 6 objects to be rearranged. The dataset covers a wide range of room types, including bedrooms, study rooms, and living rooms, along with diverse furniture types, such as chairs, tables, and plants.

Task and Motion Planning TAMP integrates high-level task planning with low-level motion planning, enabling robots to generate executable actions for complex, long-horizon tasks. Existing TAMP research has developed diverse approaches, including optimization-based, sampling-based [18], and learning-based methods [19]. However, ESRP presents unique challenges that exceed the capabilities of standard TAMP formulations. Specifically, it introduces two critical difficulties that strain existing planning methodologies: (i) Long-Horizon Task Complexity: Furniture rearrangement necessitates a sequence of interdependent actions over long horizons, surpassing the complexity of simpler manipulation tasks. (ii) 3D Spatial Reasoning Requirements: ESRP demands intricate 3D spatial understanding within the planning process, requiring precise object positioning, robust collision avoidance, and comprehensive reasoning about scene geometry. In this work, we systematically analyze the challenges posed by ESRP through the evaluation of these four baseline approaches.

III. ESRP-BENCH

A. Task Definition

ESRP defines a task in which an agent needs to reconfigure an unseen 3D indoor scene to match a target layout specified by a single top-down image. Relying solely on first-person views, the agent is required to: (i) comprehend the current scene state through egocentric observations, (ii) formulate an efficient rearrangement strategy, and (iii) execute a sequence of actions to precisely reposition furniture items. The task is deemed successful only when all rearrangeable objects are correctly placed at their designated goal positions.

Observations At each timestep, the ESRP agent receives an observation tuple comprising three components: (i) an egocentric RGB image I^{ego} (128×128) capturing the agent's current first-person view; (ii) a top-down RGB image I^g (128×128) depicting the target furniture layout; and (iii) a binary proprioceptive indicator $g \in \{0, 1\}$ denoting whether the agent is currently grasping an object.

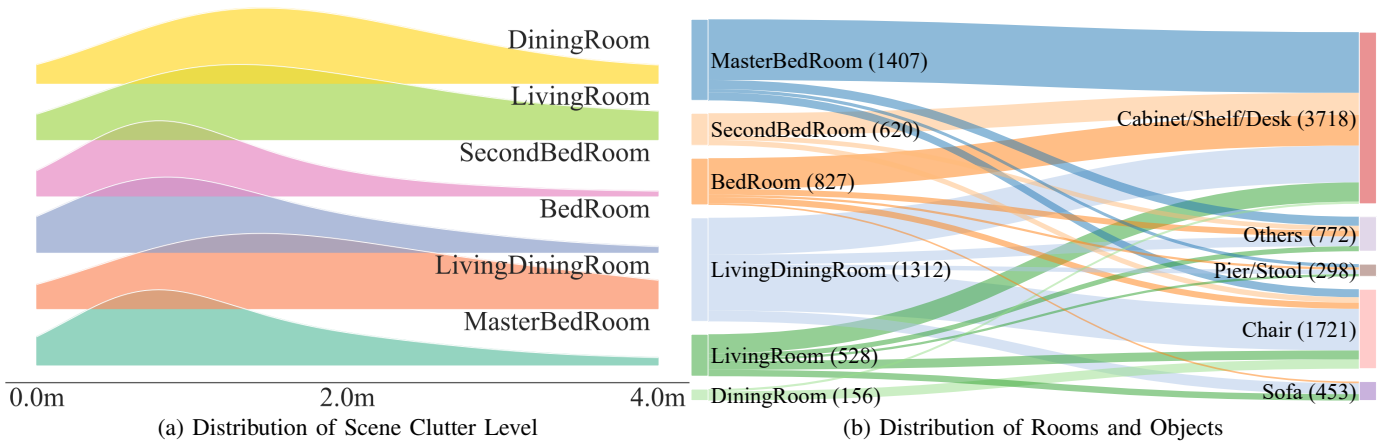


Fig. 3: **Dataset distribution.** (a) The scene clutter level in the dataset is defined as the sum of L2 distances between the initial and target positions of all rearrangeable objects. Formally, more challenging scenes have higher clutter levels. (b) Distribution of rearrangeable objects across different room types in the dataset. We report only the top six room categories in the dataset, excluding those that constitute only a small fraction. Notably, these top six categories account for over 80% of all rooms.

Actions We define the action space $\mathcal{A} = \{\text{move_forward}, \text{move_backward}, \text{turn_left}, \text{turn_right}, \text{fetch}, \text{release}\}$. The first four actions constitute the navigation primitives, while the latter two are object manipulation primitives. To balance planning efficiency with control precision, we implement discrete motion primitives: rotations are fixed in-place turns of 22.5° , and translations are defined as fixed steps of 0.25m.

Success Conditions We define episode success as the successful placement of all rearrangeable objects into their target configurations. Quantitatively, an object is considered correctly positioned if the Intersection over Union (IoU) between its current 2D bounding box and the target bounding box exceeds 0.3 on the horizontal plane. This threshold provides a robust geometric tolerance to accommodate minor control noise, while strictly ensuring that the final scene layout maintains semantic fidelity to the target.

Termination Conditions An episode terminates upon satisfying either of the following conditions: (i) the agent exhausts the maximum allowable timestep horizon L , or (ii) all rearrangeable objects are successfully positioned at their target locations. The former constitutes a timeout failure, while the latter signifies successful task completion.

B. Dataset

We introduce a comprehensive benchmark dataset tailored for ESRP, derived from the 3D-FRONT dataset [10]. This benchmark provides paired configurations with annotated initial and target layouts. In total, the dataset comprises 5,495 distinct scenes spanning 24 room types, with each scene containing between 1 and 6 rearrangeable objects from common furniture categories. Fig. 2 illustrates examples of these paired scenes across three difficulty levels: *Easy* (1 object), *Medium* (2–3 objects), and *Hard* (4–6 objects). This stratification ensures the dataset encompasses a broad range of indoor layouts, facilitating evaluation across varying degrees of spatial complexity and object interaction. Additionally, Fig. 3 presents two complementary views of the dataset statistics: the distribution of scene clutter levels and the frequency of rearrangeable objects across different room types.

Scene and Object Curation To ensure task feasibility, we applied several filtering criteria during dataset construction. First, we excluded rooms with insufficient navigational space for a furniture-carrying agent, preventing scenarios where physical constraints would make rearrangement impossible. Second, we selected furniture items with dimensions that allow effective agent manipulation, specifically choosing pieces slightly smaller than the agent to maintain maneuverability. Finally, to balance task complexity with practical solvability, we constrained each scene to contain at most 6 rearrangeable objects. These design choices create a challenging yet tractable benchmark that focuses on core rearrangement skills rather than navigation in extremely confined spaces.

Target Layout Acquisition We leverage the professionally designed layouts from the 3D-FRONT dataset [10] as the target layouts. We then capture 128×128 RGB images from a top-down view of these scenes to serve as the goal images. To represent the target layouts, we render 128×128 top-down orthographic RGB images, offering a complete spatial reference for agents. These images serve as a concise and comprehensive depiction of the desired furniture configuration, aiding agents in planning and executing their rearrangement tasks effectively.

Initial Layout Generation To ensure the rearrangement task is feasible, we generate the initial layout by perturbing the target layout through a rule-based simulation of object movements. Starting from the goal configuration, the robot iteratively moves each rearrangeable object, taking random steps forward/backward and turning, until it reaches a non-overlapping, collision-free position. This backward construction ensures, to the greatest extent possible, that a feasible solution exists between the initial and target layouts. Similar initialization strategies have been adopted in prior work [8], [17], but they typically ignore the robot’s embodiment, leading to object placements that are infeasible due to collisions during execution.

Expert Trajectories To facilitate the development of Imitation Learning approaches for ESRP, we use the reverse trajectories in the initial layout generation process as expert demonstrations. These trajectories capture rearrangement sequences executed by our rule-based agent, offering researchers high-quality behavioral data for supervised learning, behavioral cloning,

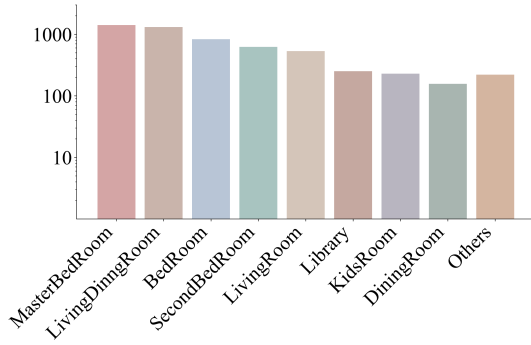


Fig. 4: **Distribution of room types.** The distribution of room categories is visualized on a logarithmic scale to account for class imbalance. Our dataset encompasses a diverse array of room types, comprehensively covering common household environments.

and other IL paradigms.

Statistics of Dataset We present comprehensive statistics of our proposed dataset. We analyze the distribution of room types and object categories, distinguishing between fixed furniture elements and rearrangeable objects to provide a detailed characterization of the dataset composition.

Fig. 4 illustrates the distribution of room types within the dataset, which encompasses 24 distinct categories, with 8 types represented by more than 100 instances each. Our dataset predominantly targets residential environments where furniture organization is most relevant; notably, *MasterBedroom* and *LivingDiningRoom* constitute the largest proportions. We further analyze the composition of movable furniture, identifying a total of 8,213 movable objects. These are categorized into: Cabinet/Shelf/Desk (3,718), Chair (1,721), Sofa (453), Pier/Stool (298), and Others (772).

C. Metrics

To quantitatively assess performance on the ESRP benchmark, we introduce a set of comprehensive evaluation metrics. Each method is evaluated across N independent episodes, with results aggregated by computing the mean performance for each metric. This approach ensures robust statistical assessment while capturing different aspects of rearrangement efficacy.

- **Success Rate (SR):** SR reflects the entire episode success rate and is computed as $SR = \frac{1}{N} \sum_{i=1}^N S_i$, where S_i is a binary indicator of success in episode i . An episode is considered successful if every object reaches its goal.
- **Object Success Rate (OSR):** To quantify the effectiveness of the rearrangement process, we introduce the OSR, which measures the proportion of objects correctly placed in their target positions. The OSR is formulated as: $OSR = \frac{1}{N} \sum_{i=1}^N M_i^{succ} / M_i$, where M_i is the number of objects that need to be rearranged in episode i and M_i^{succ} is the number of objects successfully rearranged by the end of episode i .
- **Remaining Distance Ratio (RDR):** To evaluate the proximity of the final arrangement to the goal configuration, we introduce the RDR, defined as: $RDR = \frac{1}{N} \sum_{i=1}^N (D_i^{end} / D_i^{start})$, where D_i^{start} and D_i^{end} denote the summed Euclidean (L2) distances between all objects and their target positions at the start and end of episode i , respectively. Lower RDR values indicate better performance, with values approaching zero

representing perfect alignment between all furniture items and their target positions in the goal configuration.

D. Simulation

We implement ESRP within OmniGibson [9], a physics-based simulation framework built on NVIDIA Isaac Sim.

Robot Setup Our benchmark employs the embodiment developed from the Fetch robot, in which the robotic arm is removed, and a camera is mounted overhead.

Action Execution While our primary focus lies in high-level planning and decision-making, we strictly enforce physical fidelity through rigorous collision detection. Adopting the methodology established by HomeRobot [20], we implement continuous collision checking against the full 3D meshes of the robot, furniture, and environmental structures. This process guarantees trajectory feasibility and ensures the physical plausibility of every executed action.

To streamline simulation, we abstract the low-level grasping mechanics by instantaneously attaching objects to the agent upon interaction. This simplification allows us to prioritize high-level planning over control fidelity, aligning with established methodologies in rearrangement [15] and task planning [21]. Specifically, the agent is capable of fetching objects within a 1.0 m radius from its center. This interaction range is physically grounded; considering the robot’s diameter of 0.6 m and the substantial dimensions of the target furniture, it realistically approximates the reach required for manipulating large objects in real-world scenarios.

E. Baselines

We design four baselines spanning three paradigms (Fig. 5): an IL agent *ESRP-BC*; a RL agent *ESRP-PPO*; a foundation-model agent *ESRP-VLM* that prompts a pretrained VLM in a ReAct [14] loop; and a planning-based agent *ESRP-PLAN* that uses hierarchical TAMP with privileged global state.

ESRP-BC We deploy a conditional diffusion policy [13] that predicts a sequence of future actions $a_{t:t+h-1}$, where t is the current environment step and h is the prediction horizon. At each step t , the policy is conditioned on the vector c_t by concatenating features from two ResNet-50 [22] encoders (initialized from pretrained weights) applied to the egocentric observation history $I_{t-n+1:t}^{ego}$, and the goal image I_t^g , together with embeddings of the gripper-state history $g_{t-n+1:t}$, and projecting the result with an Multi-Layer Perceptron (MLP), where n is the history length.

During training, the expert future action sequence is encoded as one-hot vectors $x_{t:t+h-1}^0$, where the superscript 0 denotes the clean sample. We sample a diffusion step k and add Gaussian noise, $x_{t:t+h-1}^k = \sqrt{\bar{\alpha}_k} x_{t:t+h-1}^0 + \sqrt{1 - \bar{\alpha}_k} \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, where $x_{t:t+h-1}^k$ is the noisy action sequence and $\bar{\alpha}_k$ is the cumulative noise coefficient. A conditional 1D U-Net predicts the noise as $\hat{\epsilon} = \epsilon_\theta(x_{t:t+h-1}^k, k, c_t)$. The objective combines diffusion denoising with a cross-entropy loss on the denoised discrete action logits: $\mathcal{L}(\theta) = \|\epsilon - \epsilon_\theta(x_{t:t+h-1}^k, k, c_t)\|_2^2 - \lambda_{CE} \sum_{\tau=t}^{t+h-1} \log p_\theta(a_\tau | \hat{x}_\tau^0, c_t)$, where τ indexes future actions and $\hat{x}_\tau^0$ denotes the denoised logits at step τ .

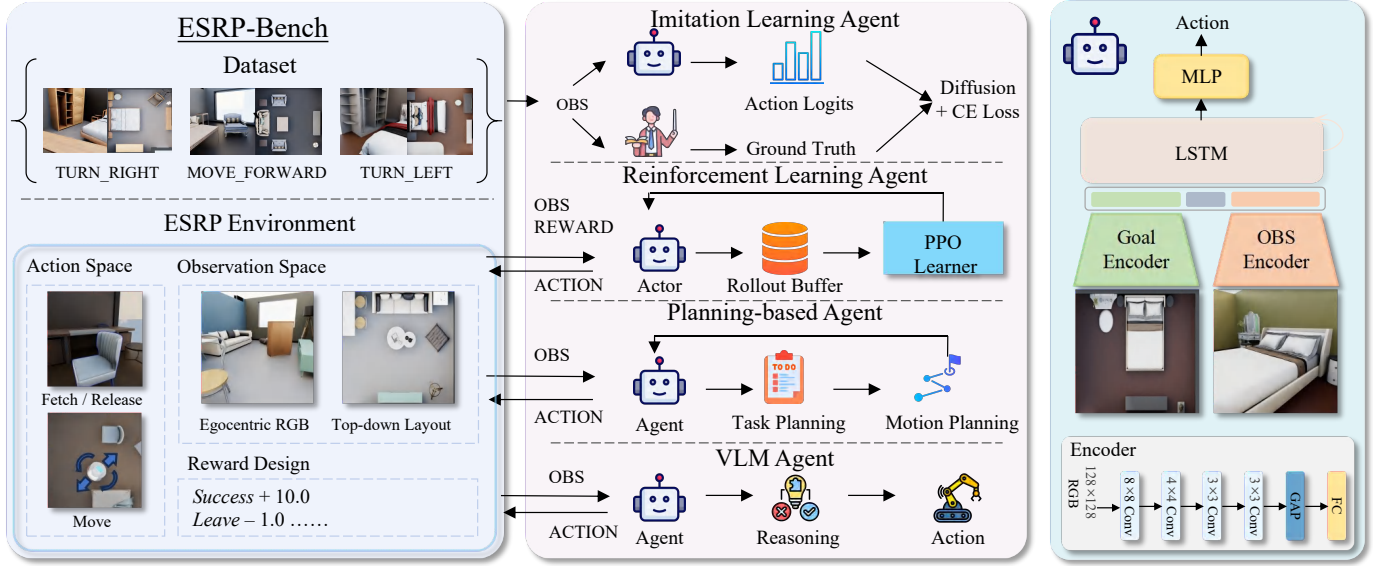


Fig. 5: **Pipeline overview.** (Left) ESRP-Bench consists of the dataset and the ESRP environment. (Middle) Four baseline models are: (1) IL agent uses expert trajectories from the dataset to train a policy that imitates expert behavior; (2) RL agent learns a policy by interacting with the ESRP environment; (3) foundation-model-based VLM agent prompts a pretrained VLM in a ReAct loop without task-specific training; and (4) planning-based agent utilizes global information to perform hierarchical task and motion planning. (Right) The model architectures for RL are shown, consisting of a CNN for visual perception and an LSTM module for memory. The gray box represents the embedding of the grasping state indicator g .

ESRP-PPO Our RL policy uses two visual encoders for I_t^{ego} and I_t^g , a learnable embedding for the gripper state g_t , and a single-layer LSTM with context length l that produces a hidden state h_t , which an MLP maps to the next action. Following prior scene-level rearrangement work [8], [21], we use a dense reward composed of: a +10 *Success* reward upon full task completion; a ± 1 *Arrival/Leave* signal when an object is released at, or displaced from, its goal; a *Potential* term $r_{\text{potential}} \cdot (D_{t-1} - D_t)$ with D_t the L2 distance from the held object to its target; a +0.01 per-step *Grasping* bonus to discourage premature releases; and a *Living* term (set to 0). Collisions incur no explicit penalty: any colliding action is rejected at the simulator level, so the agent simply forfeits the corresponding shaping rewards.

ESRP-VLM This baseline builds a multimodal ReAct [14] agent on top of the pretrained Qwen3-VL-2B-Instruct [23] VLM. At each step, it receives I_t^{ego} and I_t^g as visual inputs and a textual context that encodes g_t and a brief recent action history; following the ReAct paradigm, it alternates between a short reasoning trace and a discrete action $a_{t+1} \in \mathcal{A}$ via in-context prompting, without task-specific fine-tuning. This baseline represents the rapidly evolving family of foundation-model-based embodied agents.

ESRP-PLAN Our planning-based baseline assumes perfect perception (complete occupancy map, exact poses of all movable and unmovable objects, and target locations) and uses a hierarchical TAMP framework. The task planner determines the rearrangement order, and for each object the motion planner runs a four-phase routine: (i) BFS to the target object on a state space discretized at 0.05 m and $\pi/32$ rad, (ii) fetch, (iii) A* to the goal with Euclidean heuristic, and (iv) release. Collisions are checked against 2D polygonal obstacle representations, and a goal IoU threshold of 0.3 is used for verification.

IV. EXPERIMENTS

A. Experimental Setup

Implementation Details For the RL model, we employ PPO with a learning rate of 1.5×10^{-4} and a value function loss coefficient of 5×10^{-3} . Training is performed with a batch size of 2048, using minibatches of 64 for gradient updates. The history context length l is fixed at 64 time steps. We train the policy for 6 million environment steps until convergence.

For the IL model, we train for 500 epochs with 50 diffusion timesteps, using AdamW with a learning rate of 10^{-4} , weight decay of 10^{-6} , a batch size of 1024, a cross-entropy weight of $\lambda_{CE} = 0.2$, the prediction horizon $h = 8$, and the history length $n = 4$. At inference time, we use DDIM sampling with $D = 16$ denoising steps. The policy predicts an action sequence of length 8, while only the first 6 actions are executed, with a 0.1 probability of taking a random action at each step. This configuration is empirically found to yield the best performance.

For the planning-based model, we use BFS for navigation to objects with a maximum of 1000 iterations and A* for navigating to the goal with the same iteration limit. The state space is discretized with 0.05m grid spacing for positions and $\pi/32$ radians for orientations. Task planning uses an exhaustive search over object orderings when feasible, selecting the ordering that places the most objects successfully.

The IL model is trained on an NVIDIA A100 GPU, while all other training and inference experiments are conducted on an NVIDIA RTX 4090 GPU.

Dataset We split our dataset into training and test sets comprising 90% and 10% of the scenes, respectively. The training split is used for both RL and IL models' training.

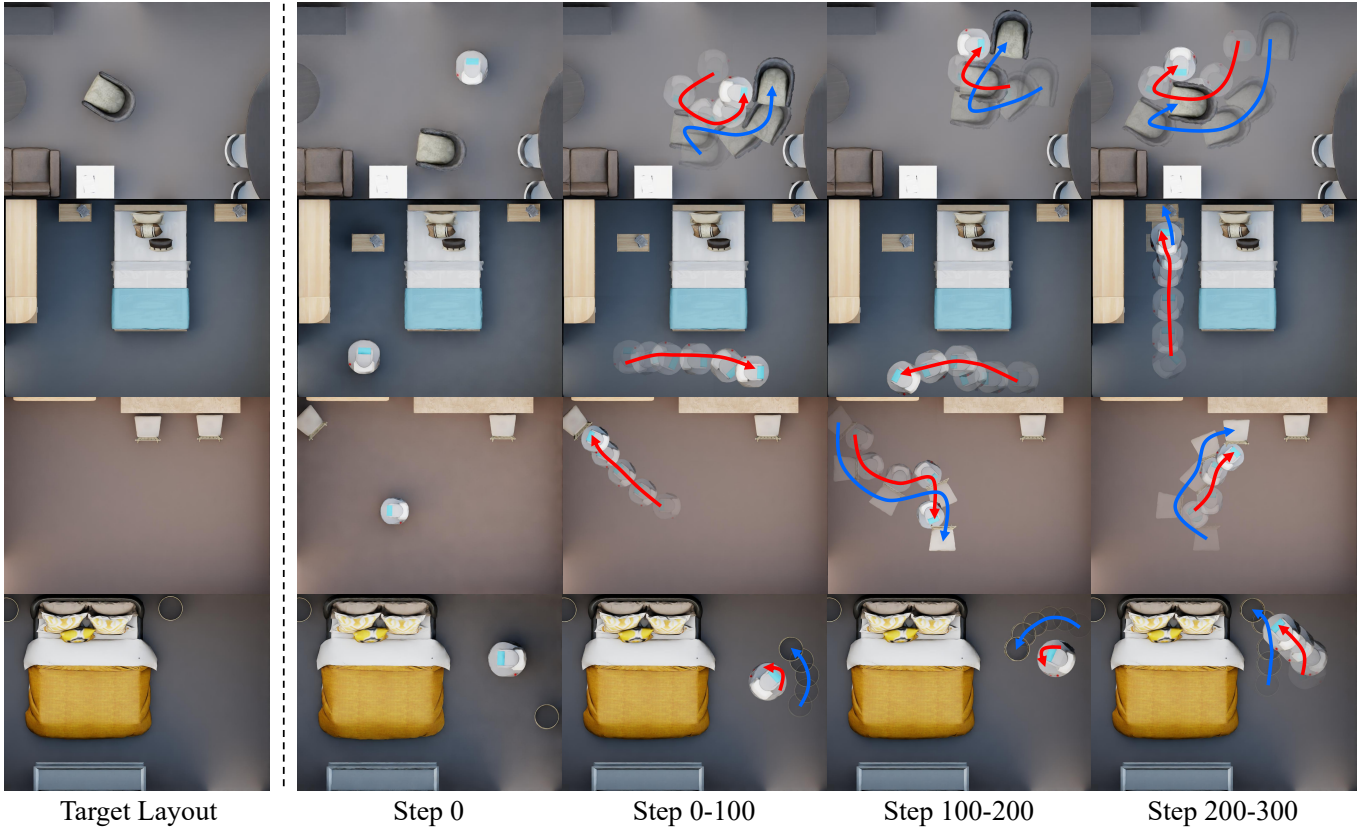


Fig. 6: **Examples of successfully rearranged scenes.** The agent successfully moves the objects to match the target layout. The red curve represents the agent’s trajectory, while the blue curve depicts the object’s trajectory.

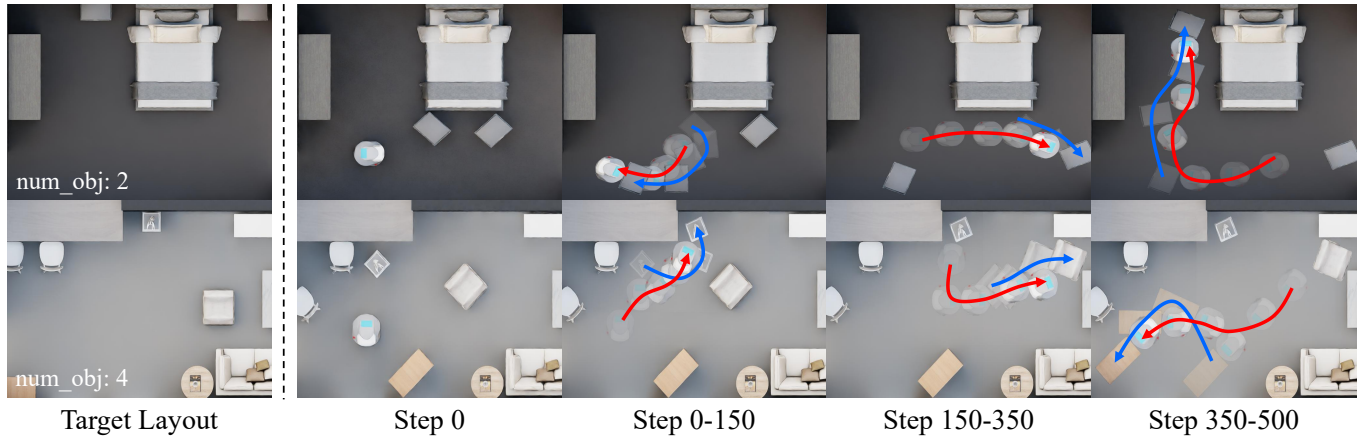


Fig. 7: **Failure cases.** As the number of rearrangeable objects increases, the task becomes more challenging due to its long-horizon complexity. In such cases, the agent rearranges the scene to approximate the target layout.

B. Results

Overall Rearrangement Results Table I demonstrates that: (i) *ESRP-PLAN* outperforms all other baselines across all three metrics. This advantage arises from its access to the global environment state, encompassing the current and target positions of all rearrangement objects, the agent’s pose, and the full obstacle occupancy of the scene, which enables explicit planning for navigation, object fetching, and obstacle avoidance. In contrast, learning-based approaches rely on partial observations and therefore exhibit weaker performance. (ii) *ESRP-PPO* consistently outperforms *ESRP-BC* across all three metrics.

While *ESRP-BC* achieves comparable SR and OSR with *ESRP-PPO*, indicating that behavior cloning has learned a certain degree of rearrangement capability from expert demonstrations, *ESRP-BC* performs worse in RDR, suggesting that errors accumulate and are amplified during long-horizon execution. In contrast, *ESRP-PPO* exhibits more comprehensive performance and maintains more stable behavior over long action sequences. (iii) *ESRP-VLM*, despite leveraging a pretrained vision-language foundation model, achieves performance comparable to *ESRP-BC* but does not surpass *ESRP-PPO*, indicating that off-the-shelf foundation-model priors alone are insufficient for the long-horizon spatial reasoning required by ESRP.

TABLE I: Performance of the baseline methods on the test set.

Baseline	SR $\uparrow$	OSR $\uparrow$	RDR $\downarrow$
ESRP-PPO	20.18%	20.35%	0.9845
ESRP-BC	19.09%	20.31%	1.0170
ESRP-VLM	17.64%	18.97%	0.9830
ESRP-PLAN	30.20%	36.59%	0.6916

TABLE II: The performance of baseline models across scenes of varying difficulty levels.

Baseline	Easy			Medium			Hard		
	SR $\uparrow$	OSR $\uparrow$	RDR $\downarrow$	SR $\uparrow$	OSR $\uparrow$	RDR $\downarrow$	SR $\uparrow$	OSR $\uparrow$	RDR $\downarrow$
ESRP-PPO	33.33%	33.33%	0.9040	0.54%	0.54%	1.0645	0.00%	0.00%	1.0048
ESRP-BC	29.44%	29.44%	0.9631	0.91%	4.20%	1.1157	0.00%	4.44%	1.0704
ESRP-VLM	27.64%	27.64%	0.9523	0.00%	3.99%	1.0372	0.00%	0.00%	1.0378
ESRP-PLAN	35.73%	35.73%	0.7101	22.40%	39.71%	0.6425	0.00%	22.00%	0.8165

Varying Difficulty Levels Table II presents the overall rearrangement performance of our baseline models across scenes of varying difficulty levels. These results reveal that current approaches struggle progressively as the number of objects requiring rearrangement increases. This trend reflects compounded challenges, including long-horizon planning under egocentric partial observability, combinatorial object ordering decisions as the number of targets grows, and dynamically evolving scenes that introduce cascading dependencies, all of which make spatial reasoning and planning increasingly difficult as task complexity scales. Performance degradation is particularly pronounced in complex scenes, highlighting the **challenge of maintaining spatial reasoning and planning capabilities as task complexity scales**.

Failure-Mode Decomposition To complement the aggregate metrics with a diagnostic view, we partition each test rollout into four mutually exclusive stages, defined by how far the agent progresses through the rearrangement pipeline (Table III). Across *ESRP-PPO*, *ESRP-BC*, and *ESRP-VLM*, the dominant failure mode is consistently Stage 3 (placement failure), accounting for 60.85%, 62.85%, and 66.55% of all rollouts, respectively, whereas only 4–6% of episodes terminate at the engagement stage and 11–13% stall during navigation. This decomposition shows that, regardless of paradigm, the principal bottleneck of ESRP is **precise placement under egocentric observability** rather than navigation or grasp initiation, providing a concrete diagnostic target for future research.

Effect of a Stronger VLM Backbone To assess whether the modest performance of *ESRP-VLM* is attributable to the specific VLM backbone rather than to the ReAct-based prompting paradigm itself, we additionally evaluate *ESRP-VLM* with Gemini 2.5 Flash-Lite [24] in place of Qwen3-VL-2B-Instruct. Replacing the backbone improves performance across all three metrics, from 17.64% SR, 18.97% OSR, and 0.9830 RDR with Qwen3-VL-2B-Instruct to 23.01% SR, 23.75% OSR, and 0.9450 RDR with Gemini 2.5 Flash-Lite. This confirms that stronger general-purpose VLM priors partially alleviate the difficulty of ESRP.

Qualitative Results Fig. 6 and Fig. 7 illustrate the rearrangement process of *ESRP-PPO*. In Fig. 6, the agent successfully identifies the designated object and moves it to the target location, completing the task within half of the available steps. Fig. 7 highlights some failure cases of our method. Although

TABLE III: Failure-mode decomposition. Each test rollout is assigned to exactly one of four mutually exclusive stages, defined by how far the agent progresses through the rearrangement pipeline. Stage 1: the agent never approaches a rearrangeable object (*navigation failure*). Stage 2: the agent reaches an object but never grasps it (*engagement failure*). Stage 3: the agent grasps an object but fails to place it at the goal (*placement failure*). Stage 4: complete success. ESRP-VLM results are based on a single run.

Baseline	Stage 1 (%)	Stage 2 (%)	Stage 3 (%)	Stage 4 (%) $\uparrow$
ESRP-PPO	13.21	5.76	60.85	20.18
ESRP-BC	11.33	6.73	62.85	19.09
ESRP-VLM	11.64	4.18	66.55	17.64

the scene is not fully rearranged successfully, the agent achieves a lower RDR by placing objects closer to their designated goals. This suggests that although the agent’s perception is adequate, **efficiently planning under time constraints and achieving robust performance from egocentric viewpoints remain critical challenges** requiring further research.

C. Discussion

Based on the experimental results, we discuss four key questions regarding the ESRP task:

Q1: What makes the task challenging, and how can these challenges be addressed?

Comparing planning- and learning-based approaches in Table I reveals two main sources of difficulty. The first is **3D Scene Understanding**: the gap between *ESRP-PLAN* (perfect perception) and learning-based methods (partial egocentric observations) reflects the difficulty of inferring global scene states from local views. The second is **Long-Horizon Task Planning**: even with perfect perception, *ESRP-PLAN* still fails on Hard-level scenes (0% SR, Table II), showing that the combinatorial complexity of multi-object rearrangement poses a challenge independent of perception.

Q2: Are the experimental settings and task abstractions reasonable?

Our design deliberately abstracts low-level control to isolate the core challenges of embodied AI. The high-level manipulation primitives we adopt (“magic snapping”), together with strict collision checks, are shared with mainstream embodied rearrangement benchmarks [15], [21], [25], and policies trained under such abstractions transfer to physical platforms with only routine sim-to-real adjustments [20]. This choice is further validated by Table II: current methods still fail on complex scenes (e.g., 0% SR in Hard scenarios) despite the simplified manipulation, confirming that the bottleneck is *long-horizon planning* and *spatial reasoning* under partial observability, not grasping mechanics – exactly the capability ESRP-Bench is designed to stress-test.

Q3: How does ESRP relate to real-world deployment, given that all experiments are conducted in simulation?

Our simulation-only evaluation follows established practice in embodied rearrangement benchmarks [15], [20], which favor controlled, reproducible comparisons before real-world deployment. The abstractions ESRP adopts – egocentric RGB observations and high-level fetch/release primitives – are shared with these benchmarks, and prior work shows that policies trained under them transfer to physical platforms with

routine sim-to-real adaptation [20]. The absence of real-robot results mainly reflects a hardware bottleneck, as commodity platforms cannot yet reliably manipulate furniture-scale objects, whereas the long-horizon planning and partial-observability understanding that ESRP isolates are exactly the capabilities future deployments will require.

Q4: Is specifying the target layout as a top-down RGB image realistic for real-world deployment?

Top-down layout images offer an information-dense, unambiguous way to specify multi-object goals: a single image conveys both *which* objects to move and *where* to place them, unlike natural language, which grows lengthy and ambiguous as object count increases. Such images are increasingly easy to acquire: a commodity RGB-D phone or LiDAR tablet can scan a room, reconstruct a 3D mesh, and segment individual furniture pieces; the user then edits the layout (e.g., dragging a sofa) and renders it from a top-down view. This scan–segment–edit–render pipeline is already validated by commercial tools such as MagicPlan [26]. Beyond user-driven editing, top-down goals can equivalently come from upstream scene synthesis modules [27], [28], letting ESRP operate on either user-authored or generated layouts.

V. CONCLUSION

In this paper, we propose ESRP, the first embodied rearrangement task that focuses on repositioning furniture within a complex 3D room with the egocentric image as the observation. To accelerate research in this emerging domain, we introduce ESRP-Bench, a comprehensive benchmark that encompasses a large-scale, diverse dataset of indoor scenes and a suite of carefully designed evaluation metrics. By providing both a realistic simulation environment and rigorous performance assessment tools, our work offers a significant step toward advancing embodied rearrangement capabilities.

Limitations This work concentrates on room-level furniture rearrangement and relies exclusively on egocentric RGB observations. We envision extending ESRP to house-level planning and incorporating additional sensor modalities, such as depth, bird’s-eye-view cameras, and LiDAR, as vital steps toward broader applicability and real-world deployment.

REFERENCES

- [1] J. E. King, M. Cagnetti, and S. S. Srinivasa, “Rearrangement planning using object-centric and robot-centric action spaces,” in *International Conference on Robotics and Automation (ICRA)*, 2016. 1, 2
- [2] Y. Labbé, S. Zagoruyko, I. Kalevatykh, I. Laptev, J. Carpentier, M. Aubry, and J. Sivic, “Monte-carlo tree search for efficient visually guided rearrangement planning,” *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 2, pp. 3715–3722, 2020. 1, 2
- [3] A. Goyal, A. Mousavian, C. Paxton, Y.-W. Chao, B. Okorn, J. Deng, and D. Fox, “Ifor: Iterative flow minimization for robotic object rearrangement,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1
- [4] H. Ren and A. H. Qureshi, “Neural rearrangement planning for object retrieval from confined spaces perceivable by robot’s in-hand rgb-d sensor,” in *International Conference on Robotics and Automation (ICRA)*, 2024. 1, 2
- [5] G. Xiong, Q. Fu, H. Fu, B. Zhou, G. Luo, and Z. Deng, “Motion planning for convertible indoor scene layout design,” *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, vol. 27, no. 12, pp. 4413–4424, 2020. 1
- [6] Z. Wang, H. Wang, and W. Liang, “Mastering scene rearrangement with expert-assisted curriculum learning and adaptive trade-off tree-search,” in *International Conference on Intelligent Robots and Systems (IROS)*, 2024. 1, 2
- [7] D. Batra, A. Chang, S. Chernova, A. Davison, J. Deng, V. Koltun, S. Levine, J. Malik, I. Mordatch, R. Mottaghi, M. Savva, and H. Su, “Rearrangement: A challenge for embodied ai,” *arXiv preprint arXiv:2011.01975*, 2020. 1
- [8] H. Wang, W. Liang, and L.-F. Yu, “Scene mover: Automatic move planning for scene arrangement by deep reinforcement learning,” *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020. 1, 2, 3, 5
- [9] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun *et al.*, “Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation,” in *Conference on Robot Learning (CoRL)*, 2023. 1, 4
- [10] H. Fu, B. Cai, L. Gao, L.-X. Zhang, J. Wang, C. Li, Q. Zeng, C. Sun, R. Jia, B. Zhao *et al.*, “3d-front: 3d furnished rooms with layouts and semantics,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 3
- [11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017. 2
- [12] N. Gavenski, F. Meneguzzi, M. Luck, and O. Rodrigues, “A survey of imitation learning methods, environments and metrics,” *arXiv preprint arXiv:2404.19456*, 2024. 2
- [13] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, “Diffusion policy: Visuomotor policy learning via action diffusion,” in *Robotics: Science and Systems (RSS)*, 2023. 2, 4
- [14] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafraan, K. Narasimhan, and Y. Cao, “ReAct: Synergizing reasoning and acting in language models,” in *International Conference on Learning Representations (ICLR)*, 2023. 2, 4, 5
- [15] L. Weihs, M. Deitke, A. Kembhavi, and R. Mottaghi, “Visual room rearrangement,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 4, 7
- [16] D. Huang, C. Tang, and H. Zhang, “Efficient object rearrangement via multi-view fusion,” in *International Conference on Robotics and Automation (ICRA)*, 2024. 2
- [17] Q. A. Wei, S. Ding, J. J. Park, R. Sajjani, A. Poulenard, S. Sridhar, and L. Guibas, “Lego-net: Learning regular rearrangements of objects in rooms,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3
- [18] A. Orthey, C. Chamzas, and L. E. Kavraki, “Sampling-based motion planning: A comparative review,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 7, 2023. 2
- [19] J. Li, W. Huang, Z. Wang, W. Liang, H. Di, and F. Liu, “Flona: Floor plan guided embodied visual navigation,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2025. 2
- [20] S. Yenamandra, A. Ramachandran, K. Yadav, A. Wang, M. Khanna, T. Gervet, T.-Y. Yang, V. Jain, A. W. Clegg, J. Turner *et al.*, “Homerobot: Open-vocabulary mobile manipulation,” *arXiv preprint arXiv:2306.11565*, 2023. 4, 7, 8
- [21] K. Mirakhor, S. Ghosh, D. Das, and B. Bhowmick, “Task planning for object rearrangement in multi-room environments,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2024. 4, 5, 7
- [22] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 4
- [23] Qwen Team, “Qwen3-VL: An open-source multimodal foundation model from the Qwen team,” <https://github.com/QwenLM/Qwen3-VL>, 2025. 5
- [24] Google DeepMind, “Gemini 2.5 Flash-Lite,” <https://deepmind.google/models/gemini/flash-lite/>, 2025. 7
- [25] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. S. Chaplot, O. Maksymets *et al.*, “Habitat 2.0: Training home assistants to rearrange their habitat,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 7
- [26] Sensopia Inc., “magicplan: Mobile floor plan scanning application,” <https://www.magicplan.app>. 8
- [27] J. Tang, Y. Nie, L. Markhasin, A. Dai, J. Thies, and M. Nießner, “Diffuscene: Denoising diffusion models for generative indoor scene synthesis,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8
- [28] J. Luiten, G. Kopanas, B. Leibe, and D. Ramanan, “Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis,” in *International Conference on 3D Vision (3DV)*, 2024. 8