Retrieval-Augmented Embodied Agents

Yichen Zhu, Zhicai Ou, Xiaofeng Mou, Jian Tang*
Midea Group, AI Lab
{zhuyc25, zhicai.ou, mouxf, jiantang22}@midea.com

Abstract

Embodied agents operating in complex and uncertain environments face considerable challenges. While some advanced agents handle complex manipulation tasks with proficiency, their success often hinges on extensive training data to develop their capabilities. In contrast, humans typically rely on recalling past experiences and analogous situations to solve new problems. Aiming to emulate this human approach in robotics, we introduce the Retrieval-Augmented Embodied Agent (RAEA). This innovative system equips robots with a form of shared memory, significantly enhancing their performance. Our approach integrates a policy retriever, allowing robots to access relevant strategies from an external policy memory bank based on multi-modal inputs. Additionally, a policy generator is employed to assimilate these strategies into the learning process, enabling robots to formulate effective responses to tasks. Extensive testing of RAEA in both simulated and real-world scenarios demonstrates its superior performance over traditional methods, representing a major leap forward in robotic technology.

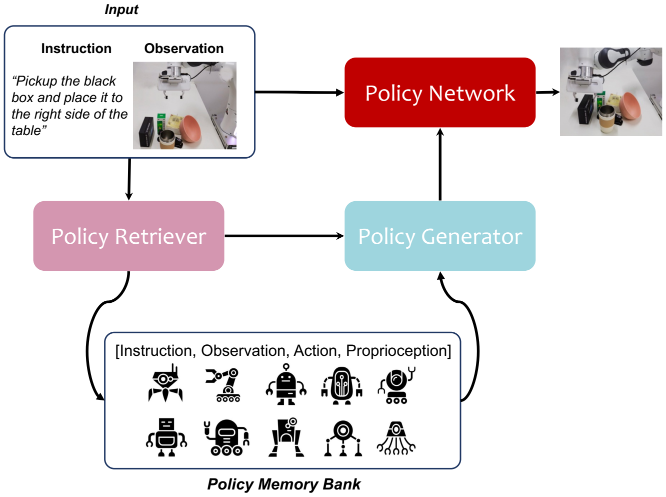
Input
Instruction
Observation
"Pickup the black box and place it to the right side of the table"
Policy Network
Policy Retriever
Policy Generator
[Instruction, Observation, Action, Proprioception]
Policy Memory Bank

Figure 1. The overview of our retrieval-augmented embodied agents. We utilize a policy retriever to extract policies from a policy memory bank, which contains large-scale robotic data across multiple embodiments. Then, we use the policy generator to reference the retrieved policy and output actions for the current input.

1. Introduction

The swift advancement of foundation models in areas like natural language processing and computer vision has sparked interest in the robotics community to create embodied agents capable of comprehending human instructions and responding aptly to their environment. Despite this enthusiasm, crafting agents that seamlessly interact with the physical world remains a formidable task. deep neural networks store knowledge—such as recognizing objects or interpreting commands—implicitly within their neural network parameters. This dependence on implicit knowledge storage demands a significant number of parameters and a wide range of training data. However, recent studies have shown that the scalability in terms of both training data and model size falls short [4, 77] when compared to foundation models in other domains, such as Large Language Models. This insight has inspired the creation of embodied agents designed to learn efficiently with limited data and model sizes. To augment their capabilities, it's becoming increasingly important for these agents to access external repositories of physical knowledge, thereby expanding their capacity to understand and interact with the world.

The ability to tap into an external repository of behavioral memory mirrors the learning process observed in human infants, who often remember and mimic the actions of adults or animals when presented with analogous scenarios from their memory. This ability is crucial for successfully navigating unknown environments and performing tasks that demand specific knowledge, like exploring unfamiliar rooms or handling new objects. Consequently, the question naturally arises: How can we harness the wealth of open-source, multi-embodiment data to enhance the precision of robots in manipulation tasks? This inquiry not only probes

[*] Corresponding author

the potential of robotic learning but also seeks to bridge the gap between human cognitive processes and robotic applications.

In this paper, we present Retrieval-Augmented Embodied Agents (RAEA), which leverage an external policy memory bank containing analogous scenarios, whether in terms of instructions, observations, or a combination of both, related to the ongoing task. We outline the overview of our framework in Figure 1. The policies retrieved from this memory, along with other relevant data, provide a rich resource for both the training and testing phases. In our methodology, we make use of the recently open-sourced Open X-Embodiment [46], a large-scale repository of robotic datasets. Open X-Embodiment contains an extensive array of tasks, applications, embodiments, and diverse environmental settings from various research labs. This extensive dataset serves as the cornerstone for our external policy banks, enriching the knowledge base of RAEA. By tapping into this vast repository, RAEA can access a broader spectrum of robotic experiences, thereby improving its adaptability and effectiveness in various tasks.

To realize our objectives, we introduce two innovative modules: a policy retriever and a policy generator. The policy retriever is adept at handling multiple input modalities, categorized into two main types: instructions and observations. For controlling robots, it accepts text and audio as instructional inputs, and images, videos, and point clouds as observational inputs. It identifies policy candidates from the memory bank that align closely with the current input. Furthermore, we have developed the policy generator, which initially processes the information in the retrieved policies, including observation, instruction, action, and proprioceptive state. It then employs a cross-attention module to integrate knowledge from various retrieved policies into the main policy networks for action prediction. In this way, the policy generator leverages the retrieved policies as contextual examples, aiding the model in producing actionable responses based on the current input.

The efficacy of our proposed Retrieval-Augmented Embodied Agent (RAEA) is demonstrated through extensive testing over two simulation benchmarks and real-world datasets, as illustrated in Figure 2. This approach not only showcases the versatility and practical ability of our framework.

In summary, our contributions are as follows:

- We present Retriveal-Augmented Embodied Agents (RAEA) that utilized the wealth of knowledge from an external policy memory bank with multiple embodiment data to facilitate prediction for robotic action.
- Our framework features a policy retriever adept at processing various input modalities. Complementing this, we have crafted a policy generator that leverages retrieved scenarios to improve the model's ability to generalize across various situations.
- To validate the efficacy of our proposed methodology, we have conducted extensive evaluations in both real-world settings and two simulated environments. The results from these tests strongly affirm the effectiveness and practicality of our approach.

Overall, our work introduces a versatile and modular retrieval-augmentation framework for embodied agents. This provides a novel and insightful perspective on the design of robotic models, integrating advanced memory capabilities.

2. Related Works

Retrieval-Augmented Models. A notable trend in Natural Language Processing (NLP) involves leveraging external memory to enhance the performance of language models. This approach retrieves documents relevant to the input text from an external database, allowing language models (generators) to use this retrieved information to make more informed predictions. Typically, the external memory consists of a collection of text passages or a structured knowledge base [69–71]. Subsequent works extend the retrieval augmentation techniques to computer vision models and multi-modal models. The most representative works including Re-Imagen [12] as a caption-to-image generator, MuRAG [11] performs question-answering using retrieved images. RA-C3M [72] uses retrieval for either text or image generation.RA-CLIP [68] and REACT [37] integrate the retrieval-augmentation technique for CLIP pretraining. Our approach, however, diverges from these prior works. While the aforementioned studies focus on enhancing language and vision models, our research is specifically geared towards robotics. We aim to search for policies that have been executed in scenarios similar to the current context, using them as in-context examples.

Models for Embodied Agents. In the field of embodied agents [1–3, 6, 10, 13, 14, 28, 30, 39, 42, 44, 49, 56–59, 62, 65] and robotics [8, 9, 18, 25, 45, 47, 63, 66, 75, 76], foundation models have become a crucial research focus, revolutionizing the interaction between AI systems and physical environments. This body of work includes studies on representation pre-training and the application of language and vision-language models [27, 29, 30, 39, 42, 55, 61] as embodied agents. Our research contributes to this growing body of knowledge, presenting a supplementary retrieval-based approach designed to enhance policy learning in robotics. This approach integrates with existing foundation models, providing a novel perspective on how to augment the capabilities of embodied agents in diverse and dynamic settings.

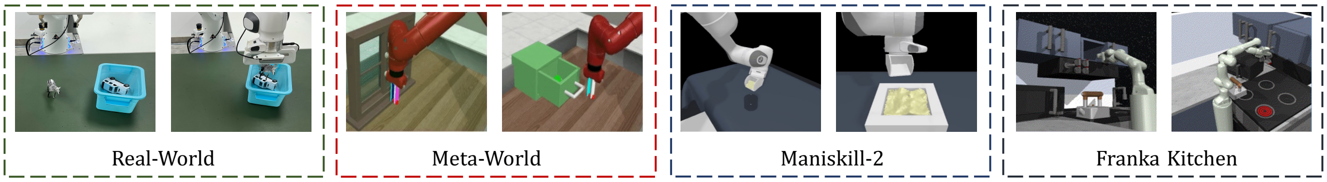

Figure 2. Examples of simulated and real-world environments that we used for evaluation.

Robotics Datasets. The robotics learning community has developed a variety of open-source datasets that are instrumental in advancing robot learning. These tasks spanning grasping [5, 16, 19, 20, 32, 35, 40], pushing interactions [17, 60, 73], sets of objects and models [75–85], and teleoperated demonstrations [8, 86–95]. Typically, these datasets are extensive and are often focused on specific robotic embodiments, exemplified by the Bridge Data [18, 34, 64] and RH20T [21]. RoboNet [15] and Open X-Embodiment [46] stand out as two large-scale datasets that incorporate multiple robotic embodiments. These datasets are frequently utilized for pretraining purposes [46], especially for the visual backbone in robotic models [7, 43, 53]. In our research, we leverage the extensive cross-embodiment data as a foundational knowledge base. This enables us to retrieve relevant policies that facilitate training in the current environment, effectively utilizing the rich diversity of the datasets to enhance our model's adaptability and performance.

3. Methodology

We introduce Retrieval-Augmented Embodied Agents, capable of retrieving relevant scenarios and generating actions based on the current scene and accompanying instructions. As illustrated in Figure 3, when presented with an input, be it an observation or an instruction-observation pair, our system employs a retriever to fetch pertinent policies from an external memory bank. Significantly, these embodied agents are equipped to interact with humans through various modalities like text and audio, and they use diverse sensors to perceive their environment. In order to broaden the spectrum of applications for our approach, we have designed a multi-modal policy retriever (in §3.2), featuring a dense retriever with a mixed-modal encoder capable of encoding diverse modalities in various combinations. Additionally, we've constructed a policy generator (in §3.3)based on the Transformer architecture. This generator processes the retrieved policies individually and leverages cross-attention to incorporate the extracted information from the retrieved policies into the primary model branch.

3.1. Preliminaries

Notations. The framework consists of a policy retriever R and a policy generator module G. The retrieval module R takes an input sequence $r = \{i, o\}$ and searches the r from an external policy memory bank. It returns a list of policy $m = \{i, o, a, p\}$, where p is the policy, i represents the instruction, o denotes the observation, a is the action, p is the proprioception. The term proprioceptive robot state is used to describe a robot's intrinsic awareness of its own positioning and movement within a given space, which includes factors like joint angles, velocity, torque, and other physical statuses. The term actions refers to the specific tasks or operations executed by the robot. The policy generator G then takes the input sequence x and the retrieved policy $M = \{m_1, m_2, \cdots, m_n\}$ and returns the action a, where a represent continuous actions that control the robots.

3.2. Policy Retriever

Overview. A policy retriever R takes a query q (e.g., the instruction-observation pair x) with multi-modal information from the policy memory bank M, and returns a relevance score $r(q, m)$. We follow prior retrieval works [33], in which the retriever r is a bi-encoder architecture,

$$r(q, m) = E_Q(q)^T E_M(m) \tag{1}$$

Here, we employ two key encoders: E_Q, responsible for encoding queries, and E_M, which encodes the memory to yield dense vectors representing query and memory policies, respectively. Given that our input and memory consist of multi-modal documents, we leverage E_Q and E_M as mixed-modal encoders capable of handling various modalities. The architecture of mixed-modal encoders can be designed in multiple ways. In our specific context, we introduce a multi-modal retrieval approach that adeptly accommodates multiple modalities. When dealing with a multi-modal input for policy learning, we partition it into two distinct components: an instruction segment and an observation segment. The instruction segment typically contains human instructions in different formats, including text or audio, while the observation segment can encompass images, videos, or point cloud data. Each modality type is separately encoded using off-the-shelf, pre-trained multi-modality encoders, which we will discuss specifically in the next section.

Multi-Modal Encoders. The primary aim of multi-modal encoders is to equip embodied agents with the

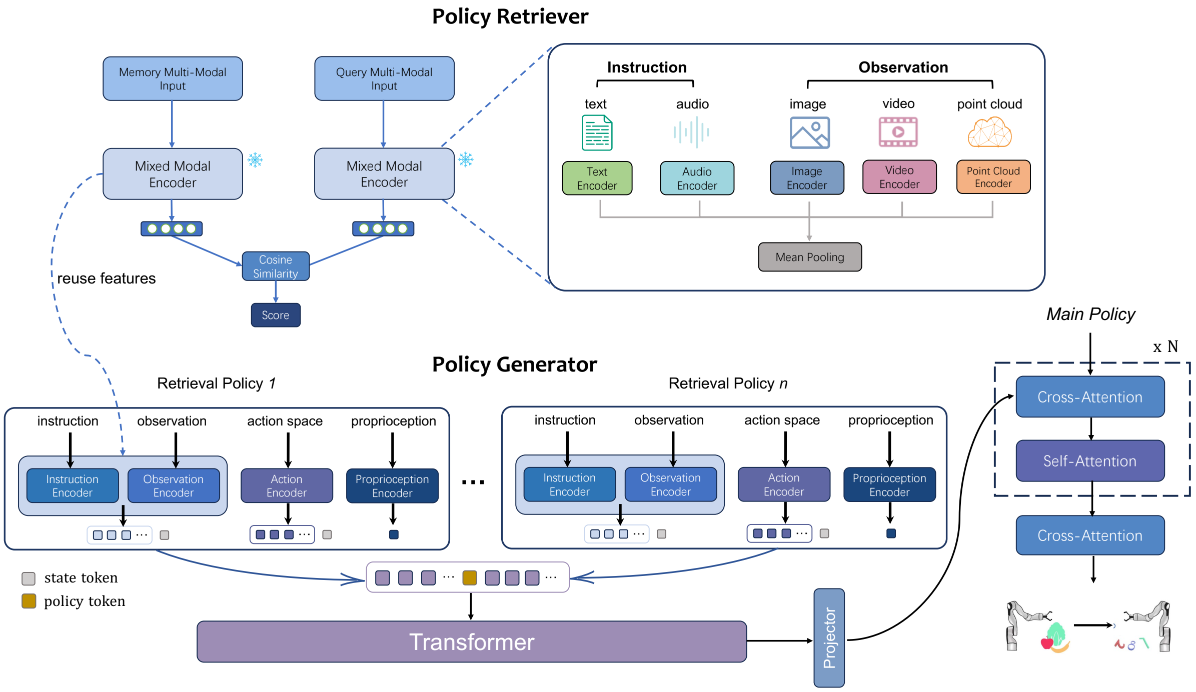

Figure 3. The framework of policy retriever (top) and policy generator (bottom) in our work. The policy retriever retrieves the relevant policy based on multi-modal input, and the policy generator processes a list of retrieved policies to help train in the current environment.

capability to handle a wide range of modalities, adapting to various scenarios. Specifically, we conceptualize the encoder as a mapping function, denoted by $P(\cdot)$. Typically, for each modality, a specialized model is needed to extract useful modality-specific representations. These are then projected onto a feature plane, ensuring uniformity in the shape of feature tensors across all modalities through projection layers. Following this, we average the vectors representing these modalities, normalizing their L2 norms to 1, thus generating a consolidated vector representation of the document. This encoding technique is uniformly applied to both E_Q and E_M. The final step involves assessing the similarity of cross-modality features between the query and the memory.

There are a number of options [67] for the multi-modal model when it comes to handling different modalities. For example, CLIP [52] or T5 [54] could be used for text and image processing. In our case, we utilize ImageBind [23], a high-performance encoder proficient across six modalities, for processing diverse input types. With the help of ImageBind, we are spared from managing many numbers of heterogeneous modal encoders. This mapping function $P(\cdot)$ maps all modalities to a unified latent embedding, greatly enhancing the efficiency of comparing feature similarity for retrieval purposes. For the retrieval process, we execute the Maximum Inner Product Search within the memory space, yielding a ranked list of candidates based on their relevance scores. From this list, we select the final k policies for further analysis and processing.

Retrieval Strategy. We discuss three key factors in obtaining/sampling informative retrieved policies for the generator in practice.

Relevance: The retrieval of policies must be closely aligned with the input sequence, covering either instructions, observations, or both. Without this alignment, the retrieved policies fail to provide meaningful contributions to the modeling of the primary input sequence. To ascertain the relevance of these policies, we employ a dense retriever score based on modality encoders.

Input Modality: Our methodology divides the input into two distinct segments: instructions and observations, each tailored to process specific input modalities. The observation component is adept at recalling comparable scenes or objects, thereby enabling the model to comprehend scenarios not encountered in its training dataset. Concurrently, the instruction segment recapitulates actions previously executed during training. This flexible framework allows for either independent or combined usage of these segments, contingent on the computational resources and specific application contexts. This approach bears a resemblance to in-context learning, wherein the

Large Language Model (LLM) is presented with scenarios that are similar, albeit not identical, to enhance response quality. Typically, we employ instruction-observation pairs at the onset of a frame, subsequently relying solely on the observation for the remainder of the frame until the next user interaction. It's noteworthy that this strategy can be integrated with recent advancements in Large Multi-Modal Models, facilitating real-time corrections of the robot's actions.

Diversity: We discovered that diversity in the policy memory bank is crucial for effective performance. Selecting the top-ranked actions based on relevance scores often leads to the inclusion of duplicate or very similar instruction and observation. This redundancy can detrimentally impact the performance of the generator. This challenge is particularly acute given that our action bank comprises fragments of videos. To optimize the efficacy of retrieval-based methods, it's essential to ensure a diverse range of in-context samples. These diverse samples are instrumental in aiding policy networks to effectively learn from demonstrations, both in training and testing stages. To address the issue of redundancy, our approach involves bypassing a candidate action if its relevance score is too closely aligned (e.g., exceeding 0.9) with the query or actions already retrieved. Additionally, to further enhance diversity, we propose a unique strategy of random token dropout from the query used in retrieval, approximately 70% of tokens. This approach serves as a regularization mechanism during training and has been observed to significantly improve the generator's performance. We also include embodiment data of embodiment that are different from the robot that we used for evaluation. We observed that even with variations in embodiment, this diverse dataset still aids the policy network in its learning process.

Data Format of Retrieved Policy. For every policy retriever, our framework covers a set of elements m that either influence or result from control processes. Subsequently, we introduce our policy generator and demonstrate how it effectively utilizes this diverse spectrum of information for enhancing policy learning.

3.3. Policy Generator

The policy generator is designed to effectively utilize the valuable information in the retrieved policy to facilitate the training of the policy for the current input. We reuse the feature representation of instruction and observation from the policy retrieval network (as in §3.2), thus avoiding redundant computations that constitute over 95% of the total. For actions and proprioceptive states, we address the variability across different robots by setting a maximum limit for both, capped at nine. We employ an action encoder and a proprioception encoder – both comprising multi-layer perceptrons (MLPs) – to generate corresponding tokens. These tokens are then integrated with the instruction-observation tokens, ensuring a seamless and effective incorporation of the retrieved policy's data into the current policy training framework. We add a state token between different states, i.e., a learnable token between action and proprioception tokens, to split the data of two states. We use absolute position embedding to ensure the tokens are in order.

Given a list of retrieved policies $M = (m_1, ..., m_K)$, we concatenated these tokens based on the relevance score. We use absolute position embedding to maintain the order of tokenized representations. We concatenate their tokens according to their relevance scores. We employ a policy token to demarcate tokens from different policies. Once these tokens are combined, we utilize the Transformer architecture as the foundation for our retrieved policy processor. To effectively incorporate the retrieved policies M into the generator, we leverage cross-attention mechanisms. This approach allows for the integration of pertinent information from the retrieved models into the main network, enhancing the overall efficacy of the system.

Particularly, give an input sequence from retrieved polices, denoted as $F^r \in \mathbb{R}^{h\times w\times c}$, and other input sequences from main input $F^x \in \mathbb{R}^{h\times w\times c}$, for simplicity, we assume two tensors have the same size. The F^r is projected into a query (Q) and key (K), and F^x is projected into value (V). Thus, we formulate our cross-attention (SC) as follows:

$$Q_i = F^r W_i^Q \tag{2}$$

$$K_i = SC(F^x, r_i)W_i^K, V_i = SC(F^x, r_i)W_i^V, \tag{3}$$

$$V_i = V_i + P(V_i) \tag{4}$$

where $SC(\cdot, r_i)$ is a MLP layer for aggregation in the i^{th} head with the down-sampling rate of r_i, and $P(\cdot)$ is a depthwise convolutional layer for projection. Finally, we calculated the attention tensor by:

$$h_i = Softmax(\frac{Q_i K_i^T}{\sqrt{d_h}} V_i) \tag{5}$$

where d_h is the dimension. The streamlined design of the cross-attention mechanism ensures that valuable representations from retrieved policies are effectively incorporated into the main network, enhancing policy learning for the current input. To optimize this process, we implement behavior cloning, using mean squared loss as our primary optimization objective.

4. Experiments

We evaluate models using multiple simulated benchmarks, including Franka Kitchen [22], MetaWorld [74], and Maniskill-2 [24] and real-world environment. We show that

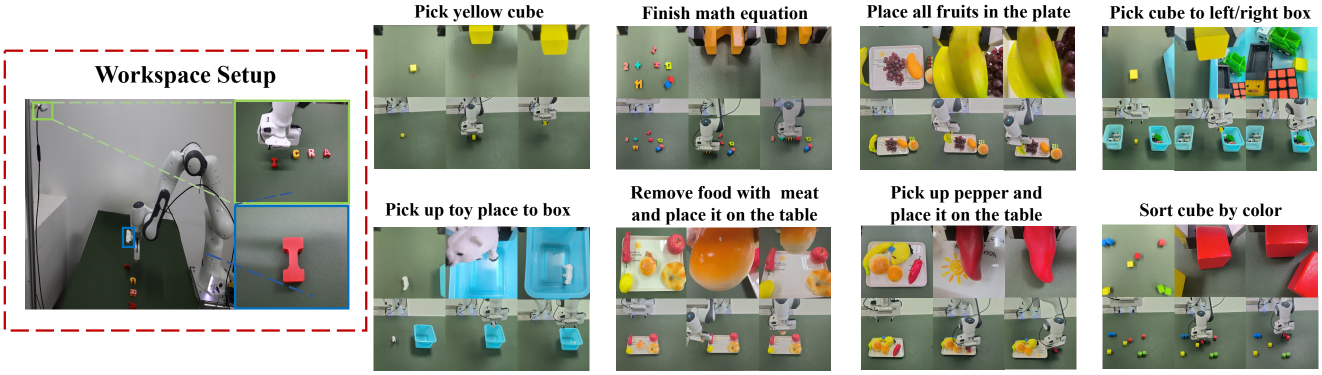
Workspace Setup
Pick yellow cube
Finish math equation
Place all fruits in the plate
Pick cube to left/right box
Pick up toy place to box
Remove food with meat and place it on the table
Pick up pepper and place it on the table
Sort cube by color

Figure 4. **Left:** The setup of our Franka real robot. **Right:** The example of some tasks that we collected.

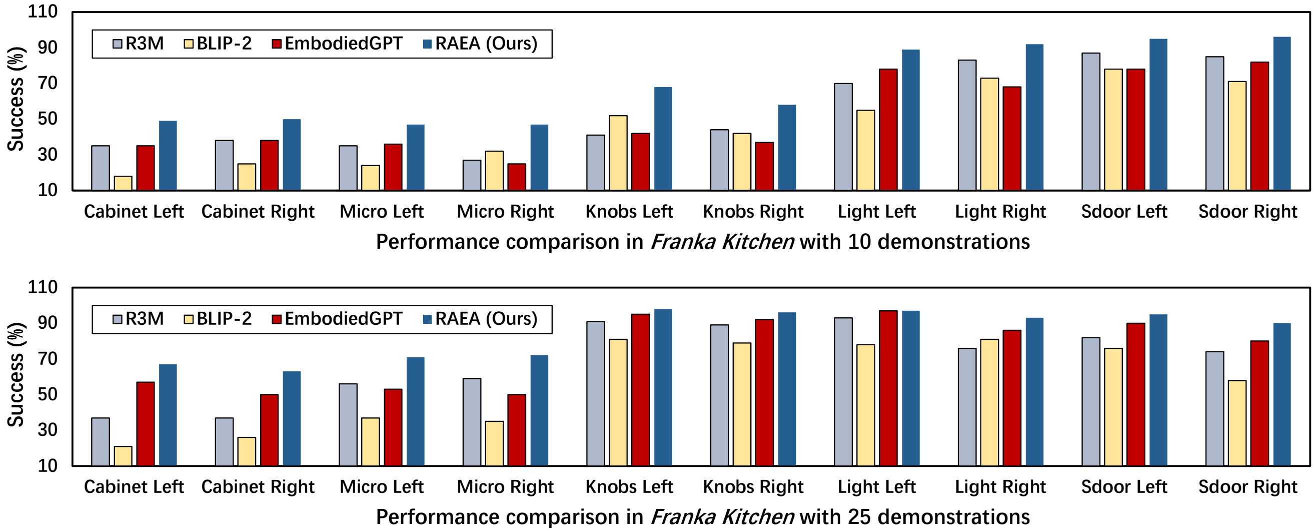

Figure 5. Performance of RAEA in Franka Kitchen with 10 or 25 demonstrations

our retrieval-augmented embodied agents significantly improve the generalization ability. Notably, we have taken precautions to ensure that the policy memory bank does not contain any data from the datasets used in our training and testing phases, thereby eliminating the possibility of bias or cheating in the test set.

4.1. Simulation Experiments.

We conduct our simulation on three benchmarks, Franka Kitchen [22], Meta-World [74], and Maniskill-2 [24]. A brief summary of these benchmarks can be found in the Appendix.

Evaluation: We assess our approach through 30 rollouts derived from the behavior cloning (BC) learned policy. Our primary metric for evaluation is the mean success rate of the final policy. Additionally, when presenting metrics for the task suite, we calculate the average mean success rate across various camera configurations.

Experimental results on Franka Kitchen & Meta-World. In our experiments, we conducted a comparative analysis of our model, RAEA, against two established state-of-the-art methods: R3M [43], popular in Franka Kitchen applications, and BLIP-2 [36], a leading vision-language model, and Embodied-GPT [41], a vision-language model designed for robotics. We trained our policy network using a few-shot learning approach, employing datasets comprising either ten or twenty-five demonstrations. The performance of these models was evaluated through 100 randomized trials across five distinct tasks in each benchmark. These evaluations were executed under two

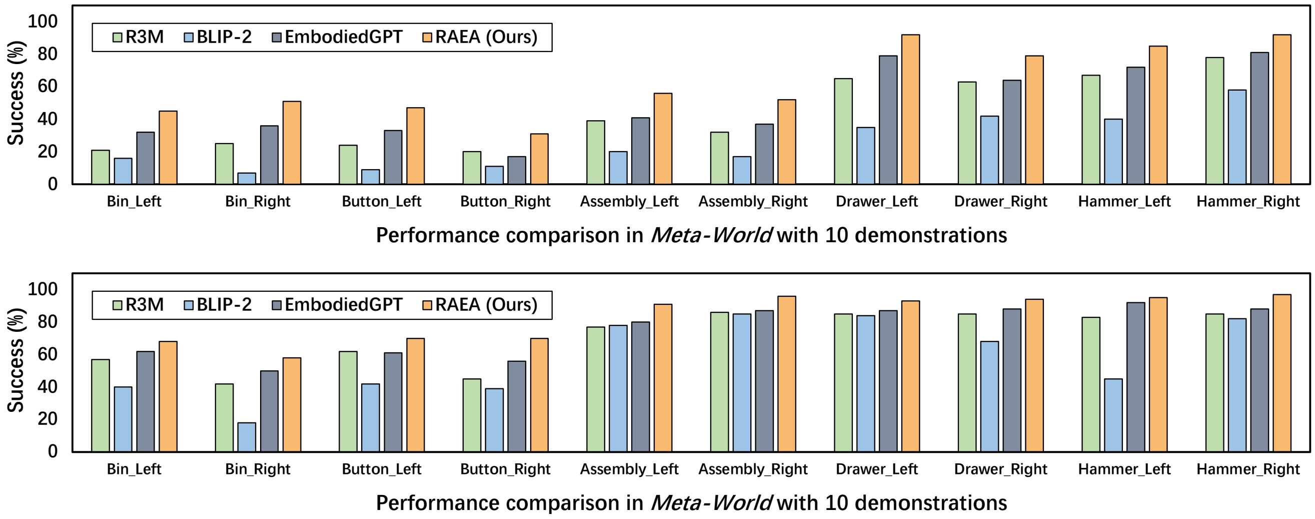

Figure 6. Performance of RAEA in Meta-World with 10 or 25 demonstrations

different settings: each involved five separate runs and was conducted from two unique camera perspectives, relying solely on visual observations. The results, depicted in Figures 5 and 6, for the Franka Kitchen and Meta-World benchmarks respectively, unequivocally demonstrate that RAEA surpasses the baseline methods in effectiveness. This superiority is particularly noticeable in low-data scenarios, such as those with only ten demonstrations, further underscoring RAEA's robustness and efficiency in environments with limited data availability.

Experimental results on Maniskill-2. We evaluated our model's performance in two distinct experimental settings: one using solely image-based observations and the other combining images with point cloud data. For comparison, we benchmarked our model against well-established methods, specifically ResNet152 [26] and Swin Transformer Base [38], both pre-trained on the ImageNet dataset. In the experiments that involved both images and point clouds, the feature representation of point cloud data for the baseline models was initially processed using PointNet [51]. This representation was subsequently integrated with the image branch, following the procedure described in Maniskill-2. In both experimental scenarios, our method consistently surpassed the performance of the two baseline models, as illustrated in Table 1. Remarkably, in the setting that utilized dual-modal observations, our RAEA model demonstrated a significantly higher average success rate. This highlights the robust generalizability and effectiveness of our approach.

4.2. Real-Robot Experiments

Datasets. Our collected dataset comprises $n = 60$ tasks. These tasks vary from straightforward pick-and-place actions, such as "pick up the yellow cube" to more complex contact-rich tasks like "open the drawer and put the pen inside," as well as tasks demanding to reason, i.e., "sort the cube with the same color." There are 70 objects in the experiments. Each task is exemplified through 30 human-collected trajectories. Further, every task is annotated with 5 distinct instructions. Figure 4 demonstrate some example and workspace setup for our real-world experiments.

Implementation details. We use an AdamW optimizer, starting with an initial learning rate of 3e-5, and implement a weight decay of 1e-6. Our learning rate scheduler is designed to linearly decay, incorporating a warm-up phase that spans the initial 2% of the total training duration. Additionally, we apply gradient clipping set at a value of 1.0 to maintain stability during training. To assess the efficacy of our approach, all experiments are conducted over 10 trials, from which we calculate the mean success rate.

Additionally, we perform ablation studies to delve into various questions related to our model's performance and capabilities.

1. Does Utilizing Multiple Modalities Improve Generalizability? Our embodied agents are equipped to support multi-modal inputs. This section examines the benefits of integrating multiple modalities. Table 2 demonstrates

Table 1. Experiments on Manisill-2 over six rigid body and soft body tasks. Our method consistently outperforms Baseline in all environments. All metrics are reported in percentage $(\%)$ with the best ones bolded.

Methods	PickCube	StackCube	PickSingleYCB	Fill	Hang	Excavate
Observation Modality: Image						
ResNet152 [26]	40.1	86.3	22.1	52.4	82.6	12.0
Swin-Base [38]	41.3	83.5	28.5	49.0	82.7	14.8
RAEA	**56.7**	**93.6**	**40.2**	**63.8**	**87.1**	**22.4**
Observation Modality: Point Cloud + Image						
ResNet152 + PointNet [26, 51]	56.6	90.7	28.4	46.9	85.7	18.4
Swin-Base + PointNet [38, 51]	44.0	90.1	30.6	45.5	86.4	17.0
RAEA	**62.7**	**91.0**	**43.8**	**71.7**	**88.1**	**24.0**

Table 2. Ablation study on the effect of using different modalities for real-world environments.

Method	Instruction	Observation	Success Rate
RAEA	Text	Image	54
	Text+Audio	Image	56
	Text+Audio	Image + Point Cloud	65
	Image	Image	58
	Image+Video	Image	63
	Image+Video	Image + Video	64
	Image+Video+Text	Image	69

Table 3. Ablation study on the effect of status information, i.e., action & proprioceptive state, in real-world data. The experiments are conducted based on text-image pairs.

Tasks	Status	Success Rate
RAEA	All	54
	Proprioception	39
	Action & Proprioception	36

Table 4. Ablation study on the data for policy memory bank using Franka-only or all embodiments.

Tasks	Embodiments	Success Rate
RAEA	All	54
	Franka-Only	48

that employing more available modalities can stably yield better performance. Notice that using the combination of language and visual as instruction significantly enhance the generalizability of the model, i.e., increase the success rate from 63 to 69. Also, adopting 3D information, such as point cloud, can be useful, which improves the success rate from 56 to 65.

2. *Does Including More Status Information in Retrieved Policies Enhance Policy Learning?* While instruction and observation are fundamental components, our research also delves into the impact of incorporating proprioception and action data. As illustrated in Table 3, there is a discernible decrease in success rate when proprioception and action are omitted from the retrieved policy, dropping from 54 to 36. This finding highlights the critical role these elements play in augmenting the efficacy of our learning approach.

3. *Is Retrieval Across Different Embodiments Beneficial?* In our experiments, we utilized the Open X-Embodiment as our primary policy memory bank, a dataset featuring a variety of embodiments. Our evaluation, depicted in Table 4, focuses on the performance implications of using data exclusively from Franka robots compared to a multi-embodiment dataset. We observed a slight decline in performance with the Franka-only data. This could be attributed to the richer diversity of environments and commands present in the broader dataset.

5. Conclusion

Training with pre-defined datasets often leads to a limited scope of ability and knowledge acquisition. In this paper, we introduce Retrieval-Augmented Embodied Agent (RAEA), a novel framework that enhances an embodied agent through a policy retriever and generation process. The primary objective of this retrieval process is to efficiently and effectively harness valuable insights from a comprehensive dataset of experiences, thereby aiding the agent in achieving its goals more proficiently. Through multiple ablation studies, we have underscored the significance of the various components within RAEA. Overall, our RAEA methodology presents an innovative and practical approach to leveraging collective knowledge from diverse datasets of different embodiments.

References

[1] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022. 2

[2] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[3] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. First person action-object detection with egonet. *arXiv preprint arXiv:1603.04908*, 2016. 2

[4] Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, et al. Robocat: A self-improving foundation agent for robotic manipulation. *arXiv preprint arXiv:2306.11706*, 2023. 1

[5] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8709–8719, 2019. 3

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 2

[7] Kaylee Burns, Tianhe Yu, Chelsea Finn, and Karol Hausman. Pre-training for manipulation: The case for shape biased vision transformers. . 3

[8] Kaylee Burns, Tianhe Yu, Chelsea Finn, and Karol Hausman. Pre-training for manipulation: The case for shape biased vision transformers. . 2

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[10] Annie S Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from" in-the-wild" human videos. *arXiv preprint arXiv:2103.16817*, 2021. 2

[11] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*, 2022. 2

[12] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 2

[13] Yuchen Cui, Scott Niekum, Abhinav Gupta, Vikash Kumar, and Aravind Rajeswaran. Can foundation models perform zero-shot task specification for robot manipulation? In *Learning for Dynamics and Control Conference*, pages 893–905. PMLR, 2022. 2

[14] Zichen Jeff Cui, Yibin Wang, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. From play to policy: Conditional behavior generation from uncurated robot data. *arXiv preprint arXiv:2210.10047*, 2022. 2

[15] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, et al. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019. 3

[16] Amaury Depierre, Emmanuel Dellandréa, and Liming Chen. Jacquard: A large scale dataset for robotic grasp detection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3511–3516. IEEE, 2018. 3

[17] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018. 3

[18] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021. 2, 3

[19] Clemens Eppner, Arsalan Mousavian, and Dieter Fox. Acronym: A large-scale grasp dataset based on simulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6222–6227. IEEE, 2021. 3

[20] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020. 3

[21] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A robotic dataset for learning diverse skills in one-shot. *arXiv preprint arXiv:2307.00595*, 2023. 3

[22] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020. 5, 6, 1

[23] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 4, 1

[24] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023. 5, 6

[25] Nicklas Hansen, Zhecheng Yuan, Yanjie Ze, Tongzhou Mu, Aravind Rajeswaran, Hao Su, Huazhe Xu, and Xiaolong Wang. On pre-training for visuo-motor control: Revisiting a learning-from-scratch baseline. *arXiv preprint arXiv:2212.05749*, 2022. 2

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7, 8

[27] Felix Hill, Sona Mokra, Nathaniel Wong, and Tim Harley. Human instruction-following with deep reinforcement learning via transfer-learning from text. *arXiv preprint arXiv:2005.09382*, 2020. 2

[28] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 2

[29] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022. 2

[30] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2022. 2

[31] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 1

[32] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018. 3

[33] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020. 3

[34] Aviral Kumar, Anikait Singh, Frederik Ebert, Mitsuhiko Nakamoto, Yanlai Yang, Chelsea Finn, and Sergey Levine. Pre-training for robots: Offline rl enables learning new tasks from a handful of trials. *arXiv preprint arXiv:2210.05178*, 2022. 3

[35] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37 (4-5):421–436, 2018. 3

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 6

[37] Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15148–15158, 2023. 2

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7, 8

[39] Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648*, 2020. 2

[40] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *arXiv preprint arXiv:1703.09312*, 2017. 3

[41] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023. 6

[42] Suraj Nair, Eric Mitchell, Kevin Chen, Silvio Savarese, Chelsea Finn, et al. Learning language-conditioned robot behavior from offline data and crowd-sourced annotation. In *Conference on Robot Learning*, pages 1303–1315. PMLR, 2022. 2

[43] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. 3, 6

[44] Soroush Nasiriany, Vitchyr Pong, Steven Lin, and Sergey Levine. Planning with goal-conditioned policies. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[45] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[46] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 2, 3, 1

[47] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. In *International Conference on Machine Learning*, pages 17359–17371. PMLR, 2022. 2

[48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[49] Deepak Pathak, Parsa Mahmoudieh, Guanghao Luo, Pulkit Agrawal, Dian Chen, Yide Shentu, Evan Shelhamer, Jitendra Malik, Alexei A Efros, and Trevor Darrell. Zero-shot visual imitation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2050–2053, 2018. 2

[50] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[51] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 7, 8

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[53] Ilija Radosavovic, Baifeng Shi, Letian Fu, Ken Goldberg, Trevor Darrell, and Jitendra Malik. Robot learning with sensorimotor pre-training. *arXiv preprint arXiv:2306.10007*, 2023. 3

[54] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 4

[55] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022. 2

[56] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018. 2

[57] Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. Mutex: Learning unified policies from multimodal task specifications. *arXiv preprint arXiv:2309.14320*, 2023.

[58] Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research*, 40(12-14):1419–1434, 2021.

[59] Pratyusha Sharma, Deepak Pathak, and Abhinav Gupta. Third-person visual imitation learning via decoupled hierarchical controller. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[60] Philip Shilane, Patrick Min, Michael Kazhdan, and Thomas Funkhouser. The princeton shape benchmark. In *Proceedings Shape Modeling Applications, 2004.*, pages 167–178. IEEE, 2004. 3

[61] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023. 2

[62] Laura Smith, Nikita Dhawan, Marvin Zhang, Pieter Abbeel, and Sergey Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos. *arXiv preprint arXiv:1912.04443*, 2019. 2

[63] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, et al. Bridgedata v2: A dataset for robot learning at scale. *arXiv preprint arXiv:2308.12952*, 2023. 2

[64] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, et al. Bridgedata v2: A dataset for robot learning at scale. *arXiv preprint arXiv:2308.12952*, 2023. 3

[65] Yixuan Wang, Zhuoran Li, Mingtong Zhang, Katherine Driggs-Campbell, Jiajun Wu, Li Fei-Fei, and Yunzhu Li. D3̂ fields: Dynamic 3d descriptor fields for zero-shot generalizable robotic manipulation. *arXiv preprint arXiv:2309.16118*, 2023. 2

[66] Junjie Wen, Yichen Zhu, Minjie Zhu, Jinming Li, Zhiyuan Xu, et al. Object-centric instruction augmentation for robotic manipulation. *arXiv preprint arXiv:2401.02814*, 2024. 2

[67] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 4

[68] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19265–19274, 2023. 2

[69] Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, et al. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*, 2022. 2

[70] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*, 2021.

[71] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*, 2022. 2

[72] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. 2023. 2

[73] Kuan-Ting Yu, Maria Bauza, Nima Fazeli, and Alberto Rodriguez. More than a million ways to be pushed. a high-fidelity experimental dataset of planar pushing. In *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 30–37. IEEE, 2016. 3

[74] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020. 5, 6, 1

[75] Zhecheng Yuan, Zhengrong Xue, Bo Yuan, Xueqian Wang, Yi Wu, Yang Gao, and Huazhe Xu. Pre-trained image encoder for generalizable visual reinforcement learning. *Advances in Neural Information Processing Systems*, 35:13022–13037, 2022. 2

[76] Minjie Zhu, Yichen Zhu, Jinming Li, Junjie Wen, Zhiyuan Xu, et al. Language-conditioned robotic manipulation with fast and slow thinking. *arXiv preprint arXiv:2401.04181*, 2024. 2

[77] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *7th Annual Conference on Robot Learning*, 2023. 1