Hierarchical Skill Retrieval for Data-Efficient Adaptation of Vision-Language-Action Models

Haoran Hao[1], Shahram Najam Syed[1], Jeff Schneider[1], Jeffrey Ichnowski[1]

***Abstract*— While Vision-Language-Action (VLA) models pretrained on large-scale robot datasets provide a strong foundation for robot manipulation, their performance can degrade when adapted to new tasks with limited task-specific demonstrations. Retrieval offers a practical way to reuse existing demonstrations for data-efficient adaptation, but existing methods often rely on visual similarity, state-action representations, or task-level language matching. These approaches may overlook the hierarchical structure of long-horizon manipulation tasks, where complete task matches are rare but reusable skills are often abundant. To address this challenge, we propose Hierarchical Skill Retrieval (HSR), a retrieval framework for data-efficient VLA adaptation. Specifically, HSR first decomposes a target task into candidate skill sequences. It evaluates each plan based on both semantic plausibility and skill reliability estimated from the prior dataset. The selected decomposition is then used for hybrid retrieval. This combines subtask-level language retrieval with behavior-feature reranking to identify demonstrations that are both semantically relevant and compatible with the target task. Finally, we adapt the policy through a two-stage pretraining and finetuning pipeline, which separates general skill acquisition from task-specific adaptation. Experiments on the LIBERO benchmark and several real-world robot manipulation tasks show that HSR improves the average success rate by 10.3% and 21.3% over the strongest baseline, respectively. These results demonstrate the effectiveness of structured skill-level retrieval for data-efficient VLA adaptation.**

I. Introduction

Foundation models pretrained on large-scale datasets have shown strong generalization ability and broad transferable knowledge [1], [2]. Motivated by recent advances in large multimodal models, vision-language-action (VLA) models [3]–[8] extend pretrained vision-language representations to robot control, aiming to transfer their general visual and linguistic understanding to downstream manipulation tasks. A common way to adapt VLA models to downstream tasks is to finetune them on target-task demonstrations [3], [9]. While effective in many settings, this approach is fundamentally limited by data scarcity: collecting high-quality robot demonstrations is expensive and time-consuming, and target datasets are often too small to fully adapt large policy models for reliable real-world deployment.

A complementary line of work improves adaptation by reusing large prior datasets as additional training data [10]–[14]. Among these methods, retrieval-based approaches are particularly effective because they can identify useful

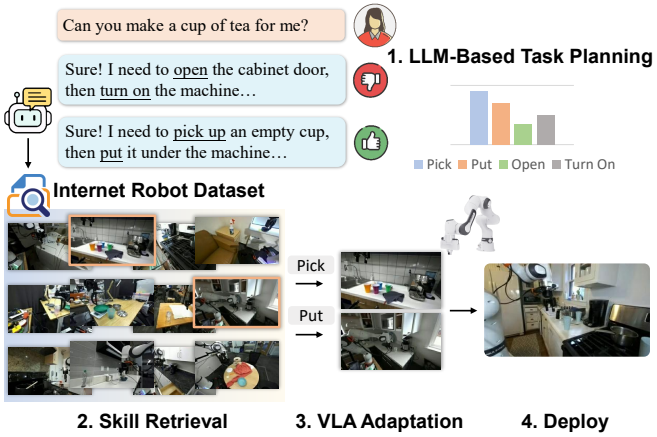


Fig. 1: Hierarchical Skill Retrieval (HSR): LLM-based task decomposition enables structured demonstration retrieval for data-efficient VLA adaptation.

demonstrations from large prior datasets such as Open X-Embodiment [6]. Existing retrieval methods typically rely on similarity metrics between target data and prior data, including distances in latent state-action spaces [10], [11] and subtrajectory-level similarity measures [13], [14]. By augmenting the target dataset with retrieved demonstrations, these methods expand the effective training set and can improve downstream task performance.

Although retrieval-based methods have shown promising results in several settings, most existing approaches are designed for vision-based policies and primarily rely on visual or motion similarity. As a result, they often overlook higher-level task semantics and relationships between tasks, and thus do not fully exploit language instructions. VLA models, on the other hand, have demonstrated strong capabilities in following language instructions [15], [16] and transferring knowledge across related tasks [4], [17], [18], such as manipulating different objects using similar actions. This suggests that language can provide an important semantic signal for improving task generalization.

However, language-based retrieval is also limited when it relies only on coarse instruction similarity. For complex long-horizon tasks, such as *"make a cup of tea"*, directly retrieving data using the full task instruction is often difficult, since complete task matches rarely exist in prior datasets. In contrast, many of the underlying skills required to complete such tasks, such as *"grasp a teapot"* or *"pick up a tea bag"*, are common and well represented. Effectively leveraging the hierarchical structure of complex tasks and retrieving data at the level of reusable subskills therefore remains an open

[1]Robotics Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213. {haoranha, jeff4, jichnows}@andrew.cmu.edu

challenge.

To address these issues, we propose Hierarchical Skill Retrieval (HSR), a framework for data-efficient VLA adaptation based on task decomposition and skill-level retrieval. Specifically, (1) we first cluster the prior dataset into reusable skills, such as Pick, Open, and Close. Given a target task, we use a Large Language Model (LLM)-based planner to decompose the task into subtasks grounded in the available skills. (2) We then score candidate plans, select the most suitable decomposition, and use the resulting subtasks as queries for retrieval. (3) Finally, we adapt the VLA model in two stages: first by pretraining on language-retrieved skill data, and then by finetuning on the target demonstrations together with feature-reranked retrieved samples.

The HSR framework leverages both the structure of the target task and the prior dataset, enabling data-efficient adaptation of pretrained VLA models under limited target data. We evaluate the proposed approach in the LIBERO [19] simulation environment and on several real-world long-horizon robot manipulation tasks. Experimental results show that HSR achieves the best average performance across a wide range of tasks and provides particularly large gains on long-horizon composite manipulation tasks. Notably, HSR improves the success rate by 10.3% on LIBERO and by 21.3% on real-world tasks relative to the strongest baseline.

Our contributions are summarized as follows:

- We propose a hierarchical retrieval framework for long-horizon manipulation, which decomposes tasks into skill sequences and retrieves demonstrations at the skill level based on task structure and semantics.
- We introduce a two-stage adaptation strategy. It first uses retrieved demonstrations to acquire transferable skills, and then finetunes on target data together with feature-reranked retrieved samples.
- We validate the proposed framework on both the LIBERO benchmark and real-world long-horizon manipulation tasks, showing improved success rates and data efficiency over strong retrieval-based baselines.

II. Related Work

A. Vision-Language Conditioned Manipulation

Earlier work on robot manipulation using deep neural networks typically restricted the input modality to either visual observations or robot states [20]. In recent years, vision-language models (VLMs) [1], [2], [21] pretrained on large-scale image-text pairs have demonstrated strong capabilities in multimodal perception and understanding. Based on these multimodal foundation models, Vision-Language-Action (VLA) policy models [3], [5]–[8], [22]–[24] aim to transfer the general knowledge learned by VLMs to robot control. By training on multi-task datasets and conditioning on both language instructions and visual observations, VLA models show improved generalization in vision-language manipulation tasks. Pretrained VLA models can leverage large-scale datasets such as Open X-Embodiment [6], enabling cross-task generalization and improved instruction understanding in robot manipulation. However, despite containing rich general knowledge, VLA models often perform suboptimally on specific downstream tasks. In many practical scenarios, downstream tasks lack sufficient data for training or finetuning. As a result, how to effectively adapt VLA models to data-scarce downstream tasks remains an open and underexplored problem.

B. Retrieval for Model Adaptation

Retrieval-based methods aim to select relevant data from existing datasets, such as DROID [25] and Open X-Embodiment [6], to improve adaptation to specific target tasks. Most prior work learns latent representations of trajectories and performs similarity-based retrieval. For example, some methods encode state-action pairs [10], optical flow [26], or sub-trajectories [13], [14], and retrieve data based on visual or motion similarity. Other approaches extend this idea by incorporating multiple modalities [12] or applying importance weighting between prior and target datasets [11]. There are also methods that retrieve relevant past experiences and directly learn from them [27]. However, these approaches mainly rely on visual or motion-level similarity, and often do not explicitly consider the semantic and hierarchical structure of manipulation tasks, which is crucial for modern VLA models. Trajectories with similar motion patterns may correspond to different subtasks, such as pushing versus picking, leading to ambiguous or less relevant retrieval results. Recent work has started to explore semantic-level reasoning in robot tasks. DROC [28] shows that LLMs can generate language-based corrections, which are then used as queries to retrieve relevant knowledge from a knowledge base. MT3 [29] decomposes tasks into alignment and interaction policies, enabling task composition and improved generalization through test-time retrieval. Motivated by these insights, we propose a framework that decomposes long-horizon manipulation tasks into subtasks and retrieves semantically related experiences from a prior dataset.

C. Robot Task Planning

Task planning aims to decompose and schedule complex tasks by combining symbolic reasoning, motion planning, and learning-based methods [30]. Effective planning enables zero-shot generalization to novel tasks by reusing existing skills [31], [32], reducing the need for additional data to learn composite tasks. Some approaches rely on human-defined transition models or learn such models from data to generate composed plans [33]–[36]. However, building such models often requires substantial human effort, large training datasets, and additional components, such as object detectors [35]. More recent work explores the reasoning capabilities of LLMs for task planning [37]–[42]. The resulting subtasks are typically executed by separate policy modules [40], [43]. Most existing approaches focus on task decomposition at test time, while placing less emphasis on how the required skills are trained or adapted. In this work, we use task decomposition as a training-time retrieval interface. Given a target task, we identify a sequence of

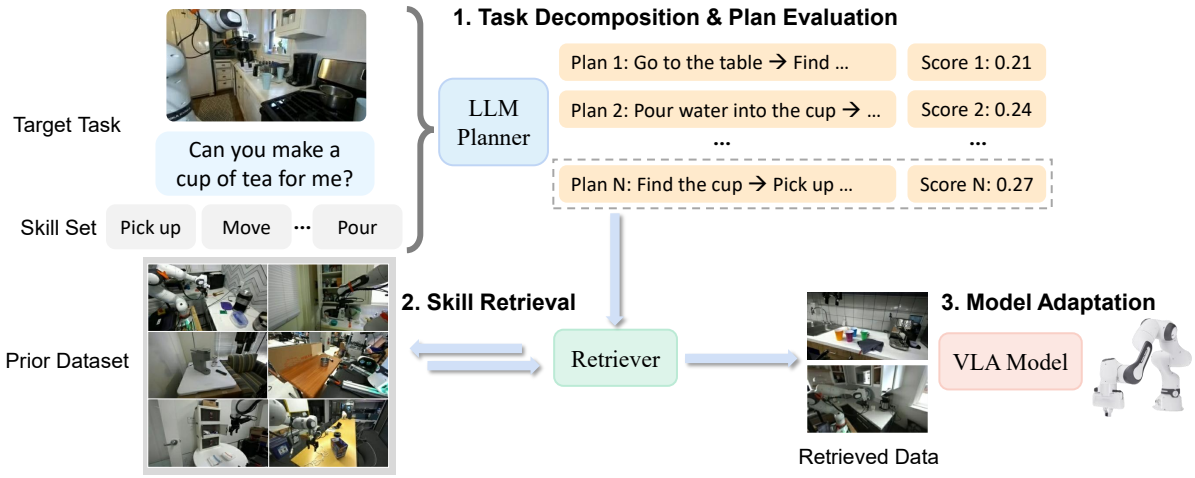


Fig. 2: **Overview of the Proposed HSR Framework**. HSR identifies existing skills in the prior dataset based on task-instruction clustering. Given a long-horizon target task, *1.* HSR first decomposes the task into several subtasks, evaluates and selects the plan with the highest score; *2.* for each subtask, HSR retrieves data based on the subtask instruction, and further reranks using representative demonstrations from the target task to filter out low-consistency data; *3.* HSR first pretrains the policy on retrieved data to learn general skills, and then finetunes it on task-related data to adapt to the target task.

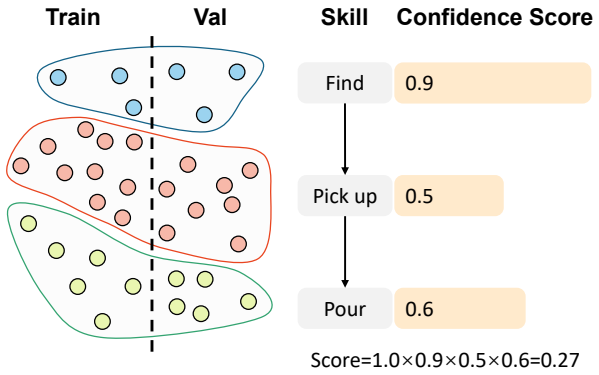


Fig. 3: **Illustration of Plan Evaluation**. We first cluster $\mathcal{D}_{\text{prior}}$ into reusable skills. Each clustered skill is then evaluated by finetuning a pretrained VLA model on the training split and measuring the BC loss on its validation set. The resulting loss serves as a lightweight proxy for skill reliability, enabling efficient plan evaluation without costly environment rollouts.

semantically relevant skills and retrieve training demonstrations that support each skill, thereby improving data-efficient adaptation of the policy.

III. Problem Statement

We study few-shot adaptation of a pretrained vision-language-conditioned policy using a small set of target demonstrations together with a large prior dataset. Given a pretrained VLA policy π_θ, the policy maps a visual observation o_t and a language instruction T to an action a_t: $a_t = \pi_\theta(o_t, T)$. For each target task, we are given a target dataset $\mathcal{D}_{\text{t}} = \{\tau_i\}_{i=1}^{N}$, where each trajectory $\tau_i = \{(o_t, a_t)\}_{t=1}^{H}$ is an expert demonstration. We assume access to a large prior dataset $\mathcal{D}_{\text{prior}}$, which contains demonstrations from diverse tasks and domains. Our objective is to identify a useful subset $\mathcal{D}_{\text{r}} \subset \mathcal{D}_{\text{prior}}$ such that adaptation on $\mathcal{D}_{\text{t}} \cup \mathcal{D}_{\text{r}}$ yields a policy that better matches expert behavior and performs well on the target task.

IV. Method

A. Hierarchical Task Decomposition

Language-based retrieval is often insufficient for robot manipulation tasks, where high-level instructions are ambiguous and rarely match complete demonstrations in the prior dataset. This limitation is particularly severe for long-horizon and compositional tasks, such as *"make a cup of tea"*, which are seldom observed as monolithic behaviors in practice. However, these tasks are typically composed of reusable subskills that recur across different tasks, such as *"pick up a cup"* and *"turn on a machine"*. Motivated by this observation, we decompose a target task into a sequence of semantic subtasks and use each subtask as a query to retrieve relevant demonstrations from $\mathcal{D}_{\text{prior}}$. This skill-level retrieval enables more precise reuse of prior experiences and improves adaptation to long-horizon manipulation. The overall HSR framework is illustrated in Fig. 2.

Although LLMs have shown strong performance in task planning, prior work such as SayCan [41] points out an important limitation: LLMs are not grounded in the robot's actual capabilities and may therefore generate plans containing subtasks that are difficult or impossible to execute. As a result, purely language-based planning may produce decompositions that are poorly matched to the policy's learned skills. To mitigate this issue, we leverage $\mathcal{D}_{\text{prior}}$ to estimate the reliability of different skills and use this information to guide task decomposition for retrieval. Our key idea is to prefer decompositions composed of skills that are both semantically relevant to the target task and well supported by the available data. Specifically, we first extract text embeddings for all task instructions in $\mathcal{D}_{\text{prior}}$ and apply K-means clustering to group them into a set $S = \{s_1, \ldots, s_K\}$, where each cluster s_i corresponds to a semantically coherent skill, such as *Find*, *Pick up*, or *Pour*.

For each s_i, we estimate its reliability using the pretrained VLA model. Rollout-based evaluation provides the most direct measure of policy performance, but is computationally

expensive and often impractical, especially in real-world settings. Instead, following DataMIL [44], we use held-out behavior cloning (BC) loss as a lightweight proxy for skill reliability. As illustrated in Fig. 3, we split $\mathcal{D}_{\text{prior}}$ into training and validation sets and finetune the pretrained VLA model on the training split to obtain $\pi_{\theta'}$. We then evaluate $\pi_{\theta'}$ on the validation trajectories associated with each skill cluster. For a validation trajectory $\tau = \{(o_t, a_t)\}_{t=1}^{H_\tau}$, let $\ell_{\text{BC}}(\pi_{\theta'}(o_t, T), a_t)$ denote the per-step BC loss. We define the average trajectory-level BC loss as

$$\mathcal{L}_{\text{traj}}(\tau) = \frac{1}{H_\tau} \sum_{t=1}^{H_\tau} \ell_{\text{BC}}(\pi_{\theta'}(o_t, T), a_t). \tag{1}$$

For a skill cluster s_i, let $\mathcal{V}_i$ denote its validation trajectories. The validation loss of s_i is computed as

$$\mathcal{L}_{\text{val}}(s_i) = \frac{1}{|\mathcal{V}_i|} \sum_{\tau \in \mathcal{V}_i} \mathcal{L}_{\text{traj}}(\tau), \tag{2}$$

and $H(s_i)$ denotes the average trajectory length in $\mathcal{V}_i$. To ensure comparability across skill clusters, we normalize both quantities using min-max normalization. Since long-horizon skills are more susceptible to compounding errors, we define the *Skill Confidence Score* as

$$C(s_i) = \exp\left(-\alpha \widetilde{\mathcal{L}}_{\text{val}}(s_i) - \beta \widetilde{H}(s_i)\right), \tag{3}$$

where $\widetilde{\mathcal{L}}_{\text{val}}(s_i)$ and $\widetilde{H}(s_i)$ denote the normalized validation loss and average trajectory length, respectively. We use fixed weights $\alpha = 0.6$ and $\beta = 0.8$ across all experiments. Rather than estimating the true success probability, $C(s_i)$ serves as a rollout-free proxy for skill reliability, favoring skills that are easier to imitate and less susceptible to error accumulation during long-horizon execution.

We then query the LLM with the target task description and the set of available skills to generate candidate decomposed plans. Each plan is represented as an ordered sequence of subtasks $p = (u_1, u_2, \ldots, u_n)$. For each subtask u_j, we assign its nearest skill cluster based on text-embedding similarity, denoted as $s(u_j)$. In addition, we prompt the LLM to provide a semantic score for each plan, reflecting its task coverage, logical consistency, and semantic plausibility with respect to the target instruction. The score is normalized to $S_{\text{sem}}(p) \in (0, 1]$ before plan ranking. We then evaluate each candidate plan by combining the semantic score with the confidence scores of its constituent skills:

$$\text{Score}(p) = S_{\text{sem}}(p) \cdot \prod_{j=1}^{n} C\big(s(u_j)\big). \tag{4}$$

We select the plan with the highest score as the final decomposition, and use it for subsequent retrieval from $\mathcal{D}_{\text{prior}}$.

B. Hybrid Skill Retrieval

Language-based retrieval is effective at identifying demonstrations that are semantically relevant to a target task. However, VLA policies are also sensitive to low-level factors such as environment layouts, lighting conditions, and camera viewpoints. As a result, demonstrations that are similar in language instruction may still differ substantially from the current execution context, which can degrade policy adaptation and execution. To address this issue, we propose a hybrid skill retrieval strategy that combines language-based retrieval with behavior-feature reranking. The key idea is to first use language similarity to ensure semantic relevance, and then use behavior similarity to select demonstrations that are more compatible with the target task.

Specifically, given a subtask, we first retrieve the top $M\%$ of episodes from $\mathcal{D}_{\text{prior}}$ according to language similarity. We then flatten these episodes into candidate frames and rerank them using behavior features. Following Behavior Retrieval [10], we train a VAE encoder on $\mathcal{D}_{\text{prior}}$ to extract their behavior features. Let z_f denote the latent feature of a candidate frame f, and let $\{z_j^t\}_{j=1}^{N_t}$ denote the latent features extracted from $\mathcal{D}_{\text{t}}$, where N_t is the number of target frames. The behavior similarity score of f is defined as

$$s(f) = \max_{1 \le j \le N_t} \left(-\|z_f - z_j^t\|_2\right), \tag{5}$$

which corresponds to the negative distance to the nearest target example in latent space. We retain the top $F\%$ of candidate frames with the highest similarity scores as the final retrieved data for policy adaptation. By combining semantic alignment at the episode level with behavior compatibility at the frame level, this hybrid retrieval strategy improves the relevance and effectiveness of the retrieved skills under varying environmental conditions.

C. Policy Adaptation

To better exploit the retrieved data, we adopt a two-stage policy adaptation scheme. Let $\mathcal{D}_{\text{lang}}$ denote the training samples from the language-retrieved episodes and let $\mathcal{D}_{\text{rerank}}$ denote the reranked frames constructed from $\mathcal{D}_{\text{lang}}$. In the first stage, we pretrain the policy on $\mathcal{D}_{\text{lang}}$ to acquire reusable behaviors. In the second stage, we finetune the policy on $\mathcal{D}_{\text{t}} \cup \mathcal{D}_{\text{rerank}}$ for task-specific adaptation. This design differs from prior single-stage co-training methods [10], [11], which jointly optimize the policy on the retrieved data $\mathcal{D}_{\text{r}}$ and the target data $\mathcal{D}_{\text{t}}$ in one step. In contrast, our two-stage strategy explicitly separates general skill acquisition from task-specific adaptation and avoids treating all retrieved data as equally informative.

V. Experiments

Models and Training. We use the pretrained SmolVLA-450M [8] model as the base policy model. During adaptation, we update only the parameters of the action expert, resulting in approximately 100M trainable parameters. In the HSR framework, we employ Qwen3-VL-4B [45] as the LLM planner for task decomposition. We use $K = 10$ clusters for both simulation and real-world experiments. For each target task, the planner generates 10 candidate plans, and we select the plan with the highest score. We use a pretrained BERT model [46] to extract text embeddings for retrieval.

Environments. We evaluate the proposed framework in both simulation and real-world settings. In simulation, we use the LIBERO benchmark [19]. For real-world evaluation,

TABLE I: **Comparison on LIBERO Benchmark**. We compare our method with behavior cloning (BC), training with all data, language-based retrieval, and four retrieval baselines: FR [26], STRAP [14], BR [10], and IWR [11]. We report the success rate (in %) and standard deviation across 3 seeds. The best results are in **bold** and the second-best results are underlined.

Category	Pick-Place					Spatial Understanding		Composite Manipulation			
Task	Soup-S	Cheese-B	Book	Soup-C	Moka-M	Mug-Mug	Mug-P	Stove-M	Bowl-C	Mug-M	Average
BC	2.0±2.0	18.0±3.5	20.7±4.2	3.3±1.2	0.0±0.0	0.0±0.0	12.7±3.1	68.0±5.3	77.3±4.2	24.7±5.8	22.7±1.1
Random	10.0±5.3	18.0±3.5	72.0±9.2	6.7±2.3	0.7±1.2	11.3±2.3	8.0±5.3	58.7±4.2	44.0±2.0	30.7±9.2	26.0±3.7
All Data	16.7±1.2	19.3±3.1	81.3±6.1	11.3±2.3	3.3±4.2	12.7±3.1	6.7±2.3	62.7±6.4	44.7±3.1	31.3±11.7	29.0±2.4
Language	16.7±8.1	20.0±5.3	88.7±5.0	**14.7±8.1**	2.7±3.1	16.0±3.5	10.7±2.3	68.0±5.3	45.3±9.0	32.7±6.1	31.5±1.1
FR	16.0±5.3	20.0±0.0	76.0±7.2	9.3±5.0	0.7±1.2	14.0±0.0	11.3±5.8	63.3±6.1	46.7±7.6	25.3±10.3	28.3±1.6
STRAP	11.3±1.2	32.0±4.0	68.0±7.2	12.0±3.5	6.7±1.2	17.3±1.2	16.7±1.2	57.3±4.6	58.0±5.3	40.0±11.1	31.9±2.1
BR	17.3±4.2	**36.7±4.6**	**95.3±2.3**	9.3±2.3	3.3±1.2	17.3±2.3	8.7±2.3	64.0±3.5	50.7±9.2	29.3±2.3	33.2±1.5
IWR	14.7±3.1	29.3±6.4	88.7±3.1	11.3±3.1	2.7±3.1	14.7±4.2	11.3±1.2	70.7±3.1	54.7±11.7	27.3±1.2	32.5±1.0
HSR	**18.0±5.3**	32.0±9.2	90.7±1.2	12.7±5.0	**15.3±3.1**	**18.0±3.5**	**26.0±2.0**	**73.3±4.2**	**81.3±11.0**	**67.3±18.1**	**43.5±3.3**

we design long-horizon manipulation tasks executed on a 7-DoF xArm robot. All experiments are conducted using a single Nvidia RTX 5090 GPU.

Simulation. We evaluate on the LIBERO long-horizon subset, which consists of 10 composite manipulation tasks requiring diverse skills such as pick-and-place and drawer closing. For each task, we construct a target dataset $\mathcal{D}_{\text{t}}$ containing 5 demonstrations. Following [11], [14], we use LIBERO-90 as $\mathcal{D}_{\text{prior}}$, which includes 90 tasks with 50 demonstrations per task. We retrieve the top 10% of episodes from $\mathcal{D}_{\text{prior}}$ in the first stage and retain the top 30% of these candidates after reranking in the second stage, corresponding approximately to 3% of frames from $\mathcal{D}_{\text{prior}}$. We evaluate the learned policy over 50 episodes using 3 random seeds and report the average success rate with standard deviation.

Real-World. We design several real-world manipulation tasks with varying levels of complexity, including *Drawer*, *Trashcan*, *Cup-Drawer*, and *Cup-Tea* (Fig. 4). For each task, we collect 20 demonstrations as $\mathcal{D}_{\text{t}}$. We randomly sample 10k trajectories from the DROID dataset [25] as $\mathcal{D}_{\text{prior}}$. In the first stage, we use the reranked top 3% of frames from $\mathcal{D}_{\text{prior}}$ for skill pretraining. In the second stage, we finetune the policy only on the target dataset $\mathcal{D}_{\text{t}}$, as the cross-embodiment gap between the source data and our xArm platform makes direct co-training less effective. We evaluate the learned policy over 20 trials per task and report the success rate. Each trial uses a different initial state, while the set of states is identical across models for fair comparison.

Baselines. We compare HSR with baselines covering target-only finetuning, naive data mixing, language-based retrieval, optical-flow retrieval, subtrajectory retrieval, and state-action representation-based retrieval.

Behavior Cloning (BC) finetunes the VLA model using only the target dataset $\mathcal{D}_{\text{t}}$.

Random randomly samples 10% of the data from the prior dataset $\mathcal{D}_{\text{prior}}$ and co-trains the VLA model with the target dataset $\mathcal{D}_{\text{t}}$.

All Data co-trains the VLA model on the entire prior dataset $\mathcal{D}_{\text{prior}}$ together with the target dataset $\mathcal{D}_{\text{t}}$.

Language performs simple language-similarity-based retrieval and co-trains the VLA model on the retrieved dataset $\mathcal{D}_{\text{r}}$ and the target dataset $\mathcal{D}_{\text{t}}$.

Flow Retrieval (FR) [26] retrieves data based on similarity computed from samples' optical flows.

STRAP [14] employs vision foundation models and dynamic time warping to retrieve sub-trajectories from $\mathcal{D}_{\text{prior}}$.

Behavior Retrieval (BR) [10] trains a VAE to encode state-action pairs and retrieves data from $\mathcal{D}_{\text{prior}}$ based on similarity in the learned latent space.

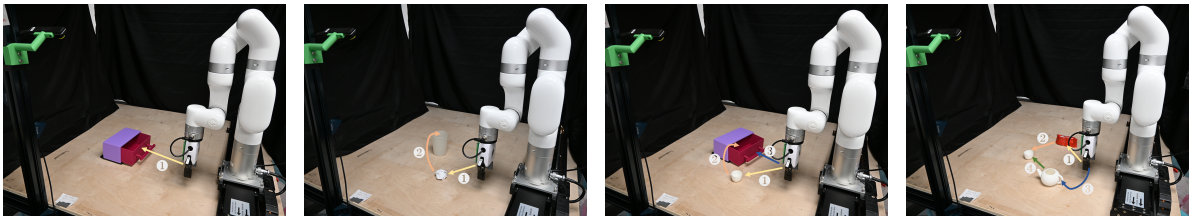
Importance Weighted Retrieval (IWR) [11] uses the same VAE encoder but retrieves data from $\mathcal{D}_{\text{prior}}$ based on estimated importance weights.

A. Simulation Experiments

Table I summarizes the results on the LIBERO benchmark. Finetuning on the target dataset alone (BC) achieves limited performance due to the small number of demonstrations. Random and All Data achieve only modest gains despite using additional prior data, suggesting that simply increasing data volume is insufficient when the prior dataset contains redundant or task-irrelevant demonstrations.

Language-based retrieval improves over random sampling, but remains limited on tasks with more complex instructions, suggesting that language similarity alone is insufficient to capture fine-grained semantics. State-action-based retrieval methods, including BR and IWR, further improve performance by leveraging trajectory-level similarity. However, their gains are smaller on composite manipulation tasks, where higher-level task structure matters more. STRAP is the closest baseline to HSR, as it also performs retrieval at the subtrajectory level. Nevertheless, its vision-based similarity criterion is less effective at capturing the semantic structure of long-horizon tasks, which limits its performance on tasks that require reusable skills.

Overall, HSR retrieves more useful demonstrations and achieves the best average performance. On tasks that mainly require a single *pick-and-place* skill, HSR performs comparably to BR, indicating that directly retrieving highly similar demonstrations can already be sufficient in simpler settings. In contrast, the advantage of HSR becomes more evident on complex composite tasks. For example, on Mug-M, HSR outperforms the strongest competing method on this task, STRAP, by 27.3%. On the challenging Bowl-C task, it is the only method that improves over pure behavior cloning. These



(a) Drawer (b) Trashcan (c) Cup-Drawer (d) Cup-Tea

Fig. 4: **Experimental Setup of Real-robot Tasks**. We evaluate using a 7-DoF xArm on several manipulation tasks: **(a) Drawer**: Close the drawer, **(b) Trashcan**: Pick up the crumpled paper and throw it in the trash can, **(c) Cup-Drawer**: Put the cup into the drawer and close it, **(d) Cup-Tea**: Put the teabag next to the cup and pour water into the cup.

TABLE II: **Experimental Results on Real-robot Manipulation**. For each task, we report the success rate over 20 trials. We also report the overall success rate across all tasks. The best results are in **bold**.

Task #Subtask	Drawer 1	Trashcan 2	Cup-Drawer 3	Cup-Tea 4	Overall
BC	12/20	7/20	5/20	3/20	27/80
STRAP	11/20	4/20	8/20	6/20	29/80
BR	15/20	3/20	4/20	5/20	27/80
IWR	16/20	4/20	7/20	4/20	31/80
HSR	**18/20**	**8/20**	**13/20**	**9/20**	**48/80**

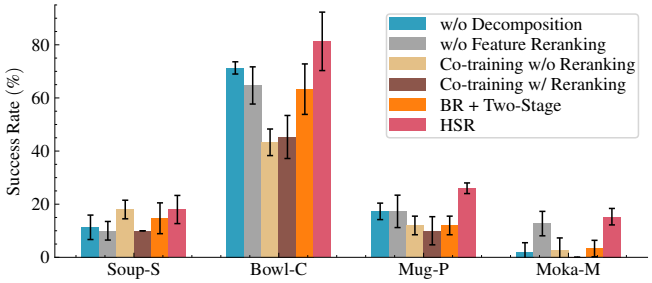


Fig. 5: **Results of Ablation Studies.** We conduct ablation studies on four LIBERO tasks to evaluate the effectiveness of key components of HSR, including task decomposition, feature reranking, and the pretraining-finetuning strategy. The results show that each component improves performance, with the full HSR framework achieving the highest success rates across all tasks.

Fig. 6: Performance under **Varying Reranking Retention Ratios**.

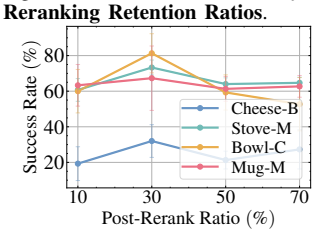


TABLE III: **Ablation on Task Decomposition**.

Method	Avg. Succ. (%)
LLM Only	37.1±2.5
LLM+Skill Set	37.7±1.2
HSR	43.5±3.3
Human	44.9±1.0

results demonstrate the effectiveness of HSR in selecting useful prior data and enabling effective policy adaptation.

B. Real-World Experiments

For real-robot tasks, data scarcity poses a significant challenge. As shown in Table II, BC trained only on target demonstrations achieves a success rate below 40 %. STRAP, BR, and IWR provide only limited improvement, suggesting that directly transferring prior data remains challenging under the substantial cross-embodiment gap between the source data and the target tasks. In contrast, HSR improves performance by retrieving relevant demonstrations in a structured manner. As a result, HSR achieves higher success rates across real-robot tasks of varying horizons. Notably, the advantage of HSR becomes more pronounced on long-horizon tasks. For the *"make a cup of tea"* task, which requires several hundred low-level actions to complete, BC trained only on the target demonstrations succeeds in 3 out of 20 trials, whereas HSR achieves 9 out of 20 successful executions, demonstrating a substantial improvement on complex multi-step manipulation.

C. Ablation Studies

Task Decomposition. We remove the task decomposition module and directly retrieve data using the original task instruction, while keeping the rest of the framework unchanged. As shown in Fig. 5, performance drops noticeably, especially on tasks with more complex instructions such as Mug-P and Moka-M. This suggests that using the full task description alone is often insufficient to retrieve demonstrations required by long-horizon manipulation. In contrast, task decomposition breaks a complex instruction into semantically meaningful subtasks, enabling more precise retrieval and improving overall performance.

We further compare different decomposition strategies in Table III, which reports the average success rate across all LIBERO-10 tasks. Raw LLM-generated plans already achieve competitive performance, indicating that LLMs can produce reasonable high-level decompositions. Incorporating the available skill set into the prompt further improves performance, suggesting that grounding the planner with executable skills leads to better task decomposition. Adding the plan evaluation module of HSR yields a further improvement, demonstrating the value of selecting decompositions that are both semantically relevant and well supported by the prior data. Manually selected plans achieve performance comparable to that of HSR, indicating that the proposed plan evaluation strategy can approach the quality of manually designed decompositions.

Feature Reranking. As shown in Fig. 5, removing the feature reranking module and using only language-retrieved data leads to a clear performance drop, especially on tasks that require precise alignment with the target execution context, such as Soup-S. This suggests that language similarity alone

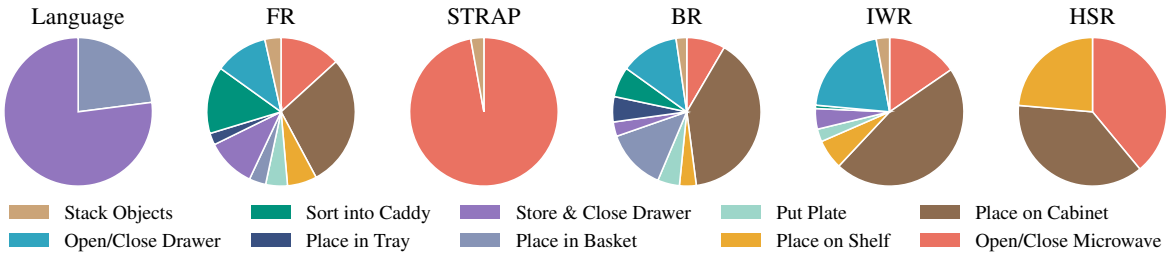


Fig. 7: **Comparison of Retrieved Data.** Distribution of retrieved data for the task *"put the yellow and white mug in the microwave and close it"*. HSR effectively retrieves demonstrations across subtasks covering both object-placing and microwave-closing skills.

is insufficient to capture low-level compatibility between the retrieved demonstrations and the target task. In contrast, the feature reranking module filters out demonstrations that are less compatible in behavior space and retains those that better match the target task, thereby improving policy adaptation.

We also ablate the retention ratio after feature reranking, with results shown in Fig. 6. The results indicate that retaining a moderate fraction of the reranked data yields the best performance on most tasks. Retaining too much data introduces additional noise, while retaining too little data may discard useful demonstrations.

Training method. Another way to leverage retrieved data is to co-train the policy on both the target and retrieved datasets in a single stage. We compare this co-training baseline with our proposed two-stage pretraining-finetuning strategy. As shown in Fig. 5, both co-training variants, with and without feature reranking, achieve lower success rates on most tasks. One possible reason is that retrieved demonstrations are typically larger in scale, noisier, and less well aligned with the target task than the target demonstrations. Co-training may therefore bias optimization toward the retrieved data, which can be suboptimal when the gap between the retrieved and target datasets is large. In contrast, our strategy first uses retrieved data to learn general skills and then finetunes on the most relevant demonstrations, which helps mitigate this issue and provides greater robustness to diverse data sources and embodiments.

We further evaluate two-stage adaptation by replacing our retrieval method with BR. This combined approach improves over the original BR baseline, but still underperforms HSR, further demonstrating the benefit of our retrieval strategy.

D. Retrieved Data Study

To illustrate the effect of different retrieval methods, we consider a representative task from the LIBERO benchmark, *"put the yellow and white mug in the microwave and close it."* The retrieved demonstrations are visualized in Fig. 7. The language-based retrieval baseline relies only on the original task description and may miss task-specific nuances. FR, STRAP, BR, and IWR incorporate additional motion- or behavior-level information, but may still retrieve demonstrations from tasks with similar motions and different goals. Such mismatched data can mislead policy learning. In contrast, HSR decomposes the long-horizon task into subtasks and retrieves demonstrations for each subtask independently. This allows the method to retrieve demonstrations that are more closely aligned with the specific requirements of each step, leading to more task-relevant retrieved data and improved performance.

VI. Conclusion

We present Hierarchical Skill Retrieval (HSR), a framework for data-efficient adaptation of VLA models to downstream manipulation tasks. By exploiting the hierarchical structure of long-horizon tasks, HSR combines high-level semantic decomposition with low-level behavior-aware retrieval to identify useful prior demonstrations. This allows the policy to reuse transferable skills from prior datasets, rather than relying on exact task-level matches. Experiments on the LIBERO benchmark and real-world long-horizon manipulation tasks show that HSR consistently improves adaptation performance under limited target data.

Discussion and Limitations. Despite these promising results, several challenges remain. First, the current framework is evaluated only on manipulation tasks and a single VLA backbone. Evaluating HSR across different policy architectures, robot embodiments, and data modalities, such as videos or images collected in diverse domains, remains an important direction for improving robustness and data efficiency. Second, although we use an LLM for task decomposition, tighter integration of language models into closed-loop control, such as online replanning and failure recovery, remains an important direction for future work.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763. 1, 2

[2] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *International Conference on Computer Vision*, 2023, pp. 11 975–11 986. 1, 2

[3] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024. 1, 2

[4] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," in *9th Annual Conference on Robot Learning*, 2025. 1

[5] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024. 1, 2

[6] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6892–6903. 1, 2

[7] S. Belkhale, T. Ding, T. Xiao, P. Sermanet, Q. Vuong, J. Tompson, Y. Chebotar, D. Dwibedi, and D. Sadigh, "Rt-h: Action hierarchies using language," *arXiv preprint arXiv:2403.01823*, 2024. 1, 2

[8] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, S. Alibert, M. Cord, T. Wolf, and R. Cadene, "Smolvla: A vision-language-action model for affordable and efficient robotics," *arXiv preprint arXiv:2506.01844*, 2025. 1, 2, 4

[9] P. Li, Y. Wu, Z. Xi, W. Li, Y. Huang, Z. Zhang, Y. Chen, J. Wang, S.-C. Zhu, T. Liu, and S. Huang, "ControlVLA: Few-shot object-centric adaptation for pre-trained vision-language-action models," in *9th Annual Conference on Robot Learning*, 2025. 1

[10] M. Du, S. Nair, D. Sadigh, and C. Finn, "Behavior retrieval: Few-shot imitation learning by querying unlabeled datasets," in *Robotics: Science and Systems*, 2023. 1, 2, 4, 5

[11] A. Xie, R. Chand, D. Sadigh, and J. Hejna, "Data retrieval with importance weights for few-shot imitation learning," in *Proceedings of The 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Lim, S. Song, and H.-W. Park, Eds., vol. 305. PMLR, 27–30 Sep 2025, pp. 1–16. 1, 2, 4, 5

[12] S. Kumar, S. Dass, G. Pavlakos, and R. Martín-Martín, "Collage: Adaptive fusion-based retrieval for augmented policy learning," in *Proceedings of The 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Lim, S. Song, and H.-W. Park, Eds., vol. 305. PMLR, 27–30 Sep 2025, pp. 4607–4624. 1, 2

[13] S. Nasiriany, T. Gao, A. Mandlekar, and Y. Zhu, "Learning and retrieval from prior data for skill-based imitation learning," in *Conference on Robot Learning (CoRL)*, 2022. 1, 2

[14] M. Memmel, J. Berg, B. Chen, A. Gupta, and J. Francis, "STRAP: Robot sub-trajectory retrieval for augmented policy learning," in *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 5

[15] L. X. Shi, B. Ichter, M. R. Equi, L. Ke, K. Pertsch, Q. Vuong, J. Tanner, A. Walling, H. Wang, N. Fusai, A. Li-Bell, D. Driess, L. Groom, S. Levine, and C. Finn, "Hi robot: Open-ended instruction following with hierarchical vision-language-action models," in *Forty-second International Conference on Machine Learning*, 2025. 1

[16] S. Yang, H. Li, Y. Chen, B. Wang, Y. Tian, T. Wang, H. Wang, F. Zhao, Y. Liao, and J. Pang, "Instructvla: Vision-language-action instruction tuning from understanding to manipulation," *arXiv preprint arXiv:2507.17520*, 2025. 1

[17] J. Barreiros, A. Beaulieu, A. Bhat, R. Cory, E. Cousineau, H. Dai, C.-H. Fang, K. Hashimoto, M. Z. Irshad, M. Itkina, *et al.*, "A careful examination of large behavior models for multitask dexterous manipulation," *arXiv preprint arXiv:2507.05331*, 2025. 1

[18] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024. 1

[19] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 44 776–44 791. 2, 4

[20] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019. 2

[21] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 34 892–34 916. 2

[22] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang, *et al.*, "Spatialvla: Exploring spatial representations for visual-language-action model," *arXiv preprint arXiv:2501.15830*, 2025. 2

[23] T. Zhang, H. Duan, H. Hao, Y. Qiao, J. Dai, and Z. Hou, "Grounding actions in camera space: Observation-centric vision-language-action policy," *arXiv preprint arXiv:2508.13103*, 2025. 2

[24] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022. 2

[25] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024. 2, 5

[26] L.-H. Lin, Y. Cui, A. Xie, T. Hua, and D. Sadigh, "Flowretrieval: Flow-guided data retrieval for few-shot imitation learning," in *8th Annual Conference on Robot Learning*. 2, 5

[27] S. N. Syed, Y. Ahuja, A. Jakobsson, and J. Ichnowski, "Expres-vla: Specializing vision-language-action models through experience replay and retrieval," *arXiv preprint arXiv:2511.06202*, 2025. 2

[28] L. Zha, Y. Cui, L.-H. Lin, M. Kwon, M. G. Arenas, A. Zeng, F. Xia, and D. Sadigh, "Distilling and retrieving generalizable knowledge for robot manipulation via language corrections," in *2024 IEEE International Conference on Robotics and Automation*, 2024. 2

[29] K. Dreczkowski, P. Vitiello, V. Vosylius, and E. Johns, "Learning a thousand tasks in a day," *Science Robotics*, vol. 10, no. 108, p. eadv7594, 2025. 2

[30] L. P. Kaelbling and T. Lozano-Pérez, "Hierarchical task and motion planning in the now," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 1470–1477. 2

[31] H. Sahni, S. Kumar, F. Tejani, and C. Isbell, "Learning to compose skills," *arXiv preprint arXiv:1711.11289*, 2017. 2

[32] M. Dalal, M. Liu, W. Talbott, C. Chen, D. Pathak, J. Zhang, and R. Salakhutdinov, "Local policies enable zero-shot long-horizon manipulation," *International Conference on Robotics and Automation*, 2025. 2

[33] L. P. Kaelbling and T. Lozano-Pérez, "Learning composable models of parameterized skills," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 886–893. 2

[34] W. Liu, N. Nie, R. Zhang, J. Mao, and J. Wu, "Learning compositional behaviors from demonstration and language," in *8th Annual Conference on Robot Learning*. 2

[35] J. Mao, T. Lozano-Pérez, J. Tenenbaum, and L. Kaelbling, "Pdsketch: Integrated domain programming, learning, and planning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 972–36 984, 2022. 2

[36] J.-J. Shao, H.-R. Hao, X.-W. Yang, and Y.-F. Li, "Abductive learning for neuro-symbolic grounded imitation," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 2025, pp. 1221–1232. 2

[37] 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. 2

[38] Y. Jin, D. Li, J. Shi, P. Hao, F. Sun, J. Zhang, B. Fang, *et al.*, "Robotgpt: Robot manipulation learning from chatgpt," *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2543–2550, 2024. 2

[39] M. Li, S. Zhao, Q. Wang, K. Wang, Y. Zhou, S. Srivastava, C. Gokmen, T. Lee, E. L. Li, R. Zhang, *et al.*, "Embodied agent interface: Benchmarking llms for embodied decision making," *Advances in Neural Information Processing Systems*, vol. 37, pp. 100 428–100 534, 2024. 2

[40] X. Li, M. Zhang, Y. Geng, H. Geng, Y. Long, Y. Shen, R. Zhang, J. Liu, and H. Dong, "Manipllm: Embodied multimodal large language model for object-centric robotic manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 061–18 070. 2

[41] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *6th Annual Conference on Robot Learning*, 2022. 2, 3

[42] Y. Meng, X. Yao, H. Ye, Y. Zhou, S. Zhang, Z. Bing, and A. Knoll, "Data-agnostic robotic long-horizon manipulation with vision-language-guided closed-loop feedback," 2025. 2

[43] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 523–11 530. 2

[44] S. Dass, A. Khaddaj, L. Engstrom, A. Madry, A. Ilyas, and R. Martín-Martín, "DataMIL: Selecting data for robot imitation learning with datamodels," in *The Fourteenth International Conference on Learning Representations*, 2026. 4

[45] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, *et al.*, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025. 4

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186. 4